18TH INTERNATIONAL SYMPOSIUM ON AVIATION PSYCHOLOGY

May 4 - May 7, 2015

WRIGHT STATE UNIVERSITY, DAYTON, OHIO

WRIGHT STATE UNIVERSITY

For your convenience, the table of contents is linked to the body of the document and is keyword searchable

Symposium: Technological Countermeasures for Spatial Disorientation

Near-Future Technological Countermeasures for Spatial Disorientation in Flight
Littman, Lawson, Brill, Rupert……...1

Recent Advances in Tactile Cueing
Rupert, Lawson……... 7

Audiotactile Aids for Improving Pilot Situation Awareness
Brill, Lawson, Rupert…… 13

Tactile Cueing Strategies to Convey Aircraft Motion or Warn of Collision
Lawson, Cholewiak, Brill, Rupert, Thompson.…….19

Requirements for Developing the Model of Spatial Orientation into an Applied Cockpit Warning System
Lawson, McGrath, Newman, Rupert……. 25

Safety and Risk Analysis

Behind the Scenes of the NAS: Human Factors Taxonomy for Investigating Service Integrity Events
Berry, Sawyer, Hinson……... 31

A Comprehensive Effort to Arrive at an Optimally Reliable Human Factors Taxonomy
King…… 37

Promoting Aviation Safety in Africa: Analysis of Air Accidents in the Region Between 2004 and 2013
Shila, Anne……. 43

Performance-Enhancing Displays

Possibilities of Using the On-Board Intelligent Voice Informing Systems in Complex Flight Situations
Petrenko…….. 49

Toward An Integrated Ecological Plan View Display For Air Traffic Controllers
Beernink, Borst, Ellerbroek, Van Paassen, Mulder…… 55

The Smart Cockpit Initiative
Smith, Larrieu……. 61

Symposium: High-Fidelity Simulation and Aviation Training to Improve Problem Solving Skills and Coordination

The Effects of Workload and Stress on Teamwork in a High Fidelity Simulation
Georgiou……. 67

After-Action Reviews: Best Practices and Application to Aerospace Education
Moffett III, Hein, McClure…… 73

Development of an Alternative Methodology for Implementation of SAGAT During Task Performance
Bridges……... 79

High Fidelity Simulation and Aviation Training to Improve Problem Solving Skills and Coordination
Lester, Craig……... 85

Excellence in Air Traffic Management

Toward a Human Performance Standard of Excellence in Air Traffic Management
Krois, Armenis, Joly, Kirwan, Marrison, May, Piccione, Schwarz……... 90

Impact of Nextgen on National Airspace Actors
Krokos, Sawyer, Berry……... 96

Performance Assessment Methods to Evaluate Discretionary ATC Safety Standards
Pounds, Galoci…….. 101

Developing Quantitative Air Traffic Risk-Benefit Pathways for Class Delta Airports: Improving Small Tower Operations
Berry, Sawyer, Hinson……. 105

UAV Selection and Assessment

The Role of Personnel Selection in Remotely Piloted Aircraft Human System Integration
Carretta, King…….. 111

Investigating UAS Operator Characteristics Influencing Mission Success
Cuevas, Kendrick, Zeigler, Hamilton……. 117

Displays for Adverse Environments

Impact of Weather Information Latency on General Aviation Pilot Situation Awareness
Caldwell, Johnson, Whitehurst, Rishukin, Udo-Imeh, Duran, Nyre, Sperlak……..... 123

Avionics Touch Screen in Turbulence: Simulation for Design
Hourlier, Servantie…….. 129

Identifying Representative Symbology for Low Visibility Operations/Surface Movement Guidance and Control System (LVO/SMGCS) Paper Charts
Sparko, Chase…….. 135

Monitoring and Supervisory Control

Human Span-of-Control in Cyber Operations: An Experimental Evaluation of Fan-Out
Mancuso, Funke, Eckold, Strang…… 141

Symposium: Terminal or En Route? That is the Question: Placement of Development Air Traffic Controllers

Perspectives of Unsuccessful Air Traffic Control Specialists
Pierce, Byrne…….. 147

An Evaluation of the Utility of AT-SAT for the Placement of New Controllers by Option
Byrne, Broach…… 153

Displays for Air and Ground Maneuvering

Learning of Location-Identity Bindings: Development of Level 1 Situation Awareness in an Air Traffic Control-Like Task
Nalbandian, Rantanen……... 159

UAV Operations

An Operational Analysis of Human Factors in an Unmanned Air System
Revell, Cutler…… 165

Experimental Evaluation of Varying Feedback of a Cognitive Agent System for UAV Mission Management
Denk, Clauss, Borchers, Werner, Schulte…….. 171

An Ecological Approach To The Supervisory Control Of UAV Swarms
Fuchs, Borst, de Croon, van Paassen, Mulder……... 177

Effect of Control Latency on Unmanned Aircraft Systems During Critical Phases of Flight
Zingale, Taylor……... 183

Cognitive Challenges in Aviation

Attentional Narrowing: A First Step Towards Controlled Studies Of A Threat to Aviation Safety
Prinet, Sarter……... 189

The Cognition of Multi-Aircraft Control (MAC): Proactive Interference and Working Memory Capacity
Amaddio, Miller, Elshaw, Finomore…….. 195

Visual Search and Target Selection Using a Bounded Optimal Model of State Estimation & Control
Perelman, Myers……. 201

Tuesday Posters:

Anticipatorily Controlled Top-Down Processes Influence the Impact of Coriolis Effects
Talker, Kallus……. 207

Procedure Used for Establishing Screening Test Cut-Points Based on Aviation Occupational Task Performance
Milburn, Chidester, Gildea, Peterson, Roberts, Perry…… 213

Conceptual and Procedural Training for Situation Awareness and Performance in an Instrument Holding Task
Dattel, Thropp…… 219

Psychological Aspects of the Organization of Information at the Instructor's Flight Simulator Workplace
Bondareva…….. 225

Concept of Flight Instructor Assistance in Helicopter Emergency Medical Service Using Pilot Trainee's Workload Determination
Maiwald, Schulte…… 230

Contribution of Multimethodology to Human Factors in Air Navigation Systems
Cabral, Estellita Lins…….. 236

A Valid and Reliable Safety Scale for Passenger's Perceptions of Airport Safety
Rice, Mehta, Winter, Oyman……. 242

Age and Trust in Flight Attendants: A Comparison Between Two Countries
Mehta, Rao, Labonte, Rice……..... 248

A Regression of Consumer Attitudes Toward Airport Water Reuse
Cremer, Rice, Winter……. 254

Development of the Air Traffic Control Tower Alerts Standard
Sierra, Buckley……... 260

A³IR-CORE and FlightProfiler: An Academic-Industry Partnership for SMS Development
Mott, Ball……... 266

Nigeria's Aviation at a Glance: The Assessment of Nigerians' Perceived Trust Level in Nigeria's Aviation Industry
Miya, Rice…….. 272

Evaluating Startle, Surprise, and Distraction: An Analysis of Aircraft Incident and Accident Reports
Talone, Rivera, Jimenez, Jentsch……... 278

Consumer Trust Ratings After an Airline Accident: An Effective Perspective
Winter, Rice, Cremer, Mehta……. 284

Panel: Techniques for the Human Centered Evaluation of Designs for the Future Aviation System

Techniques for the Human Centered Evaluation of Designs for the Future Aviation System
Smith, Abbott, Prinzel, Pritchett, Yuditsky……... 290

Controller-Pilot Interaction

Planning for the Future: Human Factors in Nextgen Air Traffic Management
Austrian, Berry, Sawyer, DeHaas……. 294

Assessing Potential Human Performance Safety Impacts Associated With Integrating Multiple Time-Based Flow Management Concepts
Sawyer, Berry, Liskey, Rohde……... 300

Computational Simulation of Authority-Responsibility Mismatches in Air-Ground Function Allocation
Ijtsma, Bhattacharyya…… 306

UAVs and ATC

Integrating UAS Operations in Class C Airspace
Truitt, Sollenberger……... 312

UAS Air Traffic Controller Acceptability Study-2: Effects of Communications Delays and Winds in Simulation
Comstock, Ghatas, Consiglio, Chamberlain, Hoffler……... 318

UAS in the NAS Air Traffic Controller Acceptability Study-1: The Effects of Horizontal Miss Distances on Simulated UAS and Manned Aircraft Encounters
Ghatas, Comstock, Jr., Consiglio, Chamberlain, Hoffler……. 324

Cognitive System Engineering Applications

Incorporating New Methods of Classifying Domain Information for Use in Safety Hazard Analysis
Leveson, Montes, Stirling……. 330

Individual Problem Representations in Distributed Work
Fernandes, Smith, Durham, Evans…… 336

Experimental Investigation of Flight Crew Strategies in Handling Unexpected Events
Field, Woltjer, Rankin, Mulder……. 342

Statistical Analysis and Quality Assurance

Statistical Errors in Aviation Psychology: Commonsense Statistics in Aviation Safety Research
Wickens…… 348

Flight Operational Quality Assurance (FOQA): Do Exceedances Tell the Whole Story?
Dillman, Wilt, Pruchnicki, Rudari, Ball, Pomeroy……... 354

Un-alerted Smoke and Fire: Checklist Content and Intended Crew Response
Burian……... 360

Crew Resource Management

Reliability of Instructor Pilots' Non-Technical Skills Ratings
Gontar, Hoermann…… 366

Behavioral Traps In Crew-Related Aviation Accidents
Velazquez, Peck, Sestak……... 372

Training

FAA Training Assessment of On-the-Job Training
McCauley…….. 378

Exploring the Mathematical Predictability of the Advanced Aircraft Training Climate
Naidoo……... 382

Airframe Parachute Knowledge and Deployment Scenarios: A Collegiate Perspective
Winter, Geske, Rice, Fanjoy, Sperlak……... 388

Wednesday Posters:

Detecting Structure in Activity Sequences: Exploring the Hot Hand Phenomenon
Hammack, Flach, Houpt…… 394

Identifying Mental Models of Search in a Simulated Flight Task Using a Pathmapping Approach
Perelman, Mueller……. 398

Multi-Gain Control: Balancing Demands for Speed and Precision
Lemasters, Flach……... 404

Haptic Guidance: Interaction Between the Guidance Model and Tuning
van Paassen, Boink, Abbink, Mulder, Mulder……. 410

Design and Evaluation of a Haptic Display for Flight Envelope Protection Systems
Ellerbroek, Rodriguez y Martin, van Paassen, Mulder……. 416

Open Source Devices for Human Factors Research
Gildea, Milburn……..... 422

Physiological Indicators of Workload in a Remotely Piloted Aircraft Simulation
Hoepf, Middendorf, Epling, Galster……. 428

EEG Data Analysis Using Artifact Separation
Credlebaugh, Middendorf, Hoepf, Galster……... 434

A Coalition Study of Warfighter Acceptance of Wearable Physiological Sensors
Menke, Best, Funke, Strang…….. 440

Modeling Task Prioritization Behaviors in a Time-Pressured Multitasking Environment
Toma, Funk……... 446

Using Augmented Reality and Computer-Generated Three-Dimensional Models to Improve Training and Technical Tasks in Aviation
Anne, Wang, Ropp……... 452

Impact of Task Load and Gaze on Situation Awareness in Unmanned Aerial Vehicle Control
Coyne, Sibley…… 458

Fusion: A Framework for Human Interaction with Flexible-Adaptive Automation Across Multiple Unmanned Systems
Rowe, Spriggs, Hooper……. 464

Visualization Methods for Communicating Unmanned Vehicle Plan Status
Behymer, Ruff, Mersch, Calhoun, Spriggs……... 470

Symposium on Cross-Cultural Pilot Selection

Panel on Cross-Cultural Pilot Selection
Damos, Rose, Martinussen, Lorenz……... 476

Automation Impacts

Inverting the Human/Automation Equation to Support Situation Awareness and Prevent Loss of Control
Kirlik, Ackerman, Seefeldt, Xargay, Talleur, Carbonari, Hovakimyan, Sha, Trujillo, Gregory…….. 482

Functional Complexity Failures and Automation Surprises: The Mysterious Case of Controlled Flight into Stall (CFIS)
Sherry, Mauro……... 488

Understanding Automation Surprise: An Analysis of ASRS Reports
Trippe, Mauro……... 494

Eye-Based Workload and SA

Flight Deck Interval Management Avionics: Eye-Tracking Analysis
Latorella, Harden……... 500

Pupillary Response as an Indicator of Processing Demands within a Supervisory Control Simulation Environment
Sibley, Coyne, Doddi, Jasper…… 506

The Electrooculogram and a New Blink Detection Algorithm
Epling, Middendorf, Hoepf, Gruenwald, Stork, Galster……... 512

Saccade Detection Using Polar Coordinates - A New Algorithm
Middendorf, Gruenwald, Stork, Epling, Hoepf, Galster……... 518

Simulation and Training

Enroute ATC Industry Perceptions of Simulation Fidelity
Dow, Histon……... 524

Follow-Up Examination of Simulator-Based Training Effectiveness
Lubner, Dattel, Henneberry, DeVivo…… 530

Evaluation of an Eye Tracking-Based Assessment and Debrief Tool for Training Next Generation Multirole Tactical Aviation Skills
Carroll, Surpris, Sidor, Bennett…… 536

A Multi-Year Study of the Safety and Training Impacts of Introducing the Live-Virtual-Constructive Training Strategy into Navy Air Combat
Sherwood, Neville, McLean, Cruit, Kaste, Walwanis, Bolton……. 542

Symposium: Manned-Unmanned Aviation Operations - A Tri-Service Overview

Training Manned-Unmanned Teaming Skills in Army Aviation
Flaherty, Bink……... 548

Optimizing Performance of Trainees for UAS Manpower, Interface and Selection (OPTUMIS): A Human Systems Integration (HSI) Approach
Pagan, Astwood, Phillips…….. 554

Army Aviation Manned-Unmanned Teaming (MUM-T): Past, Present, and Future
Taylor, Turpin……... 560

Remote-Split Operations and Virtual Presence: Why the Air Force Uses Officer Pilots to Fly RPAs
Martin……... 566

Psychophysiological Applications

Compared Evaluation of B-Alert's Encephalographic Workload Metrics Using an Operational Video Game Setup
Lini, Bey, Lecoutre, Lebour, Favier……. 572

Communications Effectiveness

'We Need Priority Please' – Mitigated Speech in the Crash of Avianca Flight 052
Cookson…… 578

The Effect of Asynchronous Data on Pilot-Controller Communication in a Dynamic Environment with Subject-Matter Experts
Lien, Histon……... 584

Safety Culture

Social Complexity: The Missing Link in a Critical Incident Reporting System
van der Westhuizen, Stanz…… 590

Pilots' Willingness to Report Aviation Incidents
Haslbeck, Schmidt-Moll, Schubert……... 596

Fatigue

Effects of Workload on Measures of Sustained Attention during a Flight Simulator Night Mission
Hoermann, Gontar, Haslbeck…… 602

The Effects of Bright Light Intervention on Flight Crew Behavioral Alertness and Cognitive Fatigue
Brown, Whitehurst…… 608

Helmet-Mounted Displays

Toward Head-Up and Head-Worn Displays for Equivalent Visual Operations
Prinzel, Arthur, Bailey, Shelton, Kramer, Jones, Williams, Harrison, Ellis…… 614

Selection Methodology

Simulator-Based Assessment of Flight-Specific Aptitudes in German Armed Forces' Aircrew Selection
Meierfrankenfeld, Greß, Vorbach…… 620

Cognitive Engineering: What's Old Is New Again
Lofaro…… 624

NEAR-FUTURE TECHNOLOGICAL COUNTERMEASURES FOR SPATIAL DISORIENTATION IN FLIGHT

Eric M. Littman, Naval Aerospace Medical Research Unit Dayton, Dayton, OH

Ben D. Lawson, U.S. Army Aeromedical Research Laboratory (USAARL), Fort Rucker, AL

J. Christopher Brill, Old Dominion University, Norfolk, VA

Angus H. Rupert, USAARL, Fort Rucker, AL

Spatial Disorientation (SD) is an important cause of deadly aircraft mishaps, despite improvements in night vision, head-up-displays, cockpit automation, etc. This paper explores several technological countermeasures for SD. This report begins by discussing the magnitude of the SD problem and the reasons why technological countermeasures are needed. The authors discuss the three main approaches that are typically used (improved selection, training, or technology) to decrease the incidence of SD, and argue that improved selection and training, although beneficial, are not sufficient by themselves to prevent SD. The authors introduce various technological solutions they are developing, including better models to predict disorientation, as well as better cockpit displays to provide accurate earth-referenced visual, auditory, or tactile cues. The authors describe how these technological approaches should benefit situational awareness, spatial localization, detection of sub-threshold vehicle motion, and prevent imminent collision with objects that are not being attended to by the pilot.

Aviation spatial disorientation (SD) is best described as a pilot's inability to correctly interpret aircraft attitude, altitude, or airspeed in relation to the earth (Benson, 2006). It is well known that SD in manned aviation can contribute to various accidents and even the loss of aircraft. Moreover, SD is the number one cause of Class A mishaps. These are incidents where the total cost of damage is $1 million or more and/or the aircraft is destroyed and/or fatal injury and/or a permanent total disability has occurred. Spatial disorientation and the loss of situation awareness (SA) also occur in unmanned aviation, even with operators on the ground. Interestingly, losses of aircraft and equipment in manned aviation over the last decade are very low in comparison to unmanned aviation where losses are high and increasing (Zirkelbach, 2007). McCauley & Matsangas (2004) showed that maintaining SA is also a key factor when operating unmanned aerial vehicles (UAVs). Although this paper is not specifically about UAVs, it is important to keep in mind that UAV operators and pilots of manned aircraft face similar challenges when it comes to maintaining SA. Three primary approaches (training, improved selection, and technology) have been used to decrease the incidence of SD. Although training and improved selection can be beneficial, they are not sufficient in and of themselves to effectively prevent SD. Technological solutions, in conjunction with effective training and selection, may provide a better means of enhancing and maintaining SA and, thus, preventing SD.

Regardless of a pilot's experience or expertise, sensory illusions can lead to perceived discrepancies between instrument indications and what the pilot feels the aircraft is doing (Zirkelbach, 2007). The subsequent mishaps are not only costly to the military, but often result in the loss of human life. For example, between 1993 and 2013, the United States Air Force (USAF) experienced 72 SD related Class A mishaps resulting in a loss of 65 aircraft for a total cost of $2.32 billion and even more important and unfortunate, the loss of 101 lives (Poisson, 2014). These consequences of SD are enormous, in the cost of lost aircraft, lost aircrew, and the cost of training new aircrew (Heinle & Ercoline, 2003; Zirkelbach, 2007). Additionally, between 1992 and 2000, SD caused 20.2% of USAF Class A mishaps. During an equivalent period, SD caused 27% of U.S. Army mishaps and 26% of U.S. Navy (USN) mishaps. In general, SD is still the most common cause of human-related aircraft accidents (Heinle & Ercoline, 2003).

Selection as a Countermeasure

Aviation, both manned and unmanned, provides numerous advantages, not just for transportation and combat, but also for intelligence collection, surveying and monitoring, etc. However, the technological complexity of cockpit designs and UAV controls, combined with stressful situations, adverse weather, and other workload drivers can cause problems for aircrew and may increase the likelihood of pilots becoming spatially disoriented. Thus, selecting the best individuals for the task is incredibly important.

Numerous pilot selection methods have been utilized since the beginning of manned flight. General screening procedures have included age, physical condition, and general intelligence. In the U. S. military, paper-and-pencil perceptual-motor and cognitive tests have served as traditional selection tools (common examples include: the Aircrew Classification Battery, the Wonderlic Personnel Test, Spatial Apperception Test, Academic Qualification Tests, Mechanical Comprehension Tests, etc.).

Traditionally, perceptual/cognitive selection test batteries consist of at least four components: 1) a general intelligence tests that has both quantitative and verbal items, 2) a spatial test (e.g. the Spatial Apperception Test), 3) a mechanical comprehension test, and 4) a background/biographical inventory. The Aviation Selection Test Battery (ASTB), used by the USN, was first created in 1942 but has since gone through a few revisions (Williams, Albert, & Blower, 1999). The 1992 version of the ASTB has six subtests: the mechanical comprehension test, the math-verbal test, the spatial apperception test (which measures spatial reasoning abilities), the aviation and nautical information test, the biographical inventory (which contains personal history and general interests questions), and the aviation interest test (Williams, Albert, & Blower, 1999).

The six ASTB subtests are weighted and combined in order to calculate three validated scores utilized in pilot selection. The Pilot Flight Aptitude Rating (PFAR) is a validated predictor of flight grades during primary flight training; the Academic Qualification Rating (AQR) is a validated predictor of academic performance during ground school; and, finally, the Pilot Biographical Inventory (PBI) is a validated predictor of attrition during primary flight training (Williams, Albert, & Blower, 1999). Data provided by the Naval Operational Medicine Institute (NOMI[1]), the organization responsible for overseeing the ASTB testing program, indicates that approximately half of the individuals who take the ASTB do not meet the minimum naval aviation selection scores (Williams, Albert, & Blower, 1999). Ultimately, after the additional steps in the selection process (e.g. physical examination, interview, board evaluation), only 15% are selected to begin training (Williams, Albert, & Blower, 1999).

Although the ASTB has been shown to be a useful and valid selection tool, it is not without its own shortcomings; namely, it and other paper and pencil tests fail to reflect the dynamic cockpit environment inside military aircraft. As a result, the Navy recently developed a Performance-Based Measurement Battery (PBMB) as a supplement to the ASTB. The PBMB is administered online through the Automatic Pilot Examination (APEX) system, which affords opportunities for more dynamic assessments. The last three of the seven subtests involve multi-tasking and/or emergency procedures scenario. These multi-tasking elements and the emergency procedures scenario are aimed at assessing dynamic skills that pilots need to effectively operate military aircraft. For example, Ostoin (2007) found that the PBMB detected important hand-eye coordinated tracking skills, something that the paper and pencil ASTB cannot assess.

It is important to note that although selection tools are helpful in general, there is currently no selection tool specifically designed to assess aviators who are particularly sensitive to SD. While all aviators with functioning vestibular organs can be made disoriented, some healthy aviators will be more

[1] On 26 October 2011, NOMI was realigned directly under Navy Medicine Support Command (NMSC) and changed its name to Navy Medicine Operational Training Center (NMOTC).

susceptible to SD than others. Moreover, some aviators who appear healthy actually may have latent vestibular pathologies rendering them much more susceptible to SD than a normal person. These gap areas need further investigation. Despite the many current unknowns in selection, selecting pilots who are best suited to handle the complexities and challenges of aviation is very important. In general, selection is helpful in identifying intelligent and resilient individuals with the "strongest foundations" on which to build. However, those ultimately selected to become experienced aviators, are not immune from losing their SA and becoming spatially disoriented. Relatedly, Matthews, Previc, and Bunting (2003) found that pilot experience was a strong predictor of reporting SD incidents. Not only were more experienced pilots more likely to have more opportunities to experience SD, but they also reported a higher frequency of each SD illusion, independent of the total number of sorties flown (Matthews et al, 2003). They suggested that experienced pilots are better able to recognize specific types of SD compared to less experienced pilots. It may also be the case that more seasoned and confident pilots are more willing to report their SD experiences, knowing that it happens to virtually all pilots. Nevertheless, increased experience, or expertise, did not translate to a reduction in SD incidence. Therefore, one must examine other approaches to decreasing the incidence of SD, like training.

Training as a Countermeasure

The U.S. Army, USAF, and USN customarily have four primary components to their SD training curricula: some initial classroom-based training, ground-based demonstrations and/or simulations of SD, flight-based demonstrations, and finally, some type of refresher course or brief (Guckenberger & Bryan, 2003). Due to high costs and time constraints the tri-services have all cut back, if not eliminated entirely, flight-based SD demonstrations.

The classroom and refresher components not only provide general overviews of the problems and dangers of SD and the importance of maintaining proper orientation, but they also focus on educating the students on topics such as the orientation-specific aspects of the visual, vestibular, and proprioceptive systems, as well as the psychological aspects of orientation. Ground-based demonstrations are important training tools because they afford the opportunity to let individuals experience how their own senses can be fooled in some of the disorientation situations. Typically, Bárány chairs, and other disorientation devices, are used to elicit various disorientation illusions. Additionally, these types of demonstrations are useful in that they are (relatively) low cost and safe from the hazards of actual flight.

Although more costly and perhaps more hazardous, in-flight demonstrations afford the opportunity to experience SD situations first hand. Flight surgeons typically fly these sorties in order to better explain and point out the various facets of SD. Moreover, these sorties are flown so that students can experience SD for themselves, but with the safeguard of having an instructor pilot (IP) aboard. These in-flight training experiences are important because they allow for the IP to teach about flying conditions that can lead to SD situations. Moreover, these flights also afford the teaching of mechanisms that can be used to cope with the illusions after they have occurred (Braithwaite, Hudgens, Estrada, & Alvarez, 1998).

Although training has been shown to be beneficial in enabling pilots to identify SD situations, it is limited by the fact that the training is often aircraft-type specific. Although some generic training is possible, it is most effective when it addresses specific aircraft and their respective designs, specific operational scenarios and specific environmental situations. For example, Williams and colleagues (2014) demonstrated that spatial strategies training can help pilots avoid low nighttime approaches, thus "combating" the black hole illusion. After the training, pilots were able to perform nighttime approaches similar to those performed during the daytime with a visible horizon. In short, this training was very successful, but it was also very specific and does not necessarily extend to other spatially disorienting situations/illusions.

Similarly, skills obtained from generic training do not necessarily translate well to other, more specific aircraft. For example, the USAF Undergraduate Pilot Training (UPT) has a generalized flight training component. Although beneficial in establishing a good foundation and starting point for pilots, one of the major shortcomings of the UPT generalized flight training was that it failed to provide student pilots with knowledge and specialized skills that they would need in order to smoothly transition to the tanker and large transport aircraft that account for one third of the USAF fleet (Weeks, Zelenski, & Carretta, 1996).

Research over the last few decades has demonstrated that SD training is effective, albeit often limited to specific aircrafts and/or specific situations and illusions. The tri-services' training curricula are very beneficial in educating students pilots about the dangers of SD—they also provide important learning opportunities for student pilots prior to actually flying their own aircraft. Selection, training, and experience are beneficial in helping pilots avoid losing their SA, but they are not foolproof. Arguably, they are necessary but not sufficient tools to combat SD. The addition of technological countermeasures may provide the key to bridging the gap, so to speak, and may provide pilots with another tool to better maintain their SA and avoid becoming spatially disoriented.

Technology as a Countermeasure

Under degraded visual conditions and especially if one is suffering from SD, pilots are taught to transition to instruments in order to discern their aircraft's true attitude. The head-down displays (HDDs; e.g. the attitude indicator, altimeter, heading and airspeed indicators), which have become standard in aircraft cockpits, are arranged in a "T" formation to facilitate a quick scan and determination of the aircraft's attitude (Albery, 2007). Numerous research studies over the last few decades have examined different display designs, layouts, configurations, etc. in order to improve the pilot's ability to determine the aircraft's true attitude and subsequently maintain SA. Although these instruments can be helpful, they do not eliminate SD. More recently, research has focused on alternative technological countermeasures that may be more successful at combating SD.

For example, Poisson (2014) examined the efficacy of an Attitude Stabilization Display (ASD), which differs from the standard attitude indicator by providing an auditory alert when the aircraft enters an unexpected attitude; the display is equipped with a "more intuitive graphical interface" and the ASD provides a very specific recommendation of a course of action to maneuver out of the unexpected attitude and return the aircraft to upright, wings level flight (Poisson, 2014). Although Poisson found mixed results when comparing the ASD to the traditional attitude indicator, he notes that the intent behind the ASD was to aid the pilot during spatially disorienting situations by alerting the pilot and drawing his or her attention toward the aircraft's instruments. Additionally, he stresses the importance that symbology and alerting systems, whatever type they may be, ought to be designed in a way that affords "quick and accurate recovery" from unexpected attitudes. Although eliminating SD may not be possible for any single instrument or device, multisensory technologies may provide pilots with the widest array of tools to help maintain SA and avoid SD.

Cockpits are loaded with display systems, dials, and meters, which inundate pilots with visual information. Traditionally, pilots have also been tasked with processing and making sense of this plethora of visual information. More recently, research has shifted toward designing more intuitive cockpit display systems, like the ASD, in order to minimize pilot workload. Additionally, multisensory technologies, like tactile and audiotactile cueing systems, have been developed in order to provide pilots with feedback that is not exclusively visual.

In degraded visual environments and under high workload conditions, the Tactile Situation Awareness System (TSAS) has been found to improve pilot performance (McGrath et al., 2004). Pilots

flying the UH-60 Black Hawk simulator were able to use the tactile cues provided by the TSAS as a secondary source of sub-threshold drift information while operating under degraded visuals and in environments prone to SD illusions (e.g. height-depth perception illusion). The TSAS allowed pilots to better concentrate on mission tasks, thus resulting in increased SA and reduced workload (McGrath et al., 2004).

Audio feedback is another type of sensory cue that pilots could use to improve their SA; however, 3-dimensional (3D) audio cues are prone to fore-aft reversals and other localization errors, whereas fore-aft localization errors do not occur with vibrotaction. One consequence of multimodal integration, i.e. concurrent auditory and visual events, is that stimuli in one modality can influence our perception of stimuli in the other (Spence & Driver, 2000). For example, the ventriloquism effect refers to perceiving auditory sounds as coming from the same direction as the visually observed object, person, etc. As Spence and Driver (2000) point out, this effect is not limited to speech and lip-movements; it can occur for any concurrent visual and auditory cues. When examining crossmodal attentional issues, numerous studies have found that visual events "never" attract auditory attention (Spence & Driver, 2000). However, Spence and Driver found an exception to this rule, namely that spatial attention can, in fact, be drawn to the illusory location of a ventriloquized sound. In other words, visual events can attract auditory attention when paired with a concurrent unlocalizable sound. Furthermore, they found that ventriloquism, even for task-irrelevant stimuli, can happen swiftly and automatically with direct consequences for objective performance.

Importantly, this effect was only found with concurrent unlocalizable sounds; when the visual cues were paired with easily-localizable sounds, no effect was observed. Because 3D audio cues are subject to localization errors, pilots may fall victim to this effect and their attention may be inappropriately drawn to irrelevant stimuli. However, combining audio and tactile feedback may provide a means of avoiding the ventriloquism effect. The additional modality information presented by the combined audiotactile cues may equip pilots with better tools to combat the disorienting nature of degraded visual environments and high workload conditions. Audiotactile cues may be better at providing targeted, localizable cues to help pilots maintain SA.

Similarly, improving vibrotactile cueing displays may enable pilots' ability to more accurately perceive aircraft motion and/or approaching obstacles during flight. For example, Amemiya, Hirota, and Ikei (2013) found that by varying the speed of front-to-back tactile array stimuli subjects would increase or decrease their estimates of their illusory vection. Tactile arrays may also be used to convey approaching obstacles. Varying the rate and location of the tactile cues may be an effective means of conveying distance from an approaching object. Ultimately, effectively synthesizing the various types of information vibrotactile cues can provide could be very beneficial in helping pilots maintain SA.

Multisensory technologies may provide pilots with the widest array of tools to help maintain SA and avoid becoming spatially disoriented. One may argue that no single technological countermeasure may be effective at eliminating SD entirely; however, if all technological countermeasures were to effectively provide pilots with quick and accurate feedback to help maintain SA, then perhaps, in combination, they might be able to eliminate SD or at the very least, drastically minimize the likelihood of its occurrence.

Disclaimer

This report is solely the opinion of the authors and does not reflect official opinions or policies of the U.S. Government nor any part thereof. Use of any trade names does not imply endorsement of products by the U.S. Government nor any part thereof. Mention of any persons or agencies does not imply their endorsement of this report.

References

Albery, W. B. (2007). Multisensory cueing for enhancing orientation information during flight. *Aviation, space, and environmental medicine*, *78*(5s), B186-B190.

Amemiya, T., Hirota, K, & Ikei, Y. (2013). In *Virtual Reality (VR), 2013 IEEE*. (pp. 141-142). Lake Buena Vista, FL: IEEE Virtual Reality Conference.

Benson, A.J. (2006). Spatial disorientation in flight. In Ernsting's Aviation Medicine, 4th edition, Gradwell & Rainford (eds). CRC Press.

Braithwaite, M. G., Hudgens, J. J., Estrada, A., & Alvarez, E. A. (1998). An evaluation of the British Army spatial disorientation sortie in US Army aviation. *Aviation, space, and environmental medicine*, *69*(8), 727-732.

Guckenberger, D. & Bryan, E. (2003). N02-171 Spatial Awareness Training System (SPATS): Phase I Final Report. Orlando, FL: Naval Air Systems Command.

Heinle, T. E., & Ercoline, W. R. (2003). *Spatial disorientation: causes, consequences and countermeasures for the USAF*. Air Force Research Laboratory Wright-Patterson AFB OH Human Effectiveness Directorate.

Matthews, R. S. J., Previc, F. & Bunting, A. (2003). *USAF Spatial Disorientation Survey*. RTO meeting proceedings, RTO Human Factors and Medicine Panel (HFM) Symposium. 81-93.

McCauley, M. E., & Matsangas, P. (2004). *Human systems integration and automation issues in small unmanned aerial vehicles* (No. NPS-OR-04-008). Naval Postgraduate School Monterey CA Department of Operations Research.

McGrath, B. J., Estrada, A., Braithwaite, M. G., Raj, A. K., & Rupert, A. H. (2004). *Tactile Situation Awareness System Flight Demonstration* (No. USAARL-2004-10). US Army Aeromedical Research Laboratory, Fort Rucker AL.

Ostoin, S. D. (2007). *An Assessment of the Performance-Based Measurement Battery (PBMB), the Navy's Psychomotor Supplement to the Aviation Selection Test Battery (ASTB)*. Naval Postgraduate School Monterey CA.

Poisson III, R. J. (2014). *Spatial Disorientation: Past, Present, and Future* (No. AFIT-ENV-14-M-50). . Air Force Institute of Technology Wright-Patterson AFB OH Graduate School of Engineering and Management.

Spence, C., & Driver, J. (2000). Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *NeuroReport*, *11*(9), 2057-2061.

Weeks, J. L., Zelenski, W. E., & Carretta, T. R. (1996). Advances in USAF pilot selection. Advances in USAF pilot selection. *Advisory Group for Aerospace Research & Development: AGARD Conference Proceedings – 588. Selection and Training Advances in Aviation* (pp. 1.1-1.11). Linthicum Heights, MD: NASA Center for Aerospace Information.

Williams, H. P., Folga, R. V., Patterson, F. R., Arnold, R. D., & Horning, D. S. (2014, May). *Spatial Disorientation Countermeasures Training: Flight Simulation for Black Hole Illusion*. Aerospace Medical Association Conference, San Diego, CA.

Williams, H. P., Albert, A. O., & Blower, D. J. (November, 1999). *Selection of Officers for U.S. Naval Aviation Training*. Presented at the 41st Conference of the International Military Testing Association and NATO Officer Selection Workshop.

Zirkelbach, T. (2007). *Pictorial display design to enhance spatial awareness of operators in unmanned aviation*. Naval Postgraduate School Monterey CA.

RECENT ADVANCES IN TACTILE CUEING

Angus H. Rupert, U.S. Army Aeromedical Research Laboratory (USAARL), Fort Rucker, AL

Ben D. Lawson, USAARL, Fort Rucker, AL

Flight tests conducted by the Army and Navy have demonstrated the utility of the Tactile Situation Awareness System (TSAS) as an adjunct to visual instruments to improve pilot performance in degraded visual environments or under conditions of high workload. The tactile stimulators (tactors) used in each of the flight tests have been incorporated into aircraft components (seat cushions and shoulder straps) and a torso garment (belt or vest). Current tactors must operate at full magnitude and a very restricted frequency range (240 to 250 Hz) in order to provide consistent and perceptible stimuli in the aviation environment. Fortunately, recent developments in piezoceramics permit the frequency of tactors to vary from 50 to 500 Hz. This wider range of tactor stimulus frequency has significantly increased the available information content for tactile cueing systems. Several recent tests of TSAS will be presented to include additional capabilities that can be expected as piezoceramic tactors are incorporated into tactile designs.

Spatial disorientation (SD) occurs when a pilot does not correctly sense the position, motion or attitude of an aircraft relative the surface of the Earth. Since the U.S. Army established their consolidated database in 1972, SD has consistently accounted for at least 20% of U.S. Army class A/B mishaps (Flightfax, 2014). Mishap reports involving SD routinely attribute the cause of the mishap to the pilot with phrases such as: “The pilot failed to maintain an adequate crosscheck of the instruments” or “The pilot failed to respond in a timely manner.”

SD events and mishaps have occurred from the time pilots entered the aerospace environment. Between the Wright brothers’ first flight in 1903 and the 1929 introduction of orientation instruments to permit “blind flight,” pilots were unable to maintain spatial orientation awareness unless there was a clear view of the ground or horizon to provide orientation. However, SD mishaps continued even after planes were equipped with instruments to include the attitude indicator, heading indicator, turn rate indicator, altimeter, vertical velocity indicator, and airspeed indicator. Despite having these visual instruments to provide all of the necessary aircraft state parameters to be aware of orientation, pilots routinely became disoriented whenever they did not frequently refresh their knowledge of aircraft state parameters. How often pilots needed to refresh orientation was a function of aircraft stability and the dynamics of the flight regime. Only a few seconds of failure to scan the instruments could result in disorientation simply because visual instruments only provide orientation while the pilot is attending to the display. The pilot has more than just the task of “aviating” and must attend to other tasks including navigating and communicating. Anytime the pilot is not attending to the orientation instruments, the aircraft can slowly depart from controlled flight.

In instrument flight conditions and without the auto pilot engaged, the pilot is constantly making minor corrections to restore straight and level flight and to maintain the desired heading and altitude. The instrument scan requires a few seconds, so by the time the pilot has completed the scan and made minor corrections to any observed deviations, it is necessary to repeat the scan if the pilot is to maintain tight control of the aircraft. During any task or off-nominal condition that distracts the pilot’s attention away from the instruments (including boredom and fatigue), the aircraft frequently departs from the desired pitch, roll, or heading requiring the pilot to make significant corrections during the next scan. Strictly speaking, the pilot is disoriented many times during typical hand-flown instrument flights.

The TSAS was developed in response to the failure of visual modality instruments to provide pilots with continuous orientation information during flight. The concept of using tactile cueing as a means of intuitively maintaining spatial orientation for pilots was introduced at the 1989 Advisory Group for Aerospace Research and Development (AGARD) meeting on Situation Awareness in Aerospace Operations. The below diagram was used to explain the relation of the tactile stimulators (tactors) to the external environment.

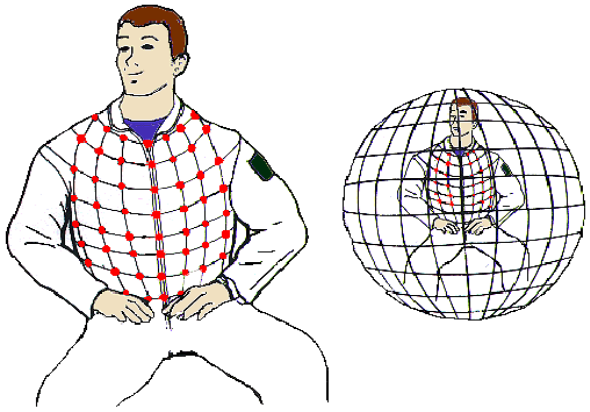


Figure 1. Columns and rows of tactile stimulators on the pilot mapped to the external environment (Rupert, 1993).

Since the pilot's torso is rigidly fixed to the aircraft via a multipoint harness, a matrix (columns and rows) of tactors incorporated into the pilots garment, harness, and seat can be mapped to the world surrounding the aircraft. Data from the aircraft orientation instruments provides the aircraft performance parameters to the pilot including the critical parameters of the direction down and the velocity vector. Most importantly since the matrix of tactors can provide pitch and roll information continuously to the pilot, it was no longer necessary to refer to the attitude indicator to maintain pitch and roll orientation information.

When pilots are flying in instrument meteorological conditions (IMC) more than 60% of their visual scan time is devoted to attending to two visual flight instruments, the directional gyro, and the attitude indicator (Simmons, Lees, and Kimbal, 1978). With the use of continuous non-visual displays, the pilot is now free to attend to other instruments that require visual attention or to other cockpit duties

Tactile displays have been proposed for use in aviation beginning as early as 1954. A few prototypes were attempted, but none met with success during in-flight trials as devices to maintain orientation. The two primary reasons early attempts at tactile cueing failed in the aviation environment were:

1. Non-intuitive displays: An early display using tactile cues on the hand to provide orientation information did not succeed since the pilot needed to devote significant attention to the tactile cues to interpret the information (Gilson and Fenton, 1974). It takes many years to become proficient at Braille, and many never succeed when it is necessary to learn this difficult task late in life. In contrast, minimal or no learning is required to present targeting information on the torso since the central nervous system is wired to reflexively interpret the location of taps on the torso and there is constant reinforcement of this experience during daily life events. This reflex is the example of a tap on the shoulder that draws attention to a point in space. The TSAS uses the same principle as a tap on the shoulder for targeting information. For pitch and roll information, TSAS provides the gravity vector in the same way as a person strapped firmly to a chair when the chair is moved in space in varying pitch and roll orientation. For this reason, minimal training is required to understand and use the system.

2. Inadequate tactile transducers: When the TSAS concept was first presented, the state of tactile transducers or tactile stimulators (tactors) was quite rudimentary. Early tactile stimulators were too large[1] and not salient enough to be appreciated in the noisy and high vibration environment of the cockpit. The first prototype transducer used for the TSAS proof of concept for torso displays were miniature speakers (Fig 2) derived from toys.

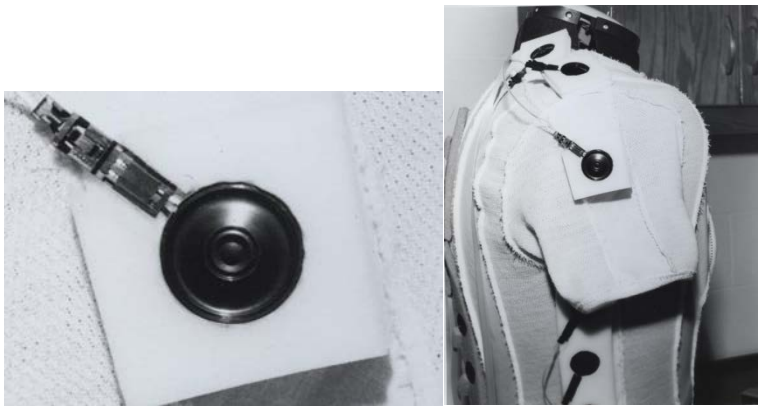


Figure 2. Miniature speaker on left and shown installed as a linear array on inside-out garment.

The fragile speakers frequently failed after short periods of use, and although they provided adequate saliency in the laboratory, they were not consistently perceived in the noisy and high vibration environment experienced in both fixed-wing aircraft and helicopters.

The second generation speakers for TSAS were custom built in-house vibrators consisting of a polyethylene block with an off-center rotating mass inside, similar to a pager motor but much larger. Again these tactors could not provide amplitude and frequency control of the stimulus. When funding was provided to support eight companies via the Small Business Innovative Research (SBIR) and Broad Agency Announcement (BAA) processes, it was possible to define the requirements for aviation tactile transducers. The requirements called for tactors that were: 1) small; 2) lightweight to permit easy integration into flight garments; 3) highly efficient to minimize power requirements and generate little heat as a by-product; 4) insensitive to contact pressure to enable tactors to be placed in seat cushions; 5) large dynamic range (50 to 500 Hz); 6) of a low failure rate to ensure reliability; 7) designed to provide minimal discomfort; 8) easily maintained by the military; 9) rugged for military environments; and 10) inexpensive.

Clearly, trade-offs were required to develop an acceptable tactor. The eight companies used different approaches to develop tactors based on varying principles. The best overall tactor was the C2 electromechanical tactor, developed by Engineering Acoustics Inc., that we have used for the past 15 years (Fig 3).

[1] An exception was direct electrical stimuli (Ross, 1973). However, the difference between a just perceptible electrical stimulus and a painful stimulus on dry skin was small and without using a paste or gel to control the interface, it was not possible to provide consistent stimuli without inducing pain.

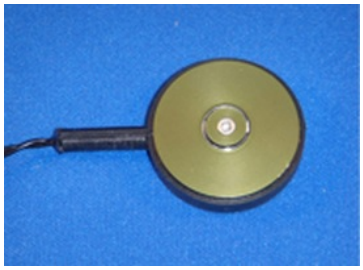


Figure 3. C-2 tactor manufactured by Engineering Acoustics Inc. The tactor is roughly the same diameter as a U.S. quarter coin.

The C-2 tactor provided consistent, highly salient stimuli for the aviation environment and was tuned for 240 Hz, which is peak sensitivity for skin vibration. The primary deficiency of the C-2 tactor was the narrow-frequency response which prevented tactile algorithm designers from using both frequency and amplitude modulation in the design of clearly interpretable tactile icons, also known as tactons.

In response to a DARPA-sponsored SBIR, the Midé Technology Corporation developed a piezoceramic tactor with a wide, dynamic response in frequency (50 to 500 Hz). When four experienced tactile researchers informally compared the C-2 electromechanical and the Midé ST-25b piezoceramic transducer, the ST-25b was felt to be more punctate and as good or better in terms of saliency.

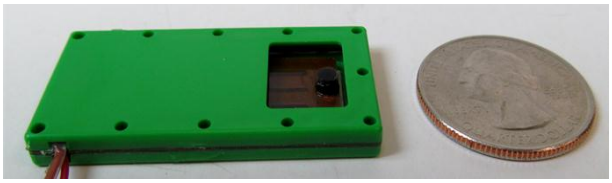


Figure 4. SHIVR™ ST-25b piezoceramic tactile stimulator (quarter for reference).

The importance of variable frequency tactors in the generation of rich tactons is demonstrated in Fig 5. By varying only the frequency and amplitude stimuli for the auditory sense, it is possible to produce variations in four different psychophysical dimensions, namely: pitch, loudness, volume, and density.

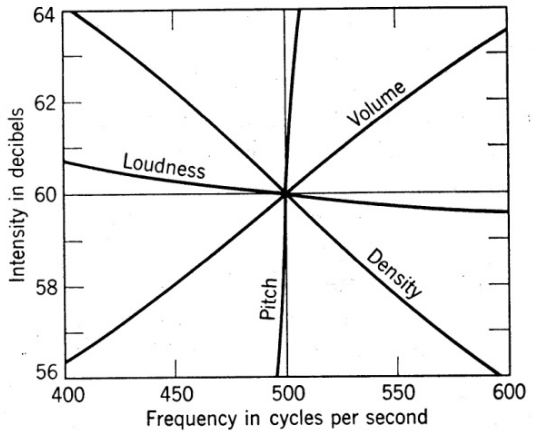


Figure 5. Isomorphic contours for pitch, loudness, volume, and density. Each contour defines the combinations of frequency and intensity at which a comparison tone will be perceived as equal in pitch or loudness or volume or density to the standard tone of 500 cps and 60 db. From Geldard (1953).

This concept was best expressed by Hans-Lukas Teuber (in Young, 1984) when he said, “The number of dimensions of perception exceeds that of the stimuli.” There are so many variables that tacton designers have to manipulate including magnitude/amplitude, frequency, waveform, pattern, duration, location, and interstimulus interval. For this reason, the range of tactile experiences is almost limitless.

When Georg von Békésy was conducting his Nobel prize-winning research on hearing mechanisms in the cochlea, he also conducted research on tactile sensation; reasoning that the inner ear is derived from the same embryologic ectoderm that produces skin receptors and so likely possesses similar mechanisms of sensory perception. He was correct but learned that the skin sensation was far more complex due to the variation in the number and types of sensory receptors and so returned to research the “trivial” system of the cochlea. This complexity of skin sensations can be used to advantage in developing rich tactile displays.

With the development of improved tactors, it will be possible to take advantage of tactile illusions that are inherent to the skin perceptual system. The “phantom sensation” (Von Békésy, 1957; Gescheider, 1965; Alles, 1970; Verrillo and Gescheider, 1975), occurs when two stimuli of equal loudness are presented at the same time to two nearby locations on the skin. The two stimuli are not felt separately but rather as a single stimulus halfway between the two stimulators. It is also possible to create the sensation of motion between two tactors by manipulating the relative intensities of two adjacent tactors. When one tactor intensity is increased while an adjacent tactor is decreased, the tactile sensation will be experienced as moving from one tactor to the other.

Another illusion providing the sensation of motion is the Rabbit illusion or cutaneous saltation (Cholewiak, 1976; Geldard, 1975). For example, a rapid sequence of 5 taps delivered first near the wrist, then halfway between the wrist and elbow, and then near the elbow, will be perceived as 15 equally spaced sequential taps “hopping” up the arm from the wrist towards the elbow and cannot be distinguished from 15 equally separated taps placed from the wrist to the forearm.

By using the Phantom and Rabbit illusions, it is possible to create “virtual tactors” located between the physical tactors which will reduce the number of tactors required in an aviation belt or garment. These illusions are a function of the separation of the tactors, the magnitude of the stimulus and the timing separation of the stimuli.

Recent tests using traditional C-2 tactors have demonstrated the capability of tactile cueing to maintain hover capabilities. Soon we will have even better capabilities with the recent development of piezoceramic tactors.

Acknowledgements

The authors would like to acknowledge DARPA for funding Midé Technology Corporation (via the Small Business Innovative Research program) to develop the piezoceramic tactor and the U.S. Army PEO Aviation for follow-on funding to permit DARPA to take the technology to the next level for aviation, robotics, and prosthesis applications. This report is solely the opinion of the authors and does not reflect official opinions or policies of the U.S. Government nor any part thereof. Use of any trade names does not imply endorsement of products by the U.S. Government nor any part thereof. Mention of any persons or agencies does not imply their endorsement of this report.

References

Alles, D. S. (1970). Information Transmission by Phantom Sensations. *Man-Machine Systems, IEEE Transactions on, MMS-11*(1), 84-91.

Cholewiak, R. W. (1976). Satiation in cutaneous saltation. *Sensory Processes, 1*, 163-175.

Flightfax. (2014). *In U.S. Army Combat Readiness Safety Center, 37*, 1-16. Retrieved from https://safety.army.mil/Portals/0/Documents/ON-DUTY/AVIATION/FLIGHTFAX/Standard/2014/May_2014_Flightfax.pdf.

Geldard, F. A. (1953). *The human senses* (1st ed.). New York: Wiley. In Stevens, S. S. (1934). The attributes of tones. *Proceedings of the National Academy of Science, 20*, 457-459. Washington DC.

Geldard, F. A. (1975). *Sensory saltation: Metastability in the perceptual world*. Hillsdale, NJ: Erlbaum.

Gescheider, G. A. (1965). Cutaneous sound Localization. *Journal of Experimental Psychology, 70*(6), 617.

Gilson, R. D., & Fenton, R. E. (1974). Kinesthetic-Tactual Information Presentations-Inflight Studies. *Systems, Man and Cybernetics, IEEE Transactions on , SMC-4*(6), 531, 535.

Ross, D., Sanneman, R., Levison, W. H., Tanner, R., & Triggs, T. J. (1973). Tactile display for aircraft control. Defense Technology Information Center Report No. AD767763. Nashua, NH: Sanders Associates, Inc.

Rupert, A. H., Mateczun, A. J., & Guedry, F. E. (1990). Maintaining spatial orientation awareness. *In Situational Awareness in aerospace operations, AGARD CP-478* (21-1-21-5). Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Rupert, A. H., Guedry, F. E., & Reshke, M. F. (1993). The use of a tactile interface to convey position and motion perceptions. *In Virtual interfaces: Research and applications, AGARD CP-541* (20-11-20-75). Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Simmons, R. R., Lees, M. A., Kimball, K. A. (1978). Visual performance/workload of helicopter pilots during instrument flight. *In Operational Helicopter Aviation Medicine, AGARD CP-255* (40-1-40-17). Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Verrillo, R. T., & Gescheider, G. A. (1975). Enhancement and summation in the perception of two successive vibrotactile stimuli. *Perception & Psychophysics*, *18*, 128-36.

Von Békésy, G. (1957). *The ear*. San Francisco, CA: W. H. Freeman.

Von Békésy, G. (1959). Similarities between hearing and skin sensations. *Psychological Review, 66*(1), 1-22.

Young, L. R. (1984). Perceptions of the body in space: mechanisms. *In Handbook of Physiology, The Nervous System, Sensory Processes, 3*(2).

AUDIOTACTILE AIDS FOR IMPROVING PILOT SITUATION AWARENESS

J. Christopher Brill, Old Dominion University, Norfolk, VA

Ben D. Lawson, U.S. Army Aeromedical Research Laboratory (USAARL), Ft. Rucker, AL

Angus H. Rupert, USAARL, Ft. Rucker, AL

Up to one-third of all aircraft mishaps are attributable to spatial disorientation (SD), costing lives and millions of dollars. One potential solution is to provide supplementary sensory cues to help improve pilots' situation awareness (SA). Given existing demands on the pilot's visual system, audition and touch present the greatest potential for success. However, accurate 3D audio perception may be problematic in noisy operational environments. To determine the effects, participants performed an azimuth cue localization task while listening to 90 dB helicopter noise. Cue modalities conditions included 3D audio, vibrotactile, and audiotactile. Accuracy was better and response times were significantly faster for tactile and audiotactile cues than for 3D audio cues alone. The results illustrate the deleterious effects of loud ambient noise on 3D audio localization and suggest audiotactile cues may offer a viable alternative non-visual display for counteracting SD.

The aviation environment poses numerous challenges for maintaining situation awareness, including spatial disorientation (SD). Up to one-third of all aircraft mishaps are attributable to SD (Gibb, Ercoline, & Scharff, 2011), costing lives and millions of dollars. One potential solution to the SD problem is to provide supplementary sensory cues to help improve a pilot's situation awareness. Given a pilot's vision is already taxed through scanning numerous displays, we sought to use non-visual cues for communicating information without increasing the burden on the information-saturated visual channel. The primary goal of the present investigation was to evaluate the effectiveness of non-visual directional cues for improving situation awareness (SA) under operationally relevant conditions, namely a noisy aircraft environment. Specifically, we sought to compare localization accuracy for three-dimensional (3D) audio, tactile, and audiotactile cues.

Literature Review of Auditory and Tactile Cues During Aviation

Investigators have recognized the utility of 3D audio as a novel way of displaying information to pilots. Examples include Begault's (1993) work on 3D audio traffic collision avoidance systems (TCAS), Brungart and Simpson's (2005) multi-talker spatial communication technology, Simpson and colleagues' work on 3D audio navigation and attitude indicators (Simpson, Brungart, Dallman, Joffrion, Presnar, & Gilkey, 2005), and recent communication systems from Garmin International, Inc. (Olathe, KS). Flight is inherently spatial, occurring in three dimensional axes (lateral, vertical, and longitudinal), and the most intuitive spatial orientation cues should reflect this, making 3D audio a promising candidate technology for a non-visual SD countermeasure system.

Despite its potential, outstanding issues currently limit widespread adoption of 3D audio for military aviation. Perhaps the most significant concerns include mixed data regarding accurate cue perception and reduced effectiveness in noisy operational environments. Binaural hearing can facilitate sound localization within ten degrees of accuracy in the horizontal plane (Senn, Kompis, Vischer, & Haeusler, 2005), and some researchers claim minimal audible angles (MAAs) of one degree (or less) are possible (Perrott & Saberi, 1990). However, these data were obtained in quiet laboratory environments using discrete sound sources (i.e., speakers) rather than headphone-based 3D audio systems. A more applicable depiction may come from Brill and Scerra (2014), who evaluated localization accuracy for eight discrete azimuth cues, each separated by 45°. Although the spatial separation between cues greatly exceeded previously published MAAs by at least 25° (see Brungart, Durlach, & Rabinowitz, 1999), localization accuracy only averaged 75%, primarily due to fore-aft reversals affecting three forward positions (0°, -45°, +45°). Brill and Scerra (2014) concluded the disparity between their results and previously published data were the result of leaving in common perceptual errors, namely fore-aft reversals, which are typically excluded from data analyses. They proposed the inclusion of fore-aft reversals was critical for accurate assessment of localization performance.

The second major concern about 3D audio in the cockpit is noise, particularly in military aircraft. In-cabin noise levels can reach 102 dB for a UH-60 "Blackhawk" helicopter, making it an extremely noisy environment. Even if 3D sound levels compensate for ambient noise, it is unclear exactly how it will affect sound localization. Good and Gilkey (1996) found that noise decreased localization accuracy for the frontal plane (i.e., fore-aft) the most. Lorenzi, Gatehouse, and Lever (1999) found similar results. They tested sound localization for high and low-frequency cues in the presence of noise and found decreased performance for the zero-degree position, irrespective of cue frequency. These results suggest frequency-based signal compensation may be of limited utility.

Tactile cues can be used as a potential alternative or supplement to 3D audio. Like hearing, touch is spatial by nature. Otherwise, we would not be able to find an itch, an insect crawling on us, or respond in the direction of a tap on the shoulder. Rupert's (2000) Tactile Situation Awareness System (TSAS) takes advantage of a tap-on-the-shoulder metaphor to provide pilots with spatial orientation cues, ranging from an Earth-centric vector (i.e., which way is down) to navigation and spatial alarms (e.g., TCAS or incoming missile). Others have developed systems based upon the TSAS theme, including Van Erp and colleagues (2006) and Rochlis (Rochlis & Newman, 2000).

Localization of torso-based tactile cues can be comparably better than 3D audio, although it greatly depends upon the circumstance. Localization accuracy for an 8-tactor circular array (i.e., belt around the torso) is 92-94% (Brill & Scerra, 2014; Cholewiak, Brill, & Schwab, 2004). However, accuracy drops with larger arrays (e.g., 74% for a 12-tactor circular array; Cholewiak et al., 2004). Whereas inaccuracies for 3D audio affect the fore and aft positions, tactile mislocalization occurs most frequently on the sides of the abdomen. Moreover, if a tactile cue is misperceived, it is typically by a single position, meaning the greatest possible error is 45° (compared to 180° for 3D audio). A comparison of localization performance for 3D audio and tactile cues reveals non-overlapping cones of confusion (Brill & Scerra, 2014). In essence, each modality's strength can potentially complement the other's weakness. Consequently, redundant bimodal cueing may yield greater performance than either modality alone.

Present Study

The present study sought to evaluate the relative effectiveness of spatial 3D audio, tactile, and combined "audiotactile" cues in the presence of noise. To improve external validity, we used a realistic operation noise stimulus: a recording of a UH-60 helicopter. It was predicted that localization accuracy and response times would be best for tactile and audiotactile cues; however, no specific predictions for relative differences between tactile and audiotactile cues were made.

Method

Participants

The experiment was reviewed and approved by Institutional Review Boards from the U.S. Army Medical Research and Materiel Command and Old Dominion University prior to participant recruitment. The participants provided written informed consent. A sample of eleven volunteers (10 males, 1 female, mean age = 31.5 years) was recruited from personnel at the U.S. Army Aeromedical Research Laboratory (USAARL) at Ft. Rucker. All had normal hearing, as confirmed by a hearing test, and normal sensorimotor functioning.

Research Design

The experiment used a within-groups design, wherein signal modality comprised each of the three conditions: auditory, tactile, and audiotactile. A within-groups design was adopted to control for potential individual differences in perception. Moreover, the nature of the experimental tasks (i.e., simple perceptual judgments) raised little concern regarding potential learning or carryover effects.

Apparatus

All signal presentation and data collection was performed using SuperLab 4.5.4 (Cedrus, Inc., San Pedro, CA) running on an MS-Windows-based laptop computer with an optical mouse. The software was setup to display on-screen instructions, present all experimental stimuli, and capture participant responses in milliseconds. The computer controlled the tactile display (see below) by sending serial strings via a USB cable. Ambient noise (90 dB)

was provided by playing a digital audio recording of a Sikorsky UH-60 "Blackhawk" helicopter in flight. The sound file was played on a desktop computer connected to a QSC model PLX-3602 stereo audio amplifier and Electro-Voice (EV) model T251+ speakers. The tactile display was comprised of an Engineering Acoustics, Inc. (EAI; Casselberry, FL) 8-tactor belt with model ATC3 controller. The belt was populated with model C2 tactors, speaker-like linear actuators for transmitting vibration into the skin. It was worn around the abdomen and secured with hook-and-loop material for a slightly snug fit for mechanically loading the tactors against the skin. The tactors were driven with a 250 Hz sinusoid at 90% power, which correlates to 51 dB, and activated for 500 ms. At this intensity, the vibrotactile pulses are easily detectable. The pulses are similar to vibration from a cell phone, but stronger and more focused. The tactile pulses were presented at eight discrete egocentric positions at 45-degree spacing: 0°(center ahead), 45°, 90° (right), 135°, 180° (aft), 225°, 270°(left), 315°.

The 3D audio display consisted of a laptop computer with a soundcard and Sennheiser HD-201 headphones for the playback of 3D spatialized digital audio files. The audio files consisted of 500 ms 150 Hz clicktrains that were processed using NASA SLAB Spatial Audio Renderer 5.8.1. They were rendered for eight azimuth positions. Each was equidistant from a central point so as to encircle the listener with discrete egocentric cues with the same 45-degree spacing as used for the tactile display. The loudness of the audio cues was calibrated using the method of adjustment in the presence of helicopter noise. A sample of five pilot participants were presented with an alternating pattern of 3D and vibrotactile signals with the task of adjusting the loudness of the 3D audio signals (using a volume knob) to match the subjective intensity of the vibrotactile signals. Each pilot participant performed the match eight times, and the average intensity value was calculated. Then, the grand mean was calculated and served as the intensity used for all study participants. Audiotactile signals were generated in a similar manner, through simultaneous (redundant) cueing via the 3D audio and tactile displays.

Procedure

Participants were welcomed to the laboratory and written informed consent was obtained. A hearing test was administered and participants were classified as having "normal" hearing if they met ANSI S3.19-1974 (ANSI, 2007). Participants were then led to a sound isolation booth and fitted with the tactile display. They were then seated in an ergonomic "kneeling" chair at a computer workstation and given an overview of the experimental task with exemplar stimuli. They were told they would be presented with a random series of audio, tactile, and audiotactile signals. They were asked to use the computer mouse to click a box on an on-screen graphic to indicate the perceived signal location. The graphic consisted of a top-down view of a human head encircled by eight boxes representing stimuli loci. Once the participant was ready to begin the experimental tasks, the experimenter left the room and began helicopter noise playback. Whenever a signal was presented, the participant would click a box to register a response. The timing of signal presentation was variable using a randomly selected inter-trial interval (2.5, 3.0, or 3.5 s) to prevent participants from getting into a response rhythm. Signals were grouped in three blocks, each comprising a modality condition. A block consisted of ten presentations of each of the eight stimulus locations, in random order, for a given sensory modality, resulting in 80 trials per block. The order of blocks was counterbalanced using a Latin Square design. After completing all three blocks, participants were thanked for their participation and dismissed from the study.

Results

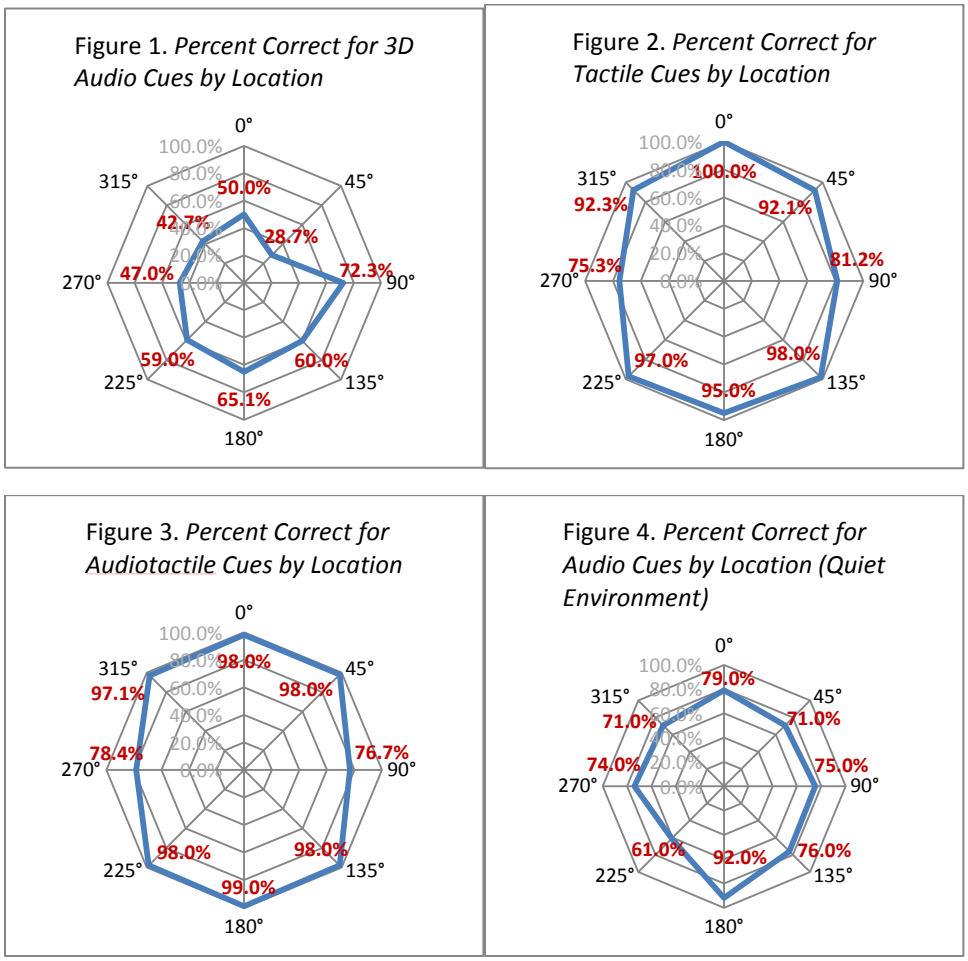
As the sample is relatively small, we chose a conservative approach to hypothesis testing by using non-parametric statistics with an alpha of .05. The data were screened for outliers, and one participant was removed for highly anomalous data, suggestive of equipment malfunction. This left a sample of ten participants for analyses. Raw performance data were coded and descriptive statistics were computed, including percent correct and mean response time by modality condition (see Table 1) and percent correct by stimulus position for each modality (Figures 1-3). To facilitate comparison, Figure 4 depicts 3D audio localization performance for the exact same cues when presented in a quiet environment (from Brill & Scerra, 2014).

Table 1.

Mean Cue Localization Accuracy and Response Time (in ms) for Azimuth Cues by Signal Modality.

Modality	Percent Correct	RT Correct Reponses	RT Errors
Audio	53.1% (16.3%)	1773 (89)	1560 (429)
Tactile	91.3% (6.3%)	1230 (645)	1867 (505)
Audiotactile	92.8% (5.3%)	1238 (170)	1546 (182)

Note: Standard deviation is in parentheses.



Figures 1-3. Percent correct localization under different cueing conditions (top-down view of eight tactors around torso).

A nonparametric test of differences (Wilcoxon Signed Rank Test for Related Samples) confirmed the hypothesis that localization accuracy (percent correct) was substantially better for tactile (M = 91.3) and audiotactile cues (M = 92.8) versus 3D audio cues (M = 53.1) ($ps < .01$). Likewise, mean response time for correct responses was significantly faster for tactile (M = 1230) and audiotactile (M = 1238) cues versus 3D audio cues (M = 1773) ($ps < .01$). However, no differences in response time were observed for errors ($p > .05$).

Discussion

The purpose of this investigation was to conduct a pilot study of the effects of operational helicopter noise on localization of discrete spatial audio, tactile, and redundant audiotactile cues in adult humans with normal hearing. Participants were, on average, 39% more accurate and 70% faster in responding to tactile or audiotactile cues versus 3D audio alone. This dramatic disparity does not just represent a statistically significant difference, but it is also a *meaningful* difference. Loud ambient noise had a dramatic effect on 3D audio performance, particularly for frontal positions, reminiscent of results by Good and Gilkey (1996) and Lorenzi et al. (1999). Nevertheless, our data exhibited an overall suppression of accuracy, whereas Lorenzi's frequency manipulation facilitated improved 3D audio performance for lateral positions. The signal we used was a 150 Hz clicktrain which was psychophysically calibrated for clear audibility above the ambient noise. However, we have yet to explore frequency-based manipulations to improve 3D audio performance. The data from this investigation will serve as a baseline for evaluating the effectiveness of frequency modulated 3D audio signals. To this end, signals representing lateral positions could be modified to contain more high frequency content to facilitate performance improvements.

Our data also suggest simultaneous vibrotactile cueing can effectively eliminate the spatial cueing inaccuracies common with 3D audio. For the human perceiver, the redundant tactile cue can help resolve ambiguity and uncertainty by providing a second piece of concordant information, particularly for positions for which perceptual reversals are likely. As in the real world, perception is rarely unimodal. We use our senses to collect multimodal information, and each sensory modality provides more information to assist perceptual guesses. The more information that is available, the less of a "guess" we make. It is the difference between seeing a friend from afar versus up close. You think the person is your friend, but it could be someone who resembles her. If she is at a distance and just standing, you can easily make an incorrect guess. In contrast, if she is a shorter distance away and you can observe her gait and hear her voice, the situation is much more data-rich to aid with identification.

Despite the impressive results of this pilot investigation, caution should be taken, as this is an ongoing investigation and more research will allow us to explore these questions further with larger sample sizes. At this stage, our results should be considered preliminary and subject to confirmation through further research. Current work is expanding the research presented here to include more azimuthal positions, elevation cues, and a second participant population: pilots with noise-induced hearing loss. Noise-induced hearing loss is a well-recognized and ongoing problem for military pilots. Pilots with hearing loss will likely have even worse 3D audio localization performance than those with normal hearing. We seek to evaluate if redundant tactile cueing is as effective for them as it is for a normal hearing population.

Given that performance for tactile cues was indistinguishable from audiotactile cues, one might suggest, why not simply use tactile cues as an alternative rather than a supplement? The answer is two-fold. First, incorporating audio and tactile cueing systems could offer a form of backup through redundancy in the event of system failure. Second, 3D audio is capable of presenting more than just noise bursts and tones. Earcons, the auditory equivalent of an icon, could be presented to aid with signal differentiation. Rather than using both touch and hearing for the equivalent of a tap-on-the-shoulder, the vibrotactile system could provide the tap, and the 3D earcon could convey target identity while also providing a concurrent directional cue.

To summarize, we presented 3D audio, tactile, and audiotactile azimuth directional cues to ten participants in the presence of 90 dB helicopter noise. Performance for tactile and audiotactile cues was significantly faster and more accurate than 3D audio cues alone. The data suggest tactile and audiotactile cues provide viable alternatives for counteracting spatial disorientation by improving pilot situation awareness.

Disclaimer

The views and opinions of authors expressed herein are those of the authors and do not necessarily state or reflect those of the United States Government or any agency thereof. This research was funded by U.S. Army Medical Research and Materiel Command. Dr. Angus Rupert is the principal investigator.

References

ANSI/ASA. (2007). *S12.68-2007 American National Standard Methods of Estimating Effective A-Weighted Sound Pressure Levels When Hearing Protectors are Worn.* Acoustical Society of America: Melville, NY.

Begault, D. R. (1993). Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation. *Human Factors*, *35*, 707-717.

Brill, J. C., & Scerra, V. E. (2014). Effectiveness of vibrotactile and spatial audio directional cues for USAF Pararescue Jumpers (PJs). *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics*, Kraków, Poland 19-23 July. Edited by T. Ahram, W. Karwowski and T. Marek.

Brungart, D. S., Durlach, N. I., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. II. Localization of a broadband source. *The Journal of the Acoustical Society of America*, *106*, 1956-1968.

Brungart, D.S.; Simpson, B.D. (2005) Improving Multitalker Speech Communication with Advanced Audio Displays. In *New Directions for Improving Audio Effectiveness* (pp. 30-1 – 30-18). Meeting Proceedings RTO-MP-HFM-123, Paper 30. Neuilly-sur-Seine, France: RTO.

Gasaway, D. C. (1986). Noise levels in cockpits of aircraft during normal cruise and considerations of auditory risk. *Aviation, Space, and Environmental Medicine*, *57*, 103-112.

Good, M. D., & Gilkey, R. H. (1996). Sound localization in noise: The effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America*, *99*, 1108-1117.

Lorenzi, C., Gatehouse, S., & Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *The Journal of the Acoustical Society of America*, *105*, 1810-1820.

Perrott, D. R., & Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, *87*, 1728-1731.

Rochlis, J. L., & Newman, D. J. (2000). A tactile display for international space station (ISS) extravehicular activity (EVA). *Aviation, Space, and Environmental Medicine*, *71*(6), 571-578.

Rupert, A. H. (2000). Tactile situation awareness system: Proprioceptive prostheses for sensory deficiencies. *Aviation, Space, and Environmental Medicine, 71*(9 suppl), A92-A99.

Senn, P., Kompis, M., Vischer, M., & Haeusler, R. (2005). Minimum audible angle, just noticeable interaural differences and speech intelligibility with bilateral cochlear implants using clinical speech processors. *Audiology and Neurotology*, *10*, 342-352.

Simpson, B. D., Brungart, D. S., Dallman, R. C., Joffrion, J., Presnar, M. D., & Gilkey, R. H. (2005, September). Spatial audio as a navigation aid and attitude indicator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 17, pp. 1602-1606). SAGE Publications.

Van Erp, J. B., Groen, E. L., Bos, J. E., & van Veen, H. A. (2006). A tactile cockpit instrument supports the control of self-motion during spatial disorientation. *Human Factors*, *48*, 219-228.

TACTILE CUEING STRATEGIES TO CONVEY AIRCRAFT MOTION OR WARN OF COLLISION

Ben D. Lawson, U.S. Army Aeromedical Research Laboratory (USAARL), Fort Rucker, AL

Roger C. Cholewiak, Cholewiak Consulting, Lawrence Township, NJ

J. Christopher Brill, Old Dominion University, Norfolk, VA

Angus H. Rupert, USAARL, Fort Rucker, AL

Linda-Brooke I. Thompson[1], USAARL, Fort Rucker, AL

This report highlights five current vibrotactile display technologies for conveying aircraft motion or approach to obstacles or waypoints, including: 1) a simple on/off vibration cue that activates when a designated position is reached; 2) vibrations whose on-off pulse rate increases as the vehicle moves faster; 3) vibrations whose fundamental frequency rises to cue the approach of an object; 4) vibrations whose body site of spatial cueing signals self-motion in a manner analogous to tactile cues during sliding along the ground; 5) vibrations whose body site of cueing expands in a manner analogous to visual looming cues. The advantages and limitations of these current approaches are discussed, a new display strategy is introduced, and recommendations are made concerning future tactile display development and research needed to provide tactile displays.

During flight, orientation and vehicle motion cues are often inadequate (e.g., due to loss of terrain visibility) or misleading (e.g., due to vestibular illusions). This report provides a brief overview of some of the currently-feasible tactile cueing strategies for restoring accurate perception of aircraft motion or the approach of obstacles or waypoints during flight. Tactile stick/rudder/collective tactile ("shaker") cues have long been used to prevent stalling or inappropriate power settings in flight. More sophisticated and informative tactile spatial orientation/motion technologies have been developed and tested which convey aircraft position and/or motion. These strategies include the following: Strategy 1: an applied vibration cue that simply turns on when one's aircraft reaches a designated waypoint; Strategy 2: an applied vibration display whose number of on-off pulses per second becomes greater as one moves faster away from a desired helicopter hover position[2]; Strategy 3: a laboratory tactile analogue of auditory perception whose vibration frequency rises to cue approach of/to a significant object (looming)[2]; Strategy 4: a laboratory tactile analogue of natural touch, whose site of spatial cueing on the body changes in a linear/laminar manner consistent with the perception of self-motion during sliding across the ground; Strategy 5: a laboratory analogue of vision, whose site of cueing expands radially to convey looming in a manner akin to the perception of visual flow. These currently-feasible cueing strategies are described briefly below, along with a new display strategy that builds upon Strategies 4 and 5 and which has been prototyped but not yet tested[2].

Strategy 1: Activating Vibration When the Target Point has been Reached

Strategy 1 employs vibration cues that simply activate when a target position has been reached or a significant event is imminent. Strategy 1 cues sometimes also convey information concerning the direction of the event, but they do not convey finely-graded cues indicative of the changing *distance* of the event relative to the vehicle operator, as he or she draws closer. For example, Van Erp and Van Veen (2004) tested a belt of eight electrically-driven vibrotactors (Figure 1) around the waist of a vehicle operator (one fast boat driver and one helicopter pilot). In this case, individual tactors activated to convey rich information concerning the direction of the desired waypoints during travel, but not their changing distance over time[3]. When the operator reached the desired waypoint, all eight tactors activated. This display aided the accuracy of wayfinding performance, despite the significant ambient vehicle vibrations present in the two vehicles studied by Van Erp and Van Veen (2004).

[1] This author is also affiliated with the Oak Ridge Institute for Science and Education.

[2] Strategies 2 and 3, and 6 were demonstrated to the audience attending this presentation.

[3] When the operator was pointing at the waypoint, the rate of vibration pulses in the front tactor would vary its frequency as well, but this was to refine directional information, not distance cues.

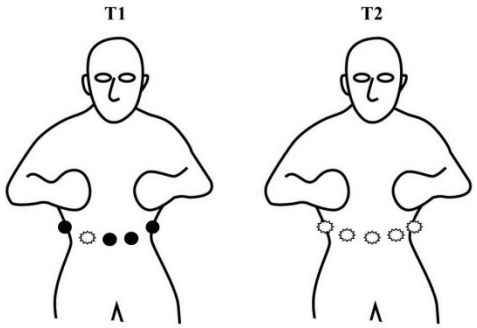


Figure 1. In this example of Strategy 1, a tactor activates (white star on man at left) to indicate the need to turn in that direction, while all tactors activate (see man at right) to indicate arrival at the waypoint. (Three additional tactors on the back are not shown.)

Strategy 2: Varying the Rate of On/Off Vibration Pulses to Indicate Self-Motion

A second tactile cueing strategy was employed by McGrath et al. (2004) during a helicopter flight demonstration (Figure 2). Two rows of eight tactors each went around the pilot's waist and supplied directional and velocity cues. When the pilots (n = 4) drifted outside their designated hover range, two vertically-oriented tactors both activated in the direction of the unwanted drift, much as a rumble strip on a freeway warns an automobile driver of lane deviations. In addition, relative cues concerning the pilots' rate of self-motion were provided by pulsing the on-off periods of vibration at a higher frequency per second as horizontal velocity of the helicopter increased. The pilots were found to have better flight control when tactile cues were available than when they were not.

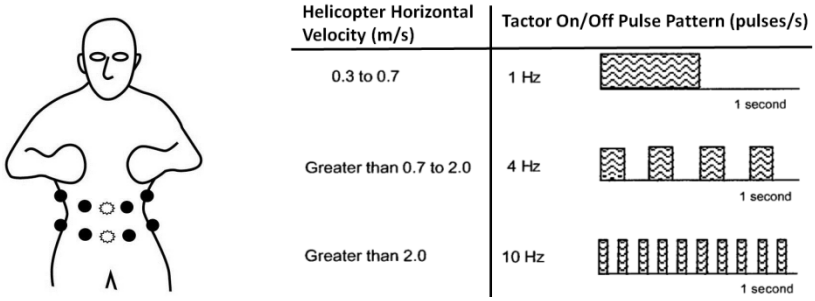


Figure 2. In this example of Strategy 2, when the helicopter drifts horizontally, two tactors (see white stars on man at left) vibrate on the side of the drift (in this case, forward). In addition, the temporal frequency of on-off vibration pulses corresponds to the velocity of the drift deviation (right, from McGrath et al., 2004).

Strategy 3: Varying the Fundamental Frequency of Vibration

Gray (2011) reported that an auditory vehicle collision warning that increases in sound intensity in a way analogous to a real sound source approaching the subject aids faster initiation of braking than other types of cues[4]. Since simple loudness and pitch changes can help to convey looming, we wondered whether a localized tactile stimulus of varying vibration frequency could convey information consistent with looming even when the spatial site of the tactile stimulus was held constant at a given body site. We sought to establish whether a varying vibration cue could be interpreted as a meaningful tactile icon or *tacton* consistent with an object that is approaching (Brewster and Brown, 2004). The relevant stimuli and findings are summarized briefly, below.

Various vibration stimuli were evaluated by thirty-five participants, using semantic differential ratings (Lawson et al., in press). The stimuli varied in terms of their tactor vibration frequency and duration of firing over the course of a three s train of twenty vibratory bursts (or beats). Two (of the six) vibration stimuli are shown in the graph in Figure 3, because these two stimuli are relevant to display Strategies 2 and 3 discussed in this report. Our study

[4] The single exception was the sound of a car horn, which unfortunately also produced a greater likelihood of false positive braking responses.

indicated that Strategy 2 (varying the rate of on/off pulses) conveyed looming less well than Strategy 3 (varying the fundamental frequency of vibration). In fact, the highest scoring condition for looming was "increasing frequency" (Strategy 3), which was interpreted differently from a random control stimulus and all other stimulus patterns evaluated, including Strategy 2 (Bonferroni-adjusted pairwise comparisons, $p < 0.01$). These findings imply that frequency (Strategy 3) may be a better aspect of the stimulus to manipulate than rate of on/off pulses (Strategy 2), at least when the purpose is to convey the concept of an object moving towards or away from the observer.

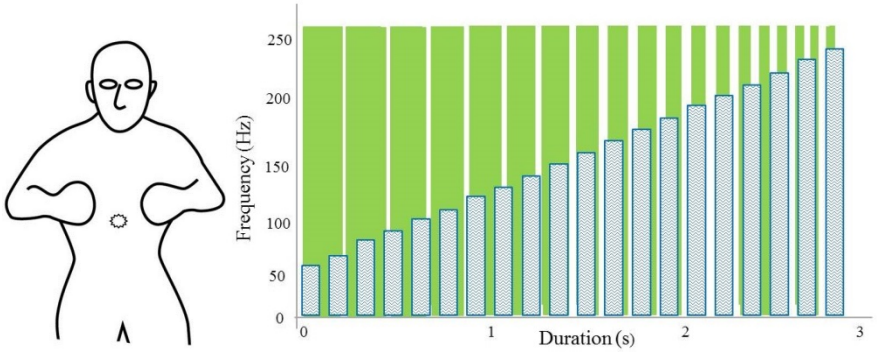


Figure 3. In this example of Strategies 2 and 3, looming is conveyed by vibrations at a given body site (see man at left) that either pulse more frequently (on/off) over time (Strategy 2, shown as green/shaded bars in background of the graph) or increase their fundamental vibration frequency over time (Strategy 3, see pattern-filled bars in the foreground). Strategy 3 was rated a better looming tacton (Lawson et al., in press).

Strategy 4: Varying the Site of Body Cueing in a Laminar or Linear Manner

Amemiya, Hirota, and Ikei (2013) presented subjects (n =7) with simulated optical flow cues (radial expansion of ~1,000 random dots on a twenty inch monitor) for forward self-motion. The found that the subjects' estimates of the speed of their illusory forward self-motion (vection) could be reduced or increased by varying the speed of a front-to-back tactile flow stimulus (i.e., by varying the inter-stimulus interval of tactor activation in a 4 x 5 array of tactors, each of which consisted of nineteen vibrating pins) across the seat of subjects' pants. This is a very natural way to convey self-motion tactually (Lawson, 2014) and an approach that could be explored for the cueing of pilots.

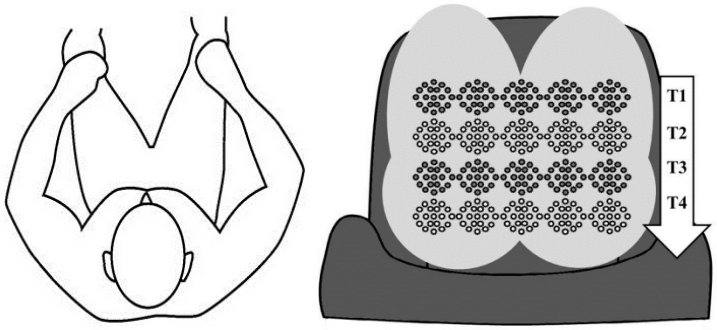


Figure 4. In this example of Strategy 4, a top-down view of a seated operator is shown at left, with the light gray shading at right approximating where the operator's posterior contacts the seat. Four rows activate at Times (T) 1, 2, 3, and 4, to provide a sweeping cue from front to back of the operator's posterior.

Strategy 5: Varying the Site of Body Cueing in a Radial Manner

In a study by Cancar et al. (2013)[5], tactile flow cues concerning the approach of a real ball[6] were sufficient (without visual cues) to enable participants (n = 12) to hit the ball at the correct time in 71% of the trials. The expanding tactile flow field Cancar et al. employed (Figure 5, partially adapted from Cancar et al.) has two possible advantages over a simple on/off vibratory "looming warning cue" (such as the vibrating mode of a cell phone). First, the expanding flow field is a logical analogue of an approaching optical target or three-dimensional auditory target (Lawson, 2014). Second, the perceived magnitude of the stimulus will increase as more tactors are activated (Cholewiak, 1979), thus increasing saliency. This approach also is worth exploring further.

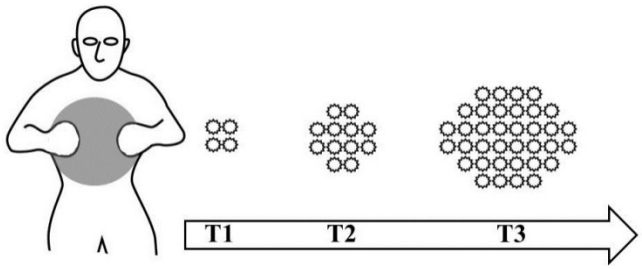


Figure 5. In this example of Strategy 5, the approach of an object is cued at three times (T1-T3) by the number of activated tactors in an expanding array on the chest (gray region).

Conclusions, Developing Strategies, and Future Recommendations

Table 1 summarizes and compares the five currently-available tactile display strategies discussed in this paper. Strategy #1 (indicating when the pilot has reached a waypoint) is simple, but cannot provide a graded warning that would be useful in a wide variety of situations. A variant of strategy #2 is employed in some current tactile cueing systems intended for aviation, but a recent study implies that Strategy #3 may be better for conveying looming. Strategy #3 may be better than Strategy 2 for conveying imminent collision. Strategies #4 and 5 are rich, but require further research evidence. The first issue that should be studied during flight is whether Strategy #3 works better than the currently-used Strategy #2. The second issue that should be studied is whether tactor arrays (Strategies #4 and 5) are superior to single-point displays (Strategies #1-3).

Table 1.

Comparison of Five Available Tactile Cueing Strategies for Conveying Aircraft Motion.

	Strategy #1	Strategy #2	Strategy #3	Strategy #4	Strategy #5
Application	Simple event warning	Rate of closure	Rate of closure	Rate of drift	Rate of closure
Simplicity:	Simple	Intermediate	Intermediate	Complex	Complex
Maturity:	High	High	Medium	Low	Low
Richness:	Low	Medium	Medium	High	High
Pros:	Low cost, time, weight, size	Tested in-flight	Optimal tacton	Intuitive	Compatible with visual displays
Cons:	Inadequate for many applications	Non-optimal tacton	Limited stimulus range	Limited testing	Limited testing

Each display strategy above has certain advantages but none will be exploited optimally unless a clear understanding of tactile display design principles for conveying aircraft motion or collision is obtained first (Lawson, 2014). In the long run, we recommend that strategies 2, 3, 4, and 5 should not continue to develop separately, but rather, coalesce into a suite providing multiple cueing capabilities within one display system. At first, such a display system may represent a smorgasbord of insufficiently integrated cues. With each round of development and testing,

[5] See also Jansson (1983).
[6] The ball was tethered on a string and swinging towards the subject.

the display should become more similar to the natural somatosensory (and ultimately, multisensory) cues that specify self-motion and imminent collisions. In addition, a wider body surface should be exploited in future displays. None of the displays presented in this report stimulated the feet or lower legs, for example. This should be corrected, since these body regions are important to the somatosensory appreciation of natural body orientation, motion, and balance, and the feet have a large representation in the brain homunculus.

Different aviation groups hold differing and deeply-entrenched preferences concerning whether currently-available visual, auditory, or tactile displays are the best choice for avoiding mishaps by conveying aircraft attitude, altitude, motion, and proximity to obstacles. Nevertheless, few experts in perception or display design would argue that the wisest choice for conveying vehicle motion to an operator is to provide a single, abstract cue that is unlike the intuitive, multisensory cues one receives during motion through the natural world. We recommend that the developers of visual, auditory, and somatosensory displays place the needs of pilots ahead of their own interests in seeing their particular unimodal display solution adopted. Instead, display experts should coordinate a multisensory display solution that feels most natural to the user and best exploits the natural functions of each sensory modality (Lawson, 2014). The goal of multisensory display systems for vehicle control, virtual environments, and teleoperation should be to provide visual, auditory, vestibular, and somatosensory cues that are similar to the cues available during real motion in the natural world. We have developed a naturalistic tactile seat display that is shown in Figure 6. The display is a working proof-of-concept that requires testing to determine whether it aids situation awareness and vehicle control. In the prototype shown, Strategies 4 (varying the site of cueing linearly) and 5 (varying the site of cueing in a radial manner) are both incorporated into one display. Seven rows of tactors fire successively (at T1-T7) to convey forward drift (left side of figure 6), while the array fires in a radially-expanding pattern (right side of Figure 6) to convey descent towards the ground (at four successive time periods, T1-T4). As with Strategy 5 discussed above, descent towards the ground triggers more tactors as one gets closer to the ground, leading to increased saliency of the critical ground collision hazard (Cholewiak, 1979).

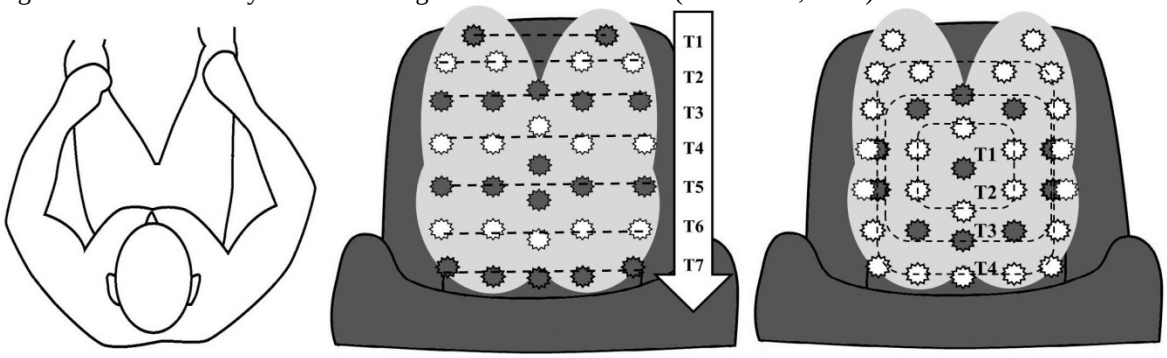


Figure 6. A top-down view of a prototype seat display that can cue horizontal drift (middle) or change in altitude (right) using the same tactor array. Light gray shading approximates where the operator's posterior contacts the seat.

A more sophisticated concept for a future tactile display is shown in Figure 7. The seated pilot is shown partially from below, to emphasize the stimulation of cutaneous receptors naturally associated with self-motion relative to the substrate. The display would incorporate large regions of skin and would intuitively convey body motion in six degrees of freedom. For clarity, Figure 7 only depicts two degrees of freedom, viz., linear translation up, down, for, or aft. Many types of cues are possible, including activating rows of vibrators linearly (to convey self-motion), radially (to convey imminent collision), and on distinct body parts successively (e.g., feet, then wrists, then chest, to convey approach to an obstacle or waypoint) (Meng et al., 2014).

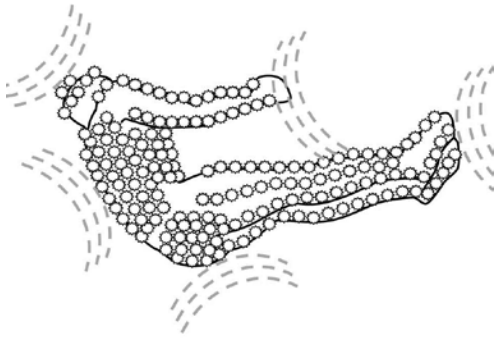


Figure 7. A tactile display concept to convey motion or approach to objects in multiple directions.

Current tactile displays are primitive compared to visual or auditory displays. However, tactile displays are rapidly growing more sophisticated, just as occurred for visual and auditory displays. Many of those who dismissed the need for a visual display of orientation in the 1920s lived to see visual attitude indicators become indispensable. Similarly, those who dismiss the usefulness of augmenting the perception of vehicle state with a tactile display will soon become routine users of high-resolution tactile displays during (real and simulated) air, sea, and ground travel.

Acknowledgements

We thank C. Harris, B. Mortimer, and G. Mort for expert support of some of the technologies described. This report is solely the opinion of the authors and does not reflect official opinions or policies of the U.S. Government nor any part thereof. Use of any trade names does not imply endorsement of products by the U.S. Government nor any part thereof. Mention of any persons or agencies does not imply their endorsement of this report.

References

Amemiya, T., Hirota, K, & Ikei, Y. (2013). In *Virtual Reality (VR), 2013 IEEE.* (pp. 141-142). Lake Buena Vista, FL: IEEE Virtual Reality Conference.

Brewster, S. & Brown, L. M. (2004). Tactons: structured tactile messages for non-visual information display. *Proceedings, 5th conference on Australian user interface, 28,* 15-23. Glasgow, UK: Australian Computer Society.

Cancar, L., Díaz, A., Barrientos, A. Travieso, D. & Jacobs, D. M. (2013). Tactile-sight: A sensory substitution device based on distance-related vibrotactile flow. *Internat. Journal of Advanced Robotic Systems, 10 (272).*

Cholewiak, R. W. (1979). Spatial factors in the perceived intensity of vibrotactile patterns. *Sensory Processes, 3,* 141-156.

Gray, R. (2011). Looming auditory collision warnings for driving. In *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(1),* 63-74.

Jansson, G. (1983). Tactile guidance of movement. *International Journal of Neuroscience, 19,* 37-46.

Lawson, B. D. (2014). Tactile displays for cueing self-motion and looming: What would Gibson Think? In Ahram, T., Karwowski, W, & Marek, T (Eds.), *Proceedings, 5th International Conference on Applied Human Factors and Ergonomics* (pp. 928-38).

Lawson, B. D., Cholewiak, R., McGee, H., Mortimer, B., & Rupert, A. (in press). Conveying looming with a localized tactile cue. Fort Rucker, AL: U. S. Army Aeromedical Research Laboratory Tech Report.

McGrath, B. J., Estrada, A., Braithwaite, M. G., Raj, A. K, & Rupert, A. H. (2004). Tactile Situation Awareness System flight demonstration final report. Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory Tech Report, Report No. 2004-10.

Meng, F, Gray, R, Ho, C., Mujthaba, A., & Spence, C. (2014). Dynamic vibrotactile signals for forward collision avoidance warning systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society.*

Van Erp, J. B. F. & Van Veen, H. A. H. C. (2004). Vibrotactile in-vehicle navigation system. *Transportation Research Part F, 7,* 247-256.

REQUIREMENTS FOR DEVELOPING THE MODEL OF SPATIAL ORIENTATION INTO AN APPLIED COCKPIT WARNING SYSTEM

Ben D. Lawson, U.S. Army Aeromedical Research Laboratory (USAARL), Fort Rucker, AL

Braden J. McGrath, University of Canberra, Griffith, Australia

Michael C. Newman, National Aerospace Training and Research Center (NASTAR), Southampton PA

Angus H. Rupert, USAARL, Fort Rucker, AL

Refinements have been made to a model of spatial disorientation (SD) to improve simulation of acceleration stimuli and visual-vestibular interactions. The improved model has been applied to aviation mishaps. The model is considered a technological countermeasure for SD because it is implemented as prototype software to aid the identification of mishap contributors in a way that should benefit didactic training. There is a more direct way this countermeasure can prevent SD, which is by adapting it for use as part of a cockpit warning system. The idea is to expand the model from one which explains mishaps post-hoc into one that warns pilots proactively whenever they are most likely to experience SD. This report introduces the general requirements that must be met to successfully expand the model for use as a proactive cockpit warning system, as well as the key criteria for determining whether it is effective.

As part of a recent U.S. Army Medical Research and Materiel Command effort, considerable refinements were made to the existing mathematical model of spatial disorientation (SD), to the point where it can analyze laboratory and in-flight acceleration stimuli and mishaps that were not feasible to analyze previously, can simulate the integration of visual-vestibular inputs better than before, and can compare the predictions from multiple theoretical models (Newman et al., 2012; McGrath, in Lawson et al., 2014; McGrath et al., 2015). These improvements constitute a technological countermeasure for SD in the sense that they have been implemented as software to improve our understanding of the specific, quantitative, and proximal factors that cause mishaps. This can have positive repercussions for aviation training. However, there is a much more direct way that this type of technological countermeasure can prevent SD, which is by developing it for use as a cockpit warning system. The idea is to adapt the same model presently used to determine whether the moment-by-moment vestibular and visual cues present prior to a flight mishap could have caused a specific disorientation illusion. This model would be refined to, instead, proactively warn pilots whenever their current vestibular and visual cues in-flight are likely to cause a specific disorientation illusion. This report introduces the general requirements such a system would need to meet and the features it must have in order to be successful. The four main requirements that must be met in order to expand the current model into an accurate cockpit display or warning system are as follows: 1) model outputs to the display must incorporate inputs concerning the state of the human user; 2) model outputs to the display must incorporate inputs concerning the state of the system within which the user must operate (i.e., the aircraft and its surrounding environment); 3) the interface (displays and controls) must be user-friendly to pilots; and 4) the usefulness of the new display must be verified during flight testing. These general requirements are elaborated in Table 1, which lists the critical and the desirable inputs and improvements needed to develop the current quantitative vestibular orientation model into a real-time, in-cockpit display or warning system.

Table 1.

Key requirements or aspects of a cockpit display based on the orientation-model.

	1: Knowing Pilot State	2: Knowing System State	3: Usability	4: Efficacy
Critical Aspects	1A. Model-display system must "know" the pilot's control inputs	2A. System must know state of the aircraft (attitude, acceleration, etc.)	Pilot-friendly user interface	Efficacy verified in-flight[1]
	1B. System must know whether pilot is looking at primary flight displays	2B. Must know environmental visibility		
Desirable Aspects	1C. Know whether pilot is cognitively attending to instruments	2C. Know flight instructions or clearances		
	1D. Know pilot's head position and motion			

Critical Pilot State Requirement #1A: Knowing the Pilot's Control Inputs

The current model is designed to explain proximal, perceptual causes of SD mishaps after the event has occurred by determining when a pilot is likely to have been disoriented and what his or her misperception was (e.g., direction of felt pitch and/or roll) during the disorientation episode. This is accomplished by comparing the quantitative predictions the model makes concerning the pilot's perceived orientation (during each second preceding a mishap) to the pilot's actual orientation (derived from the outputs of the flight data recorder), and then determining whether there is a mismatch between the two, and if so, whether the pilot's joystick control inputs at the time of comparison are consistent with the mismatch, which would imply the presence of a perceptual illusion[2]. For example, when a pilot takes off from the ground with sufficient forward acceleration and enters a cloud layer, he may experience an illusion of being pitched backward more than is actually the case (Figure 1). This is due to the well-known *somatogravic illusion*, in which the pilot perceives the direction of "down" to be dependent not solely upon the direction of gravity, but rather, the direction of the resultant between gravity and the aircraft's forward acceleration, known as the gravitoinertial force. This gravitoinertial force gives the pilot the illusion of still being pitched back when the aircraft already has leveled off; thus causing the pilot to unnecessarily move the stick forward to level off, which can result in an unrecognized dive towards the ground. Many mishaps have been attributed to the somatogravic illusion, and variants of the illusion can occur during prolonged banking turns and other maneuvers (Newman et al., 2012; McGrath, in Lawson et al., 2014; McGrath et al., 2015). The occurrence and severity of the illusion can be accurately modeled with the current orientation model based on the way that acceleration inputs to the aircraft would be processed by the vestibular system, but the model user then has to compare the model's outputs manually to the pilot's control input data to determine if there was a discrepancy.

If the present orientation model is to be used to drive a real-time cockpit display, it must receive stick inputs from the pilot and compare these automatically to the model's moment-by-moment prediction of the pilot's perceived orientation. In the somatogravic example provided in Figure 1, the model should know when the pilot is pushing the stick forward in a manner that accords with the illusion of not being leveled off yet, but does not accord with reality of already being leveled off. Fortunately, joystick inputs by the pilot are already available as data the model can acquire. This is not so for every aspect of the pilot's state that the model needs to know, however, as described below.

[1] The criteria for deciding if the display is effective are discussed later in this report and summarized in Table 2.
[2] This approach assumes the pilot has not intentionally flown the aircraft into the ground, which can usually be verified by the voice recording.

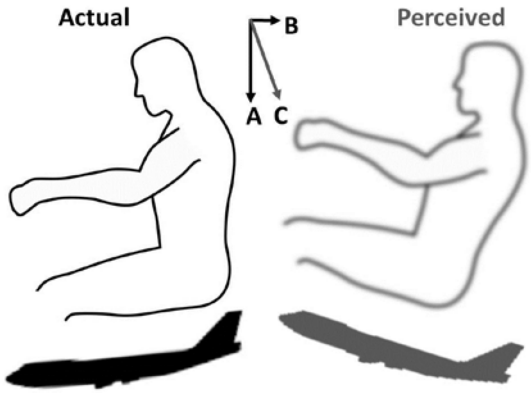


Figure 1. In this example of a somatogravic illusion, Earth's gravitational force (A) combines with the force caused by forward acceleration during take-off (B) to yield a resultant force (C) that makes the pilot feel that "down" is behind him, causing him to perceive pitch backwards (gray, right) rather than his actual orientation (black, left).

Critical Requirement #1B: Knowing if the Pilot is Looking at the Instruments

The orientation model assumes that SD usually occurs when the pilot is unable to see the outside world and has failed to continually maintain an accurate mental model of aircraft orientation by referencing the flight instruments. During mishap investigations using the model, it often must be assumed that the pilot was not attending to the instruments because the pilot's actual direction of gaze is not known. Inferences are made based on the phase of flight, the atmospheric conditions, and the voice record (e.g., indications that attempts at visual flight were made when a switch to instrument flight was warranted, indications that workload, stress, and distraction were high due to aircraft troubleshooting, navigation, or communication problems). It would be better for mishap investigation if the pilot's direction of gaze was known, and this information would aid the development of a cockpit display based on the orientation model. Direction of gaze can be inferred readily via videooculography (Stephane, 2012, in Boy. Ed.). Some aircraft systems track the pilot's head orientation, but no current systems track gaze. Since direction of gaze can vary considerably relative to head orientation, gaze data is needed to improve post-mishap reconstruction and to permit the model to become a perceived-versus-actual orientation display and/or a disorientation warning system. Suitable technology exists to fulfill this requirement and software parameters could be adjusted readily.

Desirable Aspect #1C: Knowing if the Pilot's Cognitive Attention is on the Instruments

Most of the time, tracking the pilot's gaze and knowing if it is dwelling frequently upon the primary flight displays will be sufficient to know that proper instrument flight is being maintained. Under certain circumstances, it is possible for the pilot to fix his gaze upon the instruments without cognitively attending to and processing the symbolic information provided (Mack, 2003). This can occur when a drowsy pilot stares in the direction of the instruments without cognitively processing the information provided. It can also occur when an overworked or highly stressed pilot looks at the instruments habitually without sufficient engagement of attention and working memory. Every reader will recall a time when he or she was driving an automobile and looked at a vehicle display, but then had to look again to cognitively register the information. While such cognitive lapses are not frequent, it is nevertheless true that a disorientation cockpit warning system would be more accurate if it knew not only whether the pilot was looking regularly at the instruments, but also whether the pilot was in a cognitive state that prevented the information from being processed. This goal might be achieved by collecting physiological data, e.g., to distinguish when brain activity during visual fixation is consistent with paying attention to the display or, rather, indicative of gross under- or over-arousal that generally degrades attention, or a more specific case of inattentional blindness (Turatto et al., 2002; Mack, 2003). Similarly, additional parameters of gaze could be monitored (such as saccadic velocity, fixation dwell time between saccades, number of saccades, and pupil diameter) to determine the pilot's state of arousal. Such instrumentation requires its own cycle of development and validation, however. Since arousal state information is not fully mature for cockpit adoption yet and gaze information should be sufficient, arousal state information is considered desirable in the future but not critical presently.

Desirable Aspect #1D: Knowing the Pilot's Head Position

It has long been known that making head movements during banking turns in flight is disorienting. This was initially thought to be due to the vestibular effects of simultaneous multi-axis head rotation (known as Coriolis cross-coupling), but it is now believed that the disorienting effect derives mostly from the unusually large amplitude and velocity of the otolith organs during high G head movements (known as the G-excess effect, Rupert & Guedry, 1991). The current model can predict Coriolis cross-coupling and G-excess effects, but to do so, it must know the pilot's head position and motion. Once the two critical pilot state requirements of the new display are met (#1A-1B), this readily-added capability should be incorporated. Current technology is suitable for this purpose.

Critical System State Requirement #2A: Knowing the Position and Motion of the Aircraft

To make predictions about the perceived orientation and motion of the pilot, the user inputs flight parameter data (attitude, altitude, acceleration, etc.) into the model. The model then processes these data according to the known functioning (e.g., time constants) and limitations of the vestibular organs in order to compute actual versus perceived orientation of the pilot and aircraft. The needed data is already available from many aircraft. The model is already equipped to upload and process such data, but it presently does not do so in real-time. Software parameters can be modified readily for this purpose.

Critical Requirement #2B: Knowing whether the Outside World is Visible

Currently, the model user must infer from independent information (e.g., the draft mishap report) whether visibility outside the aircraft was compromised and whether the pilot was not processing symbolic orientation information from the primary flight instruments. Common contributors to SD include unaided night flight or flight into a degraded visual environment, especially when there is high workload or distraction and the gravitoinertial force vector does not match the "down" given by gravity. For the model to become a disorientation warning system, it should know in real-time whether the pilot is not likely to be able to see the outside world sufficiently to fly under visual flight rules. This could be accomplished via the US Army Aeromedical Research Laboratory's airborne visibility indicator (Estrada et al., 2004) that could be utilized to send automated updates to the model concerning flight visibility. Model software parameters could be adjusted readily when the airborne indicator indicates that outside visual information is unreliable.

Desirable Aspect #2C: Knowing the Flight Instructions and Clearances

In the simple somatogravic example explained in Figure 1, it would be advantageous if the model knew that the pilot's intention was to fly straight-and-level once he or she had reached a certain altitude. This would help the model confirm that pilot's control inputs and aircraft motions were not only consistent with the aforementioned backward pitch illusion, but also in violation of the pilot's plan to level-off. Other aspects of the flight clearance instructions that could be input into the model include the maximum bank angle during a turn, the minimum and maximum altitude during the flight phase, etc. These inputs are listed as desirable future information for the display rather than information that is critical currently. This is because spatial disorientation usually can be inferred without this information and because certain aspects of the flight instructions and clearances for certain military missions may be too complicated or flexible to permit easy designation and incorporation into the model. Such information is readily available during many military transport operations and commercial civil airlines flights, however.

Requirement #3: A Pilot-Friendly Display

The model incorporates a convenient graphical user interface (GUI) that allows the user to select model parameters and see animations of actual versus perceived (predicted) orientation during the time leading into a mishap. However, the GUI and the virtual control buttons for altering model parameters are designed for the use of spatial orientation specialists working in an office environment. Modifications will be necessary for in-cockpit use by pilots. First, the display should conform with general vehicle display principles (Berson et al., 1981; Mejdal et al., 2001; Wickens & Carswell, 2006). Second, it should be tailored to pilots rather than orientation experts.

Requirement #4: An Effective Display

Once all the necessary elements for a model-based disorientation display are assembled and integrated, it is critical to test them in flight to confirm that the display is effective. Before such tests are done, the criteria for an effective display must be established. A few key criteria are shown in Table 2. Further detailing and prioritization of these criteria will yield a correct decision concerning whether the display is useful in maintaining situation awareness and aircraft control, or instead requires further modification.

Table 2.

Criteria for verifying the efficacy of an orientation- model-based cockpit display.

Operational Criteria	Human Factors Criteria	Scientific Criteria
Identifies the most hazardous and common types of SD	Provides salient information	Yields few false positive warnings (false threat warnings)
Prevents full entry into SD during disorientating flight maneuvers	Provides clear information	Yields almost no false negatives (failures to warn of real threats)
Permits rapid recovery after being placed into an unusual attitude		

The first operational goal mentioned in Table 2 is to develop the orientation model into an applied cockpit warning system that will be able to predict the most hazardous and prevalent SD illusions. Some key types of SD illusions with a vestibular component that are mentioned in prominent aeromedical textbooks (Rainford & Gradwell, Eds., 2006; DeHart & Davis, 2002) and key reviews (Previc & Ercoline, 2004) include: the leans, undetected (helicopter) drift, false/sloping horizon, tumbling/vertigo, graveyard spiral/spin, somatogravic illusion, inversion illusion, somatogyral illusion, elevator illusion, G-excess illusion, and vection. This list is not exhaustive, but covers most of the vestibular or visual-vestibular illusions most commonly noted by pilots. It is difficult to know which illusions are most deadly. Among these illusions, some tend to be mentioned less frequently as mishap contributors while others are frequently identified as factors in deadly mishaps. To the extent that more-frequent mention in class A mishap reports reflects how dangerous a given SD illusion is, we can conjecture that illusions such as the somatogravic illusion and undetected helicopter drift are particularly dangerous and should be included in the model-based cockpit warning system. Similarly, we infer that unrecognized (type 1) SD is inherently more dangerous than recognized SD, which would mean that tumbling/vertigo may not have to be included in the cockpit display if it cannot be modeled adequately. Currently, the model can predict over a dozen orientation illusions. Among these, we are most confident in the model's ability to predict in-flight cases of somatogravic illusion, inversion illusion, undetected drift, visual illusions that involve a vection component (e.g., induced by dust blowing during helicopter landing), and the occurrence of the leans. The model will require modifications to accurately predict variations in the duration of the leans. Once head-movement data are available, the model will be able to determine the intensity of head movement-contingent illusions such as the dynamic G-excess illusion or Coriolis cross-coupling (Rupert & Guedry, 1991). These and other modifications are being worked on as part of the Program Executive Office Aviation's Small Business Innovative Research efforts.

Possible Characteristics and Applications of the Display

The envisioned cockpit warning system would take in data from a variety of onboard sources that monitor the motion of the pilot and aircraft (e.g., attitude and heading reference system, accelerometers, gyroscopes, and head and eye position). Data from each source would be converted to the correct sensory coordinate frame and used as inputs to drive the orientation model. The model would process the sensory data in real-time and output a continuous prediction of the pilot's estimated orientation and perceived self motion. This prediction could be used for three related, in-cockpit uses. First, its outputs could augment visual displays by showing the pilot's perceived direction of "down" vs. the actual direction of down (e.g., via a simplification of Figure 1). Second, it could drive an auditory warning system that would identify when the pilot is entering a flight condition that is likely to induce SD and alert the pilot (e.g., in the situation depicted in Figure 1, the warning might say "possible illusion of backward pitch: check instruments!"). Third, in cases where continuous information is not already being provided to the pilot via a tactile situation awareness system, the disorientation display could trigger a strong vibrotactile cue on the body

providing the actual orientation of down. For example, in Figure 1, a seat vibration could be triggered under both thighs and well forward of the misleading pressure cues the pilot is getting on his buttock and lower back.

Acknowledgements

We thank Ms. Linda-Brooke Thompson for assistance with the preparation of this manuscript. This report is solely the opinion of the authors and does not reflect official opinions or policies of the U.S. Government nor any part thereof. Use of any trade names does not imply endorsement of products by the U.S. Government nor any part thereof. Mention of any persons or agencies does not imply their endorsement of this report.

References

Berson, B. L., Po-Chedley, D. A., Boucek, G. P., Hanson, D. C., & Leffler, M. F. (1981). Aircraft alerting systems standardization study, Volume II: Aircraft alerting system design guidelines. U.S. Department of Transportation (DOT) Tech. Report No. D6-49976TN. Washington, DC: Federal Aviation Administration.

DeHart, R., & Davis, J. (Eds). (2002). Fundamentals of aerospace medicine (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Estrada, A., LeDuc, P., Persson, J., Greig, J., Crowley, J., & van de Pol, C. (2004). A proof of concept of an airborne visibility indicator. Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory Tech. Report No. 2004-15.

Gradwell, D., & Rainford, J. (Eds.). (2006). *Ernsting's aviation medicine* (4th ed.). New York, NY: Oxford University Press Inc.

Guedry, F., & Rupert, A. (1991). Steady state and transient G-excess effects. *Aviation, Space, and Environmental Medicine, 62(3),* 252-253.

Mack, A. (2003). Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science, 12(5),* 180-184.

McGrath, B., Newman, M., Lawson, B., & Rupert, A. (2013). An algorithm to improve ground-based spatial disorientation training. In *Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Modeling and Simulation Technologies Conference.* Reston, VA: AIAA.

McGrath, B. (2014). Visualization of spatial disorientation mishaps in the U.S. Navy: Case study. In Lawson, B., Rupert, A., Raj, A., Parker, J., & Greskovich, C. Invited lectures from a spatial orientation symposium in honor of Frederick Guedry, Day 1. U.S. Army Aeromedical Research Laboratory Tech. Report No. 2014-10.

Mejdal, S., McCauley, M. E., & Beringer, D. B. (2001). Human factors design guidelines for multifunction displays. U. S. DOT Tech. Report No. DOT/FAA/AM-01/17. Washington, DC: Office of Aerospace Medicine.

Newman, M.C., Lawson, B.D., Rupert, A.H., & McGrath, B.J. (2012). The role of perceptual modeling in the understanding of spatial disorientation during flight and ground-based simulator training. In *Proceedings of the AIAA Modeling and Simulation of Technologies Conference*, 14 pages. Minneapolis, MN: AIAA.

Previc, F., & Ercoline, W. (2004). *Spatial disorientation in aviation (Progress in astronautics and aeronautics)* (1st ed.). Reston, VA: AIAA.

Stephane, A. L. (2012). Eye tracking from a Human Factors perspective. In Boy, G. A. (Ed), *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach*, 339-365.

Turatto, M., Angrilli, A., Mazza, V., Umilta, C., & Driver, J. (2002). Looking without seeing the background change: Electrophysiological correlates of change detection versus change blindness. *Cognition, 84(1),* B1-10.

Wickens, C. D., & Carswell, C. M. (2006). Information processing. *Handbook of human factors and ergonomics, 3,* 111-149.

BEHIND THE SCENES OF THE NAS: HUMAN FACTORS TAXONOMY FOR INVESTIGATING SERVICE INTEGRITY EVENTS

Katherine A. Berry, Fort Hill Group, LLC, Washington, DC
Michael W. Sawyer, LLC, Washington, DC
Jordan Hinson, LLC, Washington, DC

The Federal Aviation Administration (FAA) deployed the Service Integrity Risk Analysis Process (SI-RAP) with the goal of assessing the risk of technical occurrence events where the ability to provide safe air traffic management technical services is compromised. As a post-event tool, SI-RAP assesses the risk associated with an occurrence based on severity and repeatability. The SI-RAP taxonomy was developed to provide a consistent framework for supporting the assessment of event repeatability. The SI-RAP taxonomy synthesizes existing human factors taxonomies with customized factors representing the technical operations domain. The SI-RAP taxonomy is comprised of four tiers: Personnel Factors, Contextual Factors, Equipment Factors, and Systemic Factors—with each tier being composed of categories that group related taxonomy factors. An iPad application was developed to assist SI-RAP panel members in the application of the taxonomy. This paper will introduce the SI-RAP taxonomy, the SI-RAP walkthrough application, and will describe the future application of the taxonomy.

The FAA deployed the SI-RAP in October 2014. Building upon the FAA's (2013) Airborne Risk Analysis Process (Airborne RAP) and EUROCONTROL's (2013) Risk Analysis Tool (RAT), SI-RAP's primary goal is to assess the risk of technical occurrence events when the ability to provide safe air traffic management services is compromised (Berry, Sawyer, & Hinson, 2014). As a post-event analysis tool, SI-RAP assesses the risk associated with an occurrence based on severity and repeatability, with the repeatability portion incorporating a taxonomy of occurrence factors. Furthermore, the SI-RAP taxonomy (Figure 1) incorporates the human factors areas from the Air Traffic Analysis and Classification System (AirTracs) taxonomy (Berry & Sawyer, 2014).

The development of a process to examine service integrity occurrences allows for occurrences to be thoroughly and methodically examined over time. SI-RAP is applied by a panel comprised of technical operations subject matter experts (SMEs) and air traffic control (ATC) SMEs. As part of the SI-RAP process, the panel members will examine the repeatability of the occurrence to classify factors and determine the repeatability of a similar occurrence happening. The purpose of this study was to develop a standard process for assessing the repeatability of these events including the development of the SI-RAP taxonomy. The SI-RAP taxonomy will allow for factors to be identified, classified, compared, and monitored over time and across multiple occurrences. The following sections will introduce the SI-RAP taxonomy along with the associated training and tools that support the application of the taxonomy

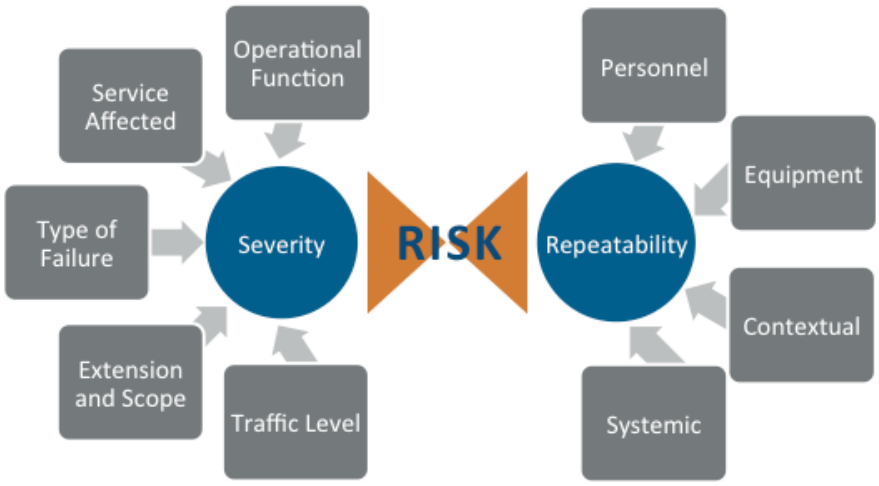


Figure 1. Components of SI-RAP

Introducing the SI-RAP Taxonomy

The SI-RAP taxonomy was developed through a process of taxonomy review, SME opinion elicitation, domain customization, and test case application. The SI-RAP taxonomy follows the structure of the RAT and the Airborne RAP taxonomies and tailors the factors to the domain-specific needs of technical occurrences. Furthermore, the SI-RAP taxonomy incorporates the human factors areas from the AirTracs taxonomy. The SI-RAP taxonomy is comprised of four tiers: Personnel Factors, Contextual Factors, Equipment Factors, and Systemic Factors. Each tier is composed of categories that group related taxonomy factors. The SI-RAP taxonomy is displayed in Figure 2 and Table 1.

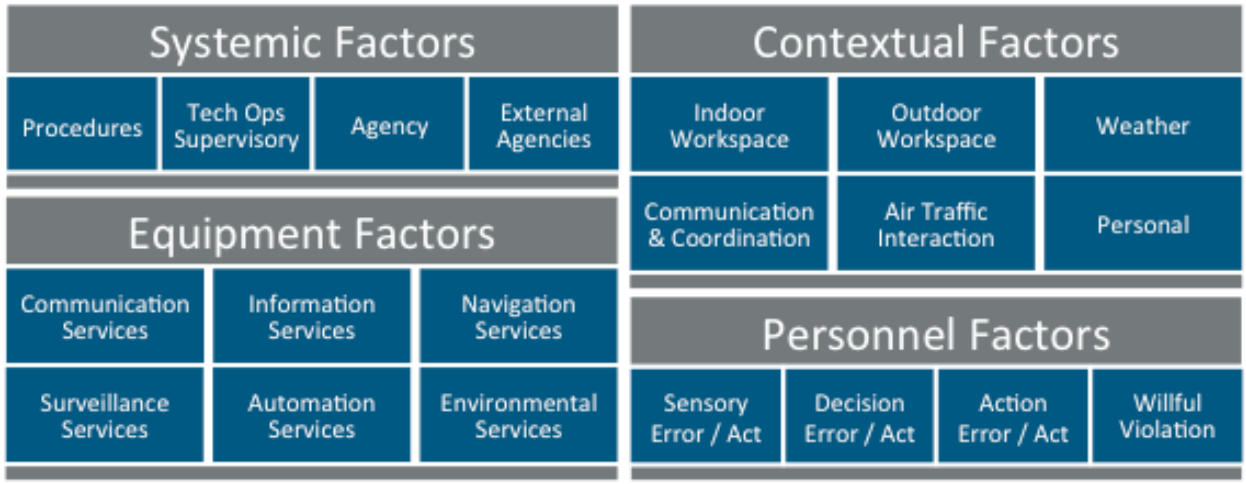


Figure 2. SI-RAP Taxonomy

Table 1. *SI-RAP Taxonomy and Factors*

Systemic Factors
Procedures Factors: Relates to the procedures, checklists, and data an ATSS must use to operate or conduct work. Factors: 6000.15, Maintenance Handbook Procedures, Technical Performance Record, Task Reference Glossary File, Facility Reference Data, Remote Monitoring and Logging System, Checklist, Standard Operating Procedures
Technical Operations Supervisory: Relates to the roles and responsibilities of Technical Operations management and supervisors at local facilities. Factors: Technician Equipment/Tool Readiness, Staffing/Personnel Scheduling, Scheduling of Equipment Outages, Oversight/Assistance, Training Resources and Availability
Agency Factors: Relates to the roles and responsibilities of Technical Operations Agency management and other Technical Operations. Factors: Facility Callback, Safety Culture, Policy, Agency Oversight, Agency Response to Occurrence
External Agency Factors: Relates to how the roles and responsibilities of external, non-FAA actors and organizations. Factors: Contractor Provided Service, Airlines, Contract Towers, Flight Service Stations, Military, Airport Authority, Other ANSPs
Equipment Factors
Communication Services: Relates to the systems, subsystems, or equipment used to transmit or receive voice or data intelligence. Factors: Air/Ground Communication - Main Radio Frequency, Air/Ground Communication - Secondary Radio Frequency, Air/Ground Communication - Backup or Emergency, NAS Voice Switch, Ground Communication –NRCS, Ground Communication – Shout Line/Indirect Access, FAA Provided Telecommunications (Telco), FTI Telco
Information Services: Relates to the systems, subsystems, or equipment used to provide meteorological information and data. Factors: Airport Weather Services - ATIS/ASOS/AWOS, Wind Equipment, Terminal Weather Services, Weather/Radar Processors, National Airspace Data Interchange Network
Navigation Services: Relates to the systems, subsystems, or equipment used to provide guidance, navigational data, or information accomplished either visually or electronically. Factors: VOR, DME, and TACR Systems, ILS and NDB Systems, Lighting - PAPI and VASI
Surveillance Services: Relates to the systems, subsystems, or equipment used for real-time detection and/or display of airborne or ground positional information for ATC. Factors: Primary Air Surveillance, Secondary Air Surveillance (Beacon), Surface Surveillance, ADS-B, Radar
Automation Services: Relates to the computerized systems, subsystems, or equipment used to provide complex automated processing of data elements used in the NAS. Automation uses hardware, software, and various data type inputs, such as communication, weather, surveillance, navigation, infrastructure, and flight information, to provide a composite NAS product. Factors: Terminal Radar Data Processing – ARTS/STARS, En Route Radar Data Processing – HOST/ERAM, Oceanic Radar Data Processing, Surface Movement Guidance and Control, Flight Data Processing, Automated Flight Service Station and FSS Systems, Traffic Management/Flow Systems, SWIM
Environment Services: Relates to the environmental and power systems, subsystems, equipment, or facilities used to support, house, or protect NAS systems, subsystems, and equipment. Factors: HVAC, Commercial Power, Critical Power Distribution System/Uninterruptible Power Supply, E/G, Fire Alarm System, Building Monitor and Control System, Access Control
Contextual Factors
Indoor Workspace: Relates to how the indoor environment, workspace, and tools in which an ATSS or other individual must operate or conduct work. Factors: Distraction – Duty Related, Distraction – Non-Duty Related, Lighting/Vision Restricted, Noise, Ergonomics, Slippery Surface, ATSS Equipment, Site Accessibility, Wildlife, Vandalism
Outdoor Workspace: Relates to how the outdoor environment, workspace, and tools in which an ATSS or other individual must operate or conduct work. Factors: Distraction – Duty Related, Distraction – Non-Duty Related, Lighting/Vision Restricted, Noise, Ergonomics, Slippery Surface, ATSS Equipment, Site Accessibility, Wildlife, Vandalism

Weather: Relates to how weather or meteorological factors can impact an ATSS, other individual, or equipment.
Factors: Fire, Flood, Fog, Glare, Ice, Rain, Snow, Temperature – High, Temperature – Low, Thunderstorm/Lightning, Visibility, Winds, Frost/Ground Heave

Communication & Coordination: Relates to the teamwork factors that are part of successful execution of maintaining the air traffic service integrity. Factors relate to the communication and coordination of planning maintenance, executing maintenance, and returning equipment to service.
Factors: Document/Record in Logs or RMLS, Misspeak/Mishear Information, Equipment Outage Reporting/Status, NOTAM Annotation/Location, Responsiveness, Supervisory Coordination

Air Traffic Interaction: Relates to the actions or inactions by the Air Traffic community (controllers, traffic managers, etc.) that directly impacted an occurrence.
Factors: Controller Misuse of Automation/Equipment, ATC Awareness of Maintenance Event, ATC Interrupts Maintenance, ATC Maintenance Moratorium, ATC Reporting of Events

Personal Factors: Relate to how an individual is impacted by internal stressors or demands.
Factors: On-the-Job Training Being Conducted, Unfamiliar Task/Procedure, Workload – High/Complex, Workload – Low/Underload, Complacency/Vigilance, Automation Reliance, Pattern Assumption/Habits, Time Pressure, Fatigue – Mental, Fatigue – Physical/Muscle, Attitude/Mood

Personnel Factors

Sensory Error/Act: Relates to a person detecting, identifying, and interpreting information through his or her senses. Sensory errors occur when a person's sensory input is degraded and a decision is made based upon faulty information.
Factors: Inspect, Monitor/Observe

Decision Error/Act: Relates to a person developing and determining a plan or response. A decision error occurs when a person's behaviors or actions proceed as intended, but the plan proves to be inadequate and results in, or contributes to, an occurrence.
Factors: Troubleshoot/Diagnose, Coordinate/Describe, Certify/Verify, Prioritization

Action Error/Act: Relates to a person executing a plan, performing a task, implementing a decision, or implementing a course of action. An Action Error/Act occurs when an individual's execution of a routine, highly practiced task relating to procedures, training, or proficiency result in an occurrence.
Factors: Modify, Align/Calibrate, Install/Upgrade, Reset/Configure, Replace/Install, Measure/Test

Willful Violation: Relates to a person willingly and knowingly deviating from rules, regulations, procedures, or policies. This factor should be classified when there is a willful violation relating to a person deliberately disregarding established rules and procedures.
Factors: Willful Violation, Situation Induced Violation

Additionally, when identifying the causal factors, the SI-RAP panel determines the classification level of each factor (Table 2). Panel members classify the factor levels as either causal, contributory, observed, or positive (Berry & Sawyer, 2014).

Table 2. *Factor Classification Levels*

Classification		Factor Definition
Adverse	Causal	An immediate/direct factor that identifies an active error or failure of critical components of equipment, systems, or human error. *Causative: If "A" occurs, then "B" will occur.*
	Contributory	An underlying/root factor that identifies latent errors or failures related to human performance, operating environment, task procedures, training, supervision, or policy that influence the presence of causal factors. *Probabilistic: If "A" occurs, then the probability of "B" occurring increases.*
Neutral	Observed	A factor that is present but the associated impact of the factor on the safety event has not been proven. It is recorded to note its potential influence on the event or actors involved and to be incorporated into trend analysis.
Beneficial	Positive	A factor that positively contributed to the safety of an event. This can include factors or actions that contributed to the detection of, or recovery from, an adverse outcome.

SI-RAP Taxonomy Application

As an accompaniment to the SI-RAP taxonomy, an iPad application was developed to assist the SI-RAP panel in the application of the SI-RAP taxonomy. When a user accesses the SI-RAP taxonomy website, the user must first request a user account and initially set up the account. After the account is approved, the SI-RAP user can access the SI-RAP application and view the homepage (as seen in Figure 3).

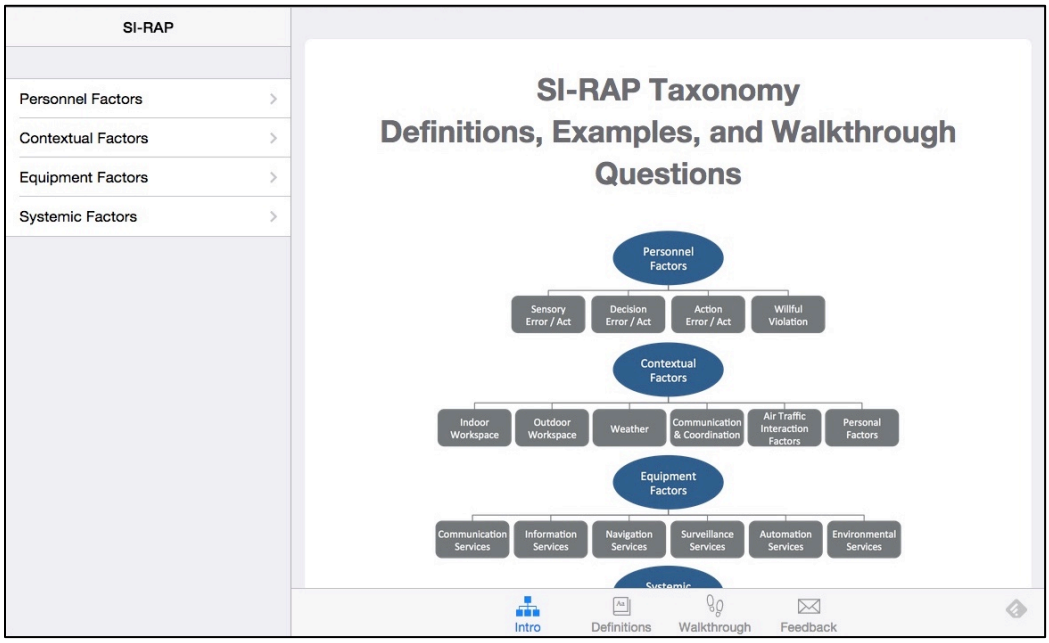


Figure 3: SI-RAP iPad Application – Homepage

From the homepage, the SI-RAP user can access definitions for the various tiers, categories, and factors. The SI-RAP user can also access the example application of each factor as well. In addition to the definitions and examples, the SI-RAP application presents the SI-RAP user with a series of questions that help users to determine which factor to select. These walkthrough questions guide the user to appropriate factors through a series of yes / no questions (Figure 4) and multiple-choice questions (Figure 5).

Figure 4: SI-RAP iPad Application – Walkthrough Question Example 1

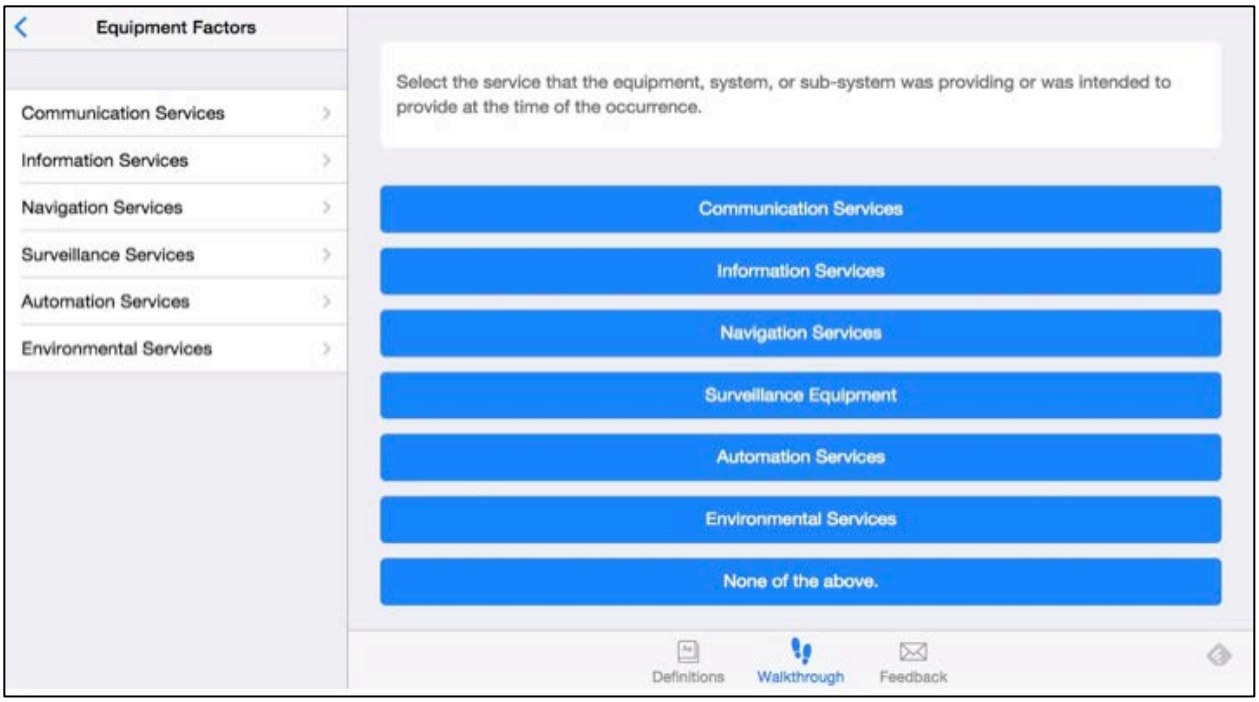


Figure 5: SI-RAP iPad Application – Walkthrough Question Example 2

Acknowledgements

We would like to acknowledge the FAA's Human Factors Research and Engineering Division (ANG-C1) for funding this project and similar work. Additionally, we would like to acknowledge the technical operations and human factors subject matter experts who provided the valuable insight necessary to develop these results. The results presented herein represent the results of this research project and do not necessarily represent the views of the FAA.

References

Berry, K. & Sawyer, M. (2014). *Air Traffic Analysis and Classification System (AirTracs): Human Factors – Safety Taxonomy Definition and Description.* Retrieved from https://www.hf.faa.gov/hfportalnew/admin/FAAAJP61/AirTracs%20Working%20Description_FAA%20Cover.pdf

Berry, K., Sawyer, M., & Hinson, J. (2014). *Incorporating Human Factors into the Service Integrity Risk Analysis Process: A SI-RAP Taxonomy and Training Program*. Retrieved from https://www.hf.faa.gov/hfportalnew/admin/FAAAJP61/Incorporating%20Human%20Factors%20into%20the%20Service%20Integrity%20Risk%20Analysis%20Process%20-%20A%20SI-RAP%20Taxonomy%20and%20Training%20Program.pdf

EUROCONTROL. (2013). *Risk Analysis Tool: Guidance Material*. Retrieved from http://www.skybrary.aero/bookshelf/books/2193.pdf

FAA. (2013). *Air Traffic Organization 2013 Safety Report*. Retrieved from http://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/safety/media/ato_2013_safety_report.pdf

A COMPREHENSIVE EFFORT TO ARRIVE AT AN OPTIMALLY RELIABLE HUMAN FACTORS TAXONOMY

Raymond E. King, Headquarters, Air Force Safety Center, Human Factors Division, Kirtland Air Force Base, NM

Department of Defense (DoD) members sought to improve the inter-rater reliability of the DoD Human Factors Analysis and Classification System, (DoDHFACS). DoDHFACS differs from the original system developed by Wiegmann and Shappell (2003), based on the work of Reason (1990), by further analyzing mishaps and hazards to a more granular level – arriving at specific "nanocodes." The steps involved in the effort included determining which of the 147 "nanocodes" were rarely/never used and collapsing nanocodes and rewriting definitions to arrive at 109 nanocodes. Next, a stepwise checklist to guide investigators through consideration of nanocodes was created. Student investigators were guided to continually test checklists and generate results to gauge inter-rater reliability. They were asked to offer constructive criticism to hone checklist questions. While inter-rater reliability results are encouraging (Fleiss' Kappa of .847 at the broadest level), additional work is necessary to realize the goal of an optimally reliable human factors taxonomy.

As will be demonstrated, "Human factors" are causal or contributory in a majority of military aviation mishaps. This paper reports on a Department of Defense (DoD) effort to improve the system used to categorize causal and contributing human factors. Specifically, recent attempts to improve coding methods with the goal of achieving better inter-rater reliability and ultimately more actionable recommendations to improve safety will be described.

Hartmann (1977), in a widely read and highly regard article, asserted that reliability is a necessary but not a sufficient basis for validity. Hartmann went on to specify that there are two methods that can be employed to determine reliability: percentage agreement reliability and reliability coefficient. Hartmann advocated for the latter over the former as percentage agreement may produce inflated estimates of reliability. Another issue to bear in mind when considering reliability is that categories must be mutually exclusive and exhaustive (that is, contain no overlapping elements and be complete) to achieve the highest reliability. Overlapping elements may result in observers using different categories for the same observation and thus finding fewer distinctions between entities being compared.

The roots of the Human Factors Analysis and Classification System (HFACS), are described in "Taxonomy of Unsafe Operations" (Shappell & Wiegmann, 1997) and catalogued in a Federal Aviation Administration (FAA) technical report (Shappell & Wiegmann, 2000) and a book "A Human Error Approach to Aviation Accident Analysis" (Wiegmann & Shappell, 2003). Their system is built upon the "Swiss cheese" model of Reason (1990). Reason recommended that a mishap investigation start with the *unsafe act(s)*, which represent(s) active failure. The investigation does not stop there, however, as latent failures and conditions are examined next. Latent conditions may exist undetected and unexpressed for years and include: *preconditions*, *unsafe supervision*, and *organizational influences*.

According to Reason (1990), unsafe acts include both errors and violations. Errors may be skill-based or may be due to decisional or perceptual factors. Violations may be routine (such as cutting the same corners that many others cut) or exceptional. Preconditions for unsafe acts include environmental (physical or technological) factors, conditions of operators (adverse mental states or adverse physiological states or physical/mental limitations) or personnel factors (crew resource management or personal readiness). Reason (1990) argued that it is also essential to investigate at the supervisory and/or organizational level because such factors have direct impact on preconditions. Addressing preconditions is likely to reveal opportunities to improve safety.

Unsafe supervision includes: inadequate supervision, supervisors planning inappropriate operations, a supervisor failing to correct a known problem, and supervisory violations. Finally, there are organization influences, to include resource management, organizational climate, and organizational process. One way to conceptualize these categories is to consider them "bins" containing the smaller units.

Results of a query of the Air Force Automated System (AFSAS) database, which is accessed via a secure website, for fiscal years 2010 through 2013 (1 October 2009 through 30 September 2013) to assess overall human factors involvement in aviation mishaps is depicted in Table 1. These numbers empirically demonstrate that human factors do, in fact, comprise a major concern for aviation safety.

Table 1.
Aviation Mishaps for FY 2010 – 2013

Class	Total Number Aviation Mishaps	Aviation Mishaps with at Least 1 Human Factors Code	Percentage of Aviation Mishaps with at Least 1 Human Factors Code	Total Number of Human Factors Codes
A*	129	113	87.60%	1,452
B*	218	113	51.83%	754
C*	2,518	895	35.54%	2,586
D*	3,142	737	23.46%	1,179
E*	28,803	1,094	3.8%	3,185
Grand Total	34,810	2,952	59.59%	9.156

*As defined in AFI 91-204, 12 February 2014.

It should be noted that DoD HFACS has not been required to be used for Classes C, D, and E mishaps (hi-lited). The reader is thus cautioned not to be misled by the lower percentages and the deflating impact on the grand total. The involvement of human factors is therefore likely heavily underestimated in USAF mishaps, particularly Class C, D, and E mishaps, as a result.

Beaubien and Baker (2002) while generally favorable in their review of HFACS, note that HFACS is a bit coarse, as it does not delineate *reasons* for the conditions it identifies. Beaubien and Baker also note that latent failures are difficult to identify in mishap analysis. The context of their review must be appreciated as they were examining coding schemes that were used with data already collected. Their final point is important: HFACS categories are nominal and not sequential and thus do not reveal a chain of events. Therefore, they do not differentiate causes from effects. That issue, however, is relatively easy to remedy in the overall scheme of an investigation. For example, the USAF constructs a mishap sequence of contributory and causal findings, and embeds DoD HFACS within it. O'Connor (2008) noted the above criticisms and detailed the efforts to address them, to include the formation of a Department of Defense (DoD) Working Group in 2003, which created DoD HFACS. DoD HFACS introduced increased granularity, an additional level of classification: "nanocodes." The original DoD HFACS included 147 nanocodes, organized under the categories (bins) delineated above (*unsafe acts, preconditions, unsafe supervision, organization influences*). O'Connor (2008) examined the reliability of DoD HFACS, version 6.2. He found that U.S. Navy and Marine aviators undergoing mishap investigator training were unable to achieve acceptable reliability, but noted that they had received only minimal training. Although the raters were able to agree on the nanocodes *not* used, they were unable to achieve consistent agreement concerning which nanocodes applied ("there were only seven nanocodes in which 50% or greater of the participants agreed to select the nanocode," *p.* 602). O'Connor noted that raters were confused by the number (147) of available nanocodes and that the nanocodes contained overlapping concepts. O'Connor found that collapsing codes improved inter-rater reliability. O'Connor therefore argued for nanocodes that are exhaustive, parsimonious, and mutually exclusive. O'Connor also noted that his research participants may not have been reading and considering the nanocodes' one-paragraph definitions, relying instead on the names of the nanocodes.

O'Connor called for subject matter experts to review the nanocodes to determine if some could be removed or combined with other nanocodes. O'Connor even went as far as to suggest that the nanocode level be abandoned if acceptable reliability could not be achieved without extensive training. A 2011 Aerospace Medical Association presentation, *DoD Human Factors Analysis and Classification System X*, prepared by human factors practitioners (Brian T. Musselman, Jeffrey D. Alton, Thomas G. Hughes, Patricia LeDuc, Richard J. Farley, & Antonio B. Carvalhais) from the three service safety centers had four expert raters code 54 USAF Class A mishaps with DoD HFACS version 6.2. They found a Kappa coefficient of .5494 with 76 out of 147 (52%) nanocodes having reliability greater than or equal to .60. The authors recommended: "Improve code definition," and development of an "organized training curriculum." Subsequent studies used DoD HFACS X, which contained fewer nanocodes (102, rather than 147). The average Kappa coefficient increased to an impressive 0.84 with expert coders, but novice coders continued to struggle, achieving Kappa coefficients of .2453 and .3239. The authors urged the development of a decision-tree algorithm, redesign of DoD HFACS into larger buckets (even if granularity would be sacrificed), and limiting coding at the nanocode level to experts only.

The steps in the current effort to improve DoD HFACS included determining the frequency that each of the nanocodes was used and considering retiring those nanocodes that were very infrequently used. Nanocodes that

were similar in the phenomena they described, as evidenced by having overlapping definitions, were merged and the definitions reworked. The goal was to reduce the number of nanocodes and to improve the mutual exclusivity of the remaining nanocodes. Specifically, AFSAS was further queried for fiscal years 2010 through 2013 for aviation and ground mishaps to determine the frequency of use of each of version 6.2's 147 nanocodes. It should be noted that AFSAS was queried to arrive at two totals: the first count tallied a specific HFACS nanocode cited once per mishap. Otherwise, a given nanocode assigned against multiple members of a crew would inflate the total. The other tally counted the grand total of HFACS nanocodes used, with no restriction on how many times a nanocode was used in any given mishap. The US Army and Navy, as members of the DoD HFACS Working Group, performed similar tallies. In the USAF, for example, PC 201 used only once for all classes of aviation and ground mishaps. Finally the DoD HFACS Working Group ensured that nanocodes were aligned in the correct bins. Nanocodes that were relocated to other bins were reassigned an alphanumeric to be consistent with the bin the new bin. Ultimately, the 147 nanocodes in version 6.2 were collapsed to109 nanocodes in version 7.0. The Working Group then developed a checklist, colloquially known as "Turbo HFACS," that uses a decision tree to guide investigators. A response of "yes" guides the investigator to the correct "bin" and suggests a list of defined nanocodes. This paper delineates the motivation to change DoD HFACS 6.2 and to document the changes made in DoD HFACS 7.0. This paper also examines the inter-rater reliability of DoD HFACS.

Method

Participants

Three hundred and forty students attending USAF aircraft mishap investigation courses served as participants. Most of the participants were pilots and maintenance personnel attending the Aircraft Mishap Investigation Course (AMIC) at the Headquarters, Air Force Safety Center (HQ AFSEC), Kirtland AFB, NM. Additional data was collected from aerospace medical personnel (flight surgeons, aerospace physiologists, and clinical psychologists) who attended the Aircraft Mishap Investigation and Prevention (AMIP) course, held at the at the USAF School of Aerospace Medicine (USAFSAM), Wright-Patterson AFB, OH.

Procedures

Participants were given approximately 45 minutes to read and code sanitized (basic identifying information had been removed) synopses of mishap reports that had been investigated by Safety Investigation Boards (SIBs). The synopses were approximately two typed, single-spaced, pages in length, using a 10-point font. To protect privilege, all mishap reports were immediately collected at the conclusion of the exercises. These mishap synopses are not published here because the degree of additional sanitizing that would have been necessary to publish them in this report would have rendered them virtually incomprehensible. The research design for this project was not strictly pre-planned, but rather evolved and capitalized on opportunities that presented themselves (see Table 2).

Table 2.
Summary of the Evolution of DoD HFACS version 7.0 Research Activities.

First Trials	Second Trials	Third Trials
Student investigator teams provided Checklist only.	Student investigator teams given answer sheets along with Checklist and required to submit responses on it,	Student investigator teams given answer sheets along with Checklist and required to submit responses on it.
Student investigator teams were directed to use only DoD HFACS version 7.0 for exercise.	Student investigator teams were directed to use DoD HFACS version 6.2 and then introduced to version 7.0 for exercise.	Student investigator teams were taught to use DoD HFACS version 6.2 and then introduced to version 7.0 for exercise.
18-question version of Checklist used.	8-question (with sub questions) version of Checklist used.	8-question (with sub questions) version of Checklist used.
	Student investigator teams asked to list the three to five (and then the five) most important HFACS nanocodes.	Student investigator teams asked to list the five most important HFACS nanocodes.

First Trials.

In the first data collections, 31 student investigator teams used an 18-question version of the DoD HFACS 7.0 Checklist to code three aircraft mishap scenarios. The author presented a brief introduction (approximately 10 minutes) to the DoD HFACS 7.0 Checklist. The participants were directed to use the questions of the Checklist and work together in teams of two or three members. Following the advice of O'Connor and Walker (2011), participants were organized into small teams rather than working alone to better simulate the conditions of a safety investigation board. Participants were instructed to not speak to any member of *another* team about the mishap during the exercise.

Second Trials.

The next series of data collection aimed to directly compare DoD HFACS 6.2 to DoD HFACS 7.0. Student investigator teams were given a mishap scenario and directed to first use version 6.2 as outlined in AFI 91-204. The student investigative teams conclusions were compared to the outcomes as determined by the actual Safety Investigation Board (SIB) and reviewed by the Memorandum of Final Evaluation (MOFE). After the student rater teams' responses using version 6.2 were collected, the teams were trained to use version 7.0 (using basically the same introduction described above) and directed to again code the scenario, without regard to what they coded using version 6.2. To encourage student investigator teams to read nanocode definitions, they were required to record their answers on sheets that only contained the alphanumeric codes, so that they would not base their decisions merely on the names of the nanocodes, without reading and considering the full definition. Moreover, rater groups were asked to list the three to five nanocodes that were the most important in the mishap, of course starting with those that they deemed causal. Following the input from the epidemiologists identified in the Acknowledgements, the participants were ultimately directed to list the five most important DoD HFACS nanocodes. The actual SIB and the MOFE found 12 DoD HFACS nanocodes to be applicable.

Third Trials.

Another data collection was held using a mishap that had been coded with fewer nanocodes by the SIB and which included only nanocodes that transitioned to version 7.0. Because the purpose of AMIC is to train investigators and not serve as a research laboratory, this AMIC class received more detailed instruction on a strategy to use DoD HFACS 6.2. The student investigators needed to be prepared to investigate mishaps immediately upon the completion of their training and there was no start date yet established for the operational transition to version 7.0. In applying version 6.2, student investigators were urged to read and consider definitions rather than just rely on the one-page wire diagram. After these student investigator teams completed their coding with version 6.2, their answer sheets were collected. These student investigator teams were then introduced to version 7.0, using basically the same instruction used in the first two trials.

Qualitative Feedback.

The feedback received from students led the authors and the rest of the Working Group to continually refine questions, eventually arriving at a solution of eight questions with sub-questions. Students were subsequently asked to provide written feedback on their opinions of the changes made in HFACS 7.0.

Results

First Trials.

During the first series of data collection, 18 of 31 (58%) rater teams selected the identical "yes" pattern when coding Scenario One using DoD HFACS version 7.0. Four of the 31 (13%) rater teams selected an identical but alternate pattern. Twenty-four (77%) rater teams selected the same nanocodes as the top three (out of 109) overall codes. The Fleiss' Kappa in considering the responses to the 18 questions was .847. The Fleiss' Kappa for the 109 nanocodes was .545 and the average Pairwise Cohen's Kappa was .543.

The 18-queston version of DoD HFACS, version7.0 did not fare as well with two other scenarios. In 25 rater teams coding Scenario Two, only four rater teams selected an identical "yes" pattern. There were two other common patterns with each being selected by two rater teams. Twenty (80%) rater teams selected the same top three codes. The Fleiss' Kappa in considering the responses to the 18 questions was .498. The Fleiss' Kappa for the 109 nanocodes was .415 and the average Pairwise Cohen's Kappa was .400.

Scenario Three had two of fifteen rater teams selecting an identical yes pattern. Five codes were selected 10 or more times by rater teams. Fifteen rater teams selected the top three overall codes. The Fleiss' Kappa in

considering the responses to the 18 questions was .550. The Fleiss' Kappa for the 109 nanocodes was .487 and the average Pairwise Cohen's Kappa was .512.

Second Trials.

As seen in Table 3, during the exercise using Mishap #1, when the student rater teams used DoD HFACS 6.2, nine out of 14 rater teams (64%) matched at least one of the above findings as being among their most important three to five DoD HFACS nanocodes. One rater team of 14 (7%) matched three nanocodes; eight rater teams (57%) had one match, and five rater teams (36%) had no matches of their top three to five DoD HFACS nanocodes to those of the SIB. Using DoD HFACS 7.0, two student rater teams (14%) had three matches; three student rater teams (21%) had two matches, nine student rater teams (64%) had one match, and zero student rater teams had no matches. Making this contrast even more stark (and more favorable to version 7.0) is the fact that two of the nanocodes identified in Mishap #1 using DoD HFAC version 6.2 did not transition to version 7.0 and thus were not available to the raters during the version 7.0 portion of the exercise.

Table 3.
Comparing DoD HFACS 6.2 to 7.0 Anchored Against Actual SIB Results, Mishap #1

Number of Matches to Actual SIB	DoD HFACS 6.2	DoD HFACS 7.0
3 Matches	1 student rater team matched SIB	2 student rater teams matched SIB
2 Matches		3 student rater teams matched SIB
1 Match	8 student rater teams matched SIB	9 student Rater teams matched SIB
0 Matches	5 student rater teams	

Third Trial.

The results of the exercise using Mishap #2 are presented in Table 4. Two rater teams elected to list only four codes as the "most significant" during the version 7.0 portion of the exercise and could not be persuaded to list more. By doing so, they lessened the opportunity to maximize matching what the SIB found.

Table 4.
Comparing DoD HFACS 6.2 to 7.0, Anchored Against Actual SIB Results, Mishap #2

Number of Matches to Actual SIB	DoD HFACS 6.2	DoD HFACS 7.0
4 Matches	3 student rater teams matched SIB	1 student rater team matched SIB
3 Matches	7 student rater teams matched SIB	5 student rater teams matched SIB
2 Matches	4 student rater teams matched SIB	5 student rater teams matched SIB
1 Match		2 student rater teams matched SIB
0 Matches	1 student rater team matched SIB	

Finally, student raters were asked to provide written feedback on their perception of the relative value of version 7.0 over version 6.2. Initially, the comments were mostly neutral as they are criticisms of both versions. The comments became much more positive. (nine out of 14, with no negative comments). Previously, comments from students were collected in a more informal fashion, but were still useful in the evolution of the checklist. Some typical themes from the student investigators included that the new version is "less intimidating" and "less subjective" and gives investigations structure. Suggestions for improvement included the observation that some of the questions are too broad and the sub questions need to be read and considered even if the instructions advise users to skip over the sub questions.

DISCUSSION

In a series of comparison using a variety of mishap scenarios, the checklist for DoD HFACS version 7.0 performed well. These encouraging results can be explained as follows: A taxonomy that has fewer nanocodes and nanocodes that have distinct meaning improves user satisfaction and may, itself, increase inter-rater reliability. While a systematic approach to considering the larger categories as well as the nanocodes likely is a key component to the improvement of inter-rater reliability, simply encouraging student investigators to consult the definitions of the nanocodes also likely improved inter-rater reliability. Systematically guiding investigators to consider all

nanocodes will increase the likelihood that the definitions of the nanocodes will be read and considered. As pointed out by previous researchers as noted in this report, requiring coding at a finer degree of granularity requires training and providing investigators with the proper resources, such as a checklist.

As noted by the participants, another issue is the correct structure of the checklist questions. Too many questions are likely to try the patience of investigators, while fewer questions with sub-questions run the risk of investigators missing significant areas that could benefit from further inquiry. The feedback gleaned from the students who graciously participated in this research suggest that investigators would be wise to not skim over sub-questions after answering "no" to the major question. While the DoD HFACS Working Group should consider honing the questions and sub-questions, DoD HFACS 7.0 is a step in the right direction according to student feedback and the results obtained in this study. A future revision should revise the questions and elevate some of the sub questions to free standing questions. Above all, any strategy that gets investigators to read and consider the definitions of the nanocodes will result in a better investigative outcome. The "yes/no" format of the questions in version 7.0 results in a clear binning (getting in the ballpark of applicable nanocodes). Such binning of causes and contributing factors represents an advancement in investigations with actionable results as it allows leaders to more accurately allocate resources to reduce future mishaps. Even if there is some disagreement as to which exact nanocode within a bin is the cause, at least the correct bin is identified and proper attention is paid to mitigation of a major cause or contributing factor of mishaps.

Future efforts should include a continued refinement of the questions, as noted above, as well as the creation of a small set of questions to assist Aviation Safety Action Program (ASAP) reporters submit reports that more clearly highlight human factors issues. ASAP reports are considered "safety without the mishap," and thus could better use of DoD HFACS actually help improve safety. Above all, investigators in training in all services must be given ample opportunity to practice investigating and coding mishaps during the "organized training curriculum" advocated by Musselman, et al.

REFERENCES

Beaubien, J. M. & Baker, D.P. (2002). A review of selected aviation human factors taxonomies accident/incident reporting systems and data collection tools. *The International Journal of Applied Aviation Studies, 2,* 11-36.

Hartmann, D.P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103-116.

O'Connor, P. (2008). HFACS with an additional layer of granularity: Validity and utility in accident analysis. *Aviation, Space and Environmental Medicine, 79,* 599 – 606.

O'Connor, P. & Walker, P. (2011). Evaluation of a human factors analysis and classification system as used by simulated mishap boards. *Aviation, Space and Environmental Medicine, 82,* 44-48.

Reason, J. (1990). *Human Error*. New York: Cambridge University Press.

Shappell, S.A. & Wiegmann, D.A. (1997). A human error approach to accident investigation: The taxonomy of unsafe operations. The International Journal of Aviation Psychology, 7(4), 269-291.

U.S. Air Force. AFI 91-204, 12 February 2014, Washington, DC: HQ SEC/SEF.

Wiegmann, D. A. & Shappell, S. A. (2003). A human error approach to aviation accident analysis. Burlington, VT: Ashgate.

Interested readers are directed to a more comprehensive treatment in an upcoming United States Air Force School of Aerospace Medicine (USAFSAM) technical report, *The Development and Inter-rater Reliability of DoD HFACS, Version 7.0* (King, Strongin, Lawson, & Kuhlmann, In Press).

PROMOTING AVIATION SAFETY IN AFRICA: ANALYSIS OF AIR ACCIDENTS IN THE REGION BETWEEN 2004 AND 2013

Jacob Joshua Shila
Purdue University
West Lafayette, Indiana

Amadou Anne
Purdue University
West Lafayette, Indiana

The international air traffic in the Africa region is projected to grow at an average annual rate of 5.1% between 2012 and 2032. The air transport industry in the region has supported about 6.9 million jobs, contributed about $ 67.8 million in economic activity, and about $ 80.5 million in GDP. However, the African continent was ranked last in the Universal Safety Oversight Audit Program (USOAP) report for the year 2012. Efforts by the International Civil Aviation Organization (ICAO), airlines, and governments, and other institutions are currently underway to promote aviation safety. Aviation safety implementation in the Africa region is essential as air transport is expected to play a key part in the region's economic growth through variety of means such as the transportation of passengers and cargo to and from the region. This study analyses the underlying causes of the air accidents for scheduled commercial air transport within the African continent that happened between 2004 and 2014. The focus of the study is to determine whether there is a significant difference between accidents caused by aircraft mechanical issues and accidents caused by other reasons such as pilot error or poor weather. Preliminary results indicate that most scheduled commercial air transport accidents in the region are likely to result from non-mechanical issues. A discussion is presented on ways to minimize the occurrence of air accidents in the region.

Aviation industry in the African continent has been growing despite the setbacks brought by safety and economy whereas between the years 2012 and 2013, the international air passenger demand grew by 3 percent with continual growth afterward (Moodley, 2013). Although the aviation industry in the Africa region is currently contributing about 2.3% of the global air passenger traffic and 3% of the Global Domestic Product due to aviation, it is estimated that air transport will continue to grow in the region at an annual rate of 5.1% hence contributing to the economy (ATAG, 2014). Safety is one of the challenges facing the aviation industry in the region despite the presence of respected carriers such as Kenya Airways, South African Airways, Ethiopian Airlines, and Egypt Air (ATAG, 2014). One of the safety concerns that have been addressed include the aging of most aircraft (about 20 years on average while the world average is about 10 years), and old technology used for navigation and air traffic management (Thomas, 2010). The growth of the aviation industry in the region, which is boomed by the growing economies, is not parallel with the improvement of the air traffic capabilities and airport handling capacities hence contributing to safety concerns (Hinshaw and Meichtry, 2014). Unreliable weather information on the flight routes and unstable radar coverage due to outdated technology and unskilled professionals seem to contribute to the flying difficulty for pilots in the region (Hinshaw and Meichtry, 2014). As of currently, about 11 countries of the 54 countries have met at least 60% of the requirements laid down by one of the International Civil Aviation Organization (ICAO)’s audits aiming at improving aviation safety focusing on areas such as ground crew training and rehabilitation of hangar facilities (Hinshaw and Meichtry, 2014). In 2012, the ICAO initiated a Strategic Acting Plan to improve Aviation Safety in Africa by focusing several initiatives and programs namely addressing and identifying Significant Safety Concerns (SSCs) through established audits such as Universal Safety Oversight Audit Program (USOAP), and encouraging the use of the Runway Safety Toolkit (ICAO, 2012). Other initiatives that are being promoted for application in the region include compliance to the International Air Transport Association (IATA) Operational Safety Audit (IOSA) which contains reference for safety management systems evaluation, Flight Data Analysis toolkit which is useful to air carriers in Africa for optimum flight operations, and the new ICAO Safety Management Manual (ICAO, 2012). African nations are also working to establish regional associations to oversee issues challenging the aviation industry such as safety (Michaels, 2007). Several regional associations such as the East African Community’s Civil Aviation Safety and Security Oversight Agency (CASSOA) have been established whose missions include to improve the safe conditions for lying in Africa (Michaels, 2007). While technology of the aircraft might be contributing to the safety development of the industry,

other areas such as air traffic control and navigation, ground handling, airport capabilities are of relatively equal importance when it comes to improving aviation safety in Africa. In this paper, commercial accidents that happened in Africa region between 2004 and 2013 are examined to determine the extent of technology as causal factor in those accidents. A discussion is also provided in light with the policies and technologies that have been adopted by most carriers in the region.

Possible Leading Causes of Air Accidents in Africa in the last decade

It is important to understand the causes of the air accidents in order to engage necessary policies and actions to mitigate the accidents. Although most accidents are the results of several causes, it is important to categorize the accidents in terms of a single major cause for both simplifications during the analysis and comparison of wider scope of accidents (Oster et al, 2013). While several approaches have been used to assign the cause of the air accidents, this paper applies a similar approach to that of Oster et al (2013) in which the cause of an accident is the element that triggered a chain of actions that led to an accident. Among several causes, equipment failure was the number one cause of accidents with highest number of both accidents and fatalities (about one-third of the accidents) for Part 121 related activities in the United States between 1990 and 2011 (Oster et al, 2013). Similarly, for Part 135 related activities, equipment failure was second major cause after pilot error for the same period in the United States (Oster et al, 2013). In the same duration, equipment failure seemed to be one of the major cause of air accidents in the Africa region (Oster et al, 2013). Understanding the major cause of air accidents might be a necessary step in determining potential policies and actions to pursue to reduce the occurrence of these accidents. This paper seeks to determine whether the leading cause of air accidents in the Africa region between 2004 and 2013 was also equipment failure or other causes. In the statistical analysis, the paper posits the hypothesis that equipment failure is the most significant cause of air accidents in the Africa region between 2004 and 2014.

Literature Review

General History of Aviation Safety in Africa Region

The aviation industry has experienced tremendous growth over the last few decades, boosted by a fast-growing economy and increasing domestic and foreign demand for air travel. In fact, over the 2003-2012 decade, there has been a 103% increase in the number of flights to and from the continent, approximately 5 times the global growth during the same period (Boehmer, 2013). According to IATA (2015), aviation in Africa affected more than 6.7 million jobs and fostered almost $70 billion in economic activity in 2012. However, the continent remains a relatively small market, only accounting for 3% of global air traffic (Pasztor, 2014).

Despite holding such a small share of worldwide air travel, Africa has historically held one of the poorest safety records among all regions. Indeed, over the periods spanning from 1996-2000 and 2001-2005, the rates of accidents in the continent were 3.6 and 5 per million departures respectively. As a comparison, the rates in the United States were 0.7 and 0.4, and the only other region to experience a rise in these rates across the same period was the Middle East (Air Safety Week, 2007). More recently, the number of accidents in Africa has remained higher than the global averages. For instance, in 2012 for Western-built jets, African airlines had one accident for every 270,000 flights whereas globally the average was one per 5,000,000 (IATA). Similarly, in 2013, approximately 20% of the aviation crashes and fatalities occurred in Africa (Pasztor, 2014). This poor safety record is undeniably the result of multiple social, economic and political factors. For instance, according to some sources, the bad state of many economies on the continent, combined with social issues like corruption and wars have considerably hindered the growth of aviation and the implementation of better infrastructure and safety initiatives (). In addition, international organizations like IATA and ICAO have pointed out that nationalistic interests and lack of cooperation between African states have led to the same results. These organizations have put in place multiple plans and initiatives that aim to solve the safety issue in African air transport . For instance, IATA's focus is reflected by its director general Tony Tyler, according to whom the majority of the problems "could be addressed by adherence to global standards and expanded cooperation among governments" (Boehmer, 2013).

Effects of Air Accidents Happening in the Region

Human injury or fatalities due to air accidents affect the economy and the reputation of the air transport industry of the particular country (Shyur, 2007). Air accidents have been observed to suppress demand and several African countries while on the other hand, negative publicity is supposed to affect the perception of the airlines' safety (Ishutkina and Hansman, 2009). Across the globe, airlines are said to annually lose about $10 billion to accidents which in turn result into other cost such as insurance claims, decline in productivity, damage of equipment and facilities, and decline in reputation of the carrier (Agabi, 2013). The air accident of Dana Air flight 9J 0992

which claimed 153 lives in Nigeria during resulted into increases of insurance premium for airlines depicting the fact that the operating cost challenge due rise of insurance rates caused by safety concerns is a challenge for most developing airlines in the region (Eze, 2012). Rose (1990), and Rhagavan et al (2005) conducted a study to determine if there was an association between airline profitability and airline safety and found out that higher profit margins were related to decrease accident and incident activities, especially for small carriers, assuming other operational factors are controlled.

Loss of lives may occasionally be accompanied with destruction of buildings and other facilities in the accident area. Such was the case for the Dana Air flight which happened in Nigeria on route from Abuja to Lagos (CNN, 2012). The aircraft crashed, about 4 kilometers from the Lagos International Airport, into a press company affecting near-by blocks of houses including a church hence (CNN, 2012). The loss of lives claimed through these accidents taint the image of the region to prospective investors who are looking forward to establish business ventures in the region. The air accidents that happened in in the border of Cameroon and Congo DRC involving a CASA C-212 twin turboprop claimed 11 lives, six of whom were Australian renowned mining investors (McCullough; et al, 2010). However, despite the safety and other challenges, connectivity in the African region is much needed to be able to access the resources in various places, hence aviation serves as a much potential transport tool (Thomas, 2010). The air accidents may also be both fueled and used for/as war catalyst such as the case of the Rwanda genocide. Despite the fact that suspicions of war had been boiling over time, the war was fully launched when both presidents were killed in a flight on April 6th, 1994 (Rosen, 2014).

Methodology: **Data Collection and Analysis**

This paper presents an analysis of the air accidents that happened between 2004 and 2014 in the African region. The type of accidents considered in this study involved commercial scheduled flights only. While there are several categories of major causes of air accidents (Airclaims, 2012; FAA, 2012), the study categorized the causes of air accidents as pilot error, technical failure, weather, and others (air traffic control, other aircraft, terrorism, crew etc.). The data for the accidents were collected through the Aviation Safety Database that is maintained by the Flight Safety Foundation Organization (ASN, 2015). A statistical comparison was conducted to determine the significance differences among the selected major causes of air accidents. Further analyses were conducted to determine differences among the major causes based on type of aircraft, and age of the aircraft.

The Aviation Safety Database contained information on the 13 factors that were used for this research. These factors were: date, location, aircraft model, date of its first flight, operator, engines equipped, on-board fatalities, total number of occupants, total number of fatalities (including on ground), type of flight (cargo, passenger, military), phase of flight, and summary from which the causes were extracted.

Analysis of variance was performed to determine if there was a significant difference between the technical causes of accidents (such as power plants, maintenance, or other technical failure) and other non-technical cause including pilot error, weather related accidents, other causes (including air traffic control failures), and other unknown causes. All statistical analysis was conducted using the Minitab software under the Purdue University license.

Results and Discussion

A total of 132 accidents fitting the selection criteria were retrieved from the database, spanning from January 3rd 2004 to November 29, 2013. An analysis of the location of crashes revealed that most accidents occurred in the Democratic Republic of Congo (42 accidents or 32%), the geographic area of Sudan and South Sudan (25 or 19%), Kenya and Tanzania (respectively 9 and 8 or 7% and 6%). Thus, the overwhelming majority (80%) of all accidents took place in the Eastern and Southern parts of the continent.

In terms of the causes of these accidents, 5 main categories were identified: technical failure, pilot error, weather, other and unknown. Indeed, in 23 cases, there was no conclusive information available regarding the cause of the accidents for different reasons such as the wreckage not having been found or the lack of accident investigation data. However, among the accidents for which the causes were identified, the two primary factors were technical failures and pilot errors, which accounted for 30% and 28% of all the accidents that happened during the time respectively. Weather related accidents contributed 10% of the all the accidents while other known and unknown accidents contributed about 32% of all the accidents. Figure (1) indicates; however, the frequency started to decrease although not very significantly as time progresses.

The high rate of technical failures could be linked to two other factors that this research analyzed: the origin of the aircrafts and their age. In fact, half of the aircrafts involved in these crashes were of Russian origin (Antonov, Ilyushin, etc.), while American aircrafts (Cessna, Boeing) made for 32% of the population and European-

made planes consisted of 14% (Figure 2). In addition, the median age of aircrafts that suffered these accidents, from their first flight to the date of the accidents, was 24 years, and 38% of the planes were more than 30 years old.

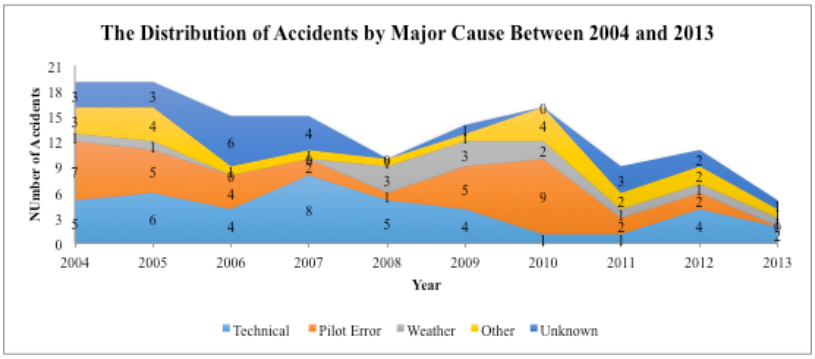


Figure 1. Distribution of Accidents by Major Causes Between 2004 and 2013

Figure 2. Countries of Origin of Aircraft Involved in Accidents in Africa Between 2004 and 2013

A further analysis of variance was performed to determine if there was any statistical significant difference among the causes. Pair comparison analysis for each pair of causes was performed using Tukey Method with significance level of 0.05. Table (1) and (2) indicate that the technical causes of accidents were not statistically significantly different from other causes except for weather related accidents. Hence, the fact that perhaps the concern of safety should be directed to all areas of improvements including aircraft technology, air traffic control regulations, air crew regulations and others.

Table 1.
Analysis of Variance for the Major Causes of Air Accidents Over the 2004-2013 Period

Source	DF	Seq	SS Contribution	Adjusted Sum of Squares	Adjusted Means Squares	F-Value	P-Value
Cause	4	52.92	23.50%	52.92	13.23	3.46	0.015
Error	45	172.3	76.50%	172.3	3.829		
Total	49	225.22	100.00%				

Table 2.
Confidence Intervals of the Causes of Accidents

Cause	N	Mean	StDev	95% Confidence Interval
Other	10	2	1.247	(0.754, 3.246)
Pilot Error	10	3.7	2.83	(2.454, 4.946)
Technical	10	4	2.211	(2.754, 5.246)
Unknown	10	2.3	1.889	(1.054, 3.546)
Weather	10	1.3	1.059	(0.054, 2.546)

Table 3.
Pair Comparison of the Major Causes of Accidents

Cause	N	Mean	Grouping	
Technical	10	4	A	
Pilot Error	10	3.7	A	B
Unknown	10	2.3	A	B
Other	10	2	A	B
Weather	10	1.3		B

In terms of fatalities, a total of 1204 ai8rcraft occupants died from these 132 accidents and there were 72 deaths on the ground. Overall then, the total number of direct casualties from aircraft accidents that were analyzed in this research amounts to 1276 individuals.

The results indicate that despite the fact most of aircraft in the region are of old, safety emphasis should be equally placed on all areas surrounding the safety factor.

References

Agabi, Chris. (2013). Africa: Airlines Lose n1.6 Trillion Annually to Accidents, Daily Trust, All Africa (Online Newspaper). Retrieved from: http://allafrica.com/stories/201311130681.html

Airclaims. (2012). World aircraft accident summary (WAAS), 1990 - 2012. CAP 479 issue 167. London: Airclaims Ltd.

Air Safety Week (2007). Improving Aviation Safety in Africa. Air Safety Week Vol 21 Issue 36. Retrieved from: https://docs.google.com/file/d/0B1skhITKOW-WZzJlME1PVDlGVlE/edit

Air Transport Action Group (ATAG). (2014). Aviation Benefits Beyond Borders, ATAG. URL: http://aviationbenefits.org/media/26786/ATAG__AviationBenefits2014_FULL_LowRes.pdf

Aviation Safety Network [ASN]. (2015). ASN Aviation Safety Network Database. Retrieved from: http://aviation-safety.net/database/

Boehmer, Jay (2013). Africa: Land of Aviation Opportunities and Obstacles. Business Travel News Vol 30 Issue 18 pg 21. Retrieved from: https://docs.google.com/file/d/0B1skhITKOW-WU05GWnhxSm4tajg/edit

CNN Wire Staff. (2012). Official: 153 on plane, at least 10 on ground dead after Nigeria crash, Cable News Network [CNN]. Retrieved from: http://www.cnn.com/2012/06/03/world/africa/nigeria-plane-crash/

Eze, Chinedu. (2012). Nigeria: Dana Crash Increases Insurance Premium for Airlines, This Day, All Africa (Online Newspaper). Retrieved from: http://allafrica.com/stories/201206220789.html

Federal Aviation Administration [FAA]. (2012). Aviation safety information analysis and sharing system (ASIAS). Retrieved from: http://www.asias.faa.gov/portal/page/portal/asias_pages/asias_home/datainfo:database:k-o

Hinshaw, Drew; Meichtry, Stacy. (2014). World News: Mali Crash Flags Africa Safety Flaws, Wall Street Journal, Eastern Edition, p. A9, Dow Jones & Company Inc, New York, US

International Civil Aviation Organization (ICAO). (2012). Strategic Action Plan To Improve Aviation Safety in Africa, ICAO News Brief, URL: www.icao.int

Ishutkina, Mariya A.; Hansman, R. John. (2009). Analysis of the Interaction Between Air Transportation and Economic Activity: A Worldwide Perspective, Doctoral Dissertation, Massachusetts Institute of Technology, Report No. ICAT-2009-2. Cambridge, MA.

McCullough, James; et al. (2010). Mining magnate Ken Talbot feared dead in plane crash over Congo, The Courier Mail, Brisbane. Retrieved from: http://www.couriermail.com.au/news/mining-magnate-ken-talbot-feared-killed-in-plane-crash-over-congo/story-e6freon6-1225881942528?nk=1355d84997ee8d58e1b786ff7a5f05af

Moodley, Kiran. (2013). When Will Africa Be the Next Aviation Hotspot? Special to CNBC.com, Paris Airshow 2013: A CNBC Special Report. URL: http://www.cnbc.com/id/100823725#.

Oster (Jr), Clinton V.; Strong, John S.; Zorn, C. Kurt. (2013). Analyzing aviation safety: Problems, challenges, opportunities, Journal of Research in Transportation Economics, 43(2013) 148-164

Thomas, G. (2010). AFRICA'S safety travails. Air Transport World, 47 (10), 35 – 38. Retrieved from http://search.proquest.com/docview/757875621?accountid=13360

Rose, Nancy L. (1990). Profitability and Product Quality: Economic Determinants of Airline Safety Performance, The Journal of Political Economy, Vol. 98, No. 5, Part (Oct., 1990), 944 – 964.

Rosen, Jon. (2014). The President's Assassins: A mansion. A crash site. And the spark that ignited the Rwandan genocide, Slate Magazine. Retrieved from: http://www.slate.com/articles/news_and_politics/roads/2014/04/rwandan_genocide_20th_anniversary_touring_juv_nal_habyarimana_s_crash_site.html

Raghavan, Sunder; and Rhoades, Dawna L. (2005). Revisiting the relationship between profitability and air carrier safety in the US airline industry, Journal of Air Transport Management 11 (2005) 283 – 290.

Shyur, Huan-Jyh. (2007). A quantitative model for aviation safety risk assessment, Journal of Computers and Industrial Engineering 54 (2008) 34 – 44. Retrieved from: doi: 10.1016/j.cie.2007.06.032

Pasztor, Andy (2014). African Air Safety Trails Rest of World. Wall Street Journal Online Edition. Retrieved from: http://www.wsj.com/articles/SB10001424052702304887104579302904075115882

POSSIBILITIES OF USING THE ON-BOARD INTELLIGENT VOICE INFORMING SYSTEMS IN COMPLEX FLIGHT SITUATIONS

Oleksandr Petrenko
National Aviation University
Kyiv, Ukraine

The paper discusses expediency and possibility of using the on-board voice informing systems aimed at issuing messages which help recognize hazardous scenarios and work out correct strategies of flight crew activities. Important possibilities of man-to-machine speech interactions in intelligent cockpits analyzed taking into account the wide range of psychological characteristics of speech. Voice channel of machine-to-man interaction regarded as one of the tools of effective conjugation of capabilities of intelligent cockpit and human's special heuristic potential. In the context of cockpit intellectualization prerequisites for the voice information reporting system functions transformation appear. The concept and flowchart of Intelligent Voice Support System (IVSS) is described and proposed as a mean of assistance for realization of human potential in a perspective socio-technical systems.

In the current context, automation is characterized by the brand new features which bring about changes in approaches to the management and support of a man-machine interaction. The systemic thinking becomes one of the most important approaches to hazard analysis and scenario forecasting (Leveson, 2013). That also requires implementation a new technologies of informing a crew for control of a situation. These technologies must comply with peculiarity of work at the modern automated workplace. Information technologies expansion led to the automation of the most complicated control processes traditionally subject to being managed by human beings only. Automation systems get the artificial intelligence characteristics somehow approximating to the human abilities. Nevertheless, the exceptionally high potential of a human being in the process of heuristic problems solving makes it necessary to construct such a man-machine interaction system which could be optimal in the context of on-board systems intellectualization still having a human being as an active control manager.

A review of recent papers shows that current man-machine systems and interfaces construction concepts reflect approaches aimed at providing a human with information of general nature as, for example, in a brand new multimodal avionics cockpit called "intelligent cockpit". The idea of intelligent cockpit finds its implementation in the construction of "Intelligent Situation-Aware Crew Assistant System" and a man-machine interface "Anticipation Support for Aeronautical Planning" (Mouthaan, Ehlert, Rothkrantz, 2003). As we can see, all above-mentioned approaches build bridges to brand new changes in the crews professional activity process structuring and transformation of their activity operational nature.

Given that the main special quality of man is his ability to heuristic activity, it is strategically correct that automated cockpit is evolving towards best conditions for such tasks as scenario forecasting, understanding the flight situation as a whole, planning, development of new algorithms for action, decision-making under conditions of lack of information (Petrenko, 2013). This is an activity which a human performs better than a machine but in order that man could realize their full potential, it must be sufficiently man-machine conjugation.

It is obvious that man-machine interaction construction approaches should provides means of this interaction and nature of the problems solved in the process of this interaction to be in harmony with each other. There are researches which make it possible to impose a task of monitoring and considering of a current human operational abilities on the on-board systems (Dorneich, Passinger, Beekhuyzen, Hamblin, Keinrath, Whitlow & Vašek, 2011). These developments allow flexible redistribution of the control tasks among crew members taking into account each human being individual status, as well as setting aside of low grade tasks reducing information stream intension with respect to the operational situation priorities.

Low grade tasks elimination may be very important at certain moments, but it is not a goal in itself. Moreover, solving problems of such type may be useful for a human being as for keeping the best control over pace of possible changes and being ready to react immediately to any deviation (P. Schutte). Nevertheless, under the conditions of technologies and automation systems becoming more and more reliable and correct, the main safety threat results from the human assessments, priorities and strategies system. Therefore, the question what and how information for human should provide the intelligent cockpit systems, must be addressed from the standpoint of creating the conditions for the disclosure of exceptional opportunities of a man. We should talk not only about unloading a human giving them more opportunities to solve heuristic tasks but providing human beings with a direct assistance in the process of tackling in the new situation of almost partnerships human and intelligent machine. But the question is what kind of tasks such a cockpit must perform and how intelligent onboard systems should convey to human the results of their capabilities.

Considering a relation between human intelligence and a speech function the use of a voice channel for conjugation of capabilities of intelligent cockpit with the potential of human is causing a particular interest. It can be assumed that verbal modality is one of the most promising channels of machine-to-man interaction in conditions of intelligent cockpit. The validity of this opinion becomes even more apparent when we begin to analyze the psychological characteristics of speech and compare them with the requirements for the channels of machine-to-man interaction in such workplaces.

The meaning and difficulties of use a speech in machine-to-man interaction

Voice information reporting system available performs the task of issuing a message of a clearly defined nature and time point. Typically such message contains information on the event or an important parameter value which requires corresponding actions to be taken, as well as a prompt about necessary current action to be applied. In all these cases the choice of a voice modality for information submitting is determined by considerations of information processing visual channel saturation and switching to the less loaded information channel, situation urgency and immediate important command performance, data collection time-saving, signal correct acceptation provision, information perception provision under the conditions of multitasking activity.

The analysis of both the crew activities in the context of the highly automated cockpit and new intelligent cockpit possibilities connected with achievements in the IT field allows to emphasize new different aspects of the voice information reporting system usage ideology relying on the role of speech for the human activities.

Ability to speak is a key feature of a human being. It is the speech that corresponds best to all represented as highly polysemious and infinite because it’s connected with the human

intelligence ability for abstract thinking. Speech has large information capacity. It meets the requirement of submitting highly generalized information to the best advantage.

It was noted that man's verbal abilities are very intimately related to his planning abilities (Miller, Galanter, Pribram, 1960, p. 38). Speech properties allow its successful usage for the human consciousness meta-structures control. The sample of such a meta-structure is a generalized flight image which contains a number of interrelated components including motivational and emotional ones. Speech makes it possible to represent a generalized event forecasting or describe preferable acting strategies, form a correct decision quickly or direct the data collection process to the necessary way.

Speech also has a great suggestive potential owing to which a spoken message is able to run through the mind dominance. Speech can help in stereotypes coping, liberating from illusion, mobilization of individual resources. Thus, automated *voice* instructing may be helpful while executing the complex manoeuvre when the sensory component of immediate perception within the multicomponent acting image can interfere with a pilot activity, especially in nontypical situation. Conceptual element integrated to the acting image allows its rational correction when necessary.

Speech is also connected with the adaptation to the social structure. It is considered that the need in such adaptation is one of the Homo Sapiens language abilities actualization factors, and social and psychological adaptation is connected with finding a common language in a team. Ability to speak is also the way for a human to identify a speech partner of a like nature.

In this regard, it is interesting to note that a number of researches comprehends changes not only of a person’s activity nature while interacting with equipment, but also of operator teams and crews functioning nature while working in the specific environment of the intelligent technological systems. This refers, in particular, to the “hybrid team” (Eschen-Léguedé, Knappe, Keye, 2011), as a significant phenomenon, when a machine becomes a symbiotic partner of a human and is perceived by them as another crew member or as an extension of their own mind.

It was shown that working successfully in highly automated Human-Machine-Interfaces in a “hybrid team” conditions demands different aspects of personality and attitudes. It should be noted that the concept of "team" is directly related to the concept of "communication" and also "personal communication», which has speech as a main modality. That’s why the usage of speech in man-machine interaction is expected to promote the garmonization of the “hybrid team”.

Approaches to the construction of onboard intelligent voice support system

The foregoing affords ground for giving more thought to the creating of a brand new man-machine verbal interaction systems being different from the voice information reporting systems available and realizing perfectly the speech capacities in man-machine interactions. The essence of the difference is seen in using voice modality not just as one of the alternative channels for the reliable single information signal deliverance but as specific means of transmitting of highly generalized capacious information which reflects the combination of the current situation aspects and tendencies. Development of such systems presents severe difficulties. Speech capacities when deal with polysemous notional units run into the problem of possible information understanding distortion risk. There arises the task to overcome controversy between need to use complex notional units and provision of the sense distortion elimination.

The problem of a correct message understanding in the “human-human” interaction is solved due to the process of dialogue when the ambiguity can be eliminated in the context of the

discourse. Similarly, discourse is important for machine-to-man interaction. Therefore machine-to-man interaction should be realized as a continuous process of tracking of crew activity, rather than separate episodes. Still another approach to the providing of the accurate meaning transmission, to our mind, may involve issuing tuple of several alternative meaning-like statements outlining the conceptual field that allows neutralize subjective perception variations of meanings.

The voice informing of a human in the intelligent cockpit should be fulfilled taking into account the human being state. It obviously deals with both message meaning and form. Creating of the adaptive voice information reporting systems suggests using message construction algorithms based on the situation pattern processing and a human being state data in the aggregate.

Interaction suggests two-way communications. There naturally arises a question on the expedience of uniting of the voice information reporting system and the crew voice messages perception system. Poll held by us showed that psychologically to a greater extent air line pilots are ready to work with intelligent voice information reporting systems rather than with voice control systems [3]. Another common factor represented the correlation between the positive attitude to the opportunity of verbal interaction with on-board systems and positive attitude to the cockpit automation. Sceptical attitude to the prospects of possible verbal interaction with on-board intelligent systems was expressed only by 27% of the pollees. Greater degree of readiness among pilots to work with intelligent voice information reporting systems rather than with the voice control systems can be explained by the fact that for the operator the voice reaction task is more difficult than the task to react to the voice. This fact was confirmed in the process of our laboratory study when the multitasking voice-inclusive operator acting conditions simulator was used. Differential voice information processing was experimentally followed by errors of 8% of the testees whereas differential voice commands issuing errors were made by 20,6% of the same testees.

It's obvious that the active voice function is more effort-taking for a human being and in an extremely tense situation it the one which would suffer more than the voice perception function. It affords ground to view the voice information reporting systems as the top-priority means of the man-machine verbal interaction assigning a support role to the crew members voice processing systems with no need for the crew to issue any specific voice commands. It seems possible for the whole crew voice activity to be specifically processed. Its analysis will allow to assess the crew members functional state and their subjective perception of the situation urgency. This speech analysis is informative in terms of tempo, intensity, intonations, frequency spectrum as well as content and conceptual harmony (speech act completeness, clarity, timeliness, adequacy). Analysis of the crew verbal interaction actual implementation correspondence to the regulations of standards and current circumstances allows to consider the beginning of the professional experience deformation process as the breakdown presage. Here we can refer to the professional experience deformation model empirically confirmed (Karapetian, Mikhailik, Pichko, Prokof'ev, 1989). According to this model, the professional experience has a multilayered structure and in tense acting situations the process of this structure deformation spreads from the outer layers to the deeper ones. Furthermore, the very first to be destructed is the outer layer, the one of the crew members interaction experience including its immanent speech component, then the cognitive, voluntary and motor layers as follows.

Figure 1 shows the Intelligent Voice Support System (IVSS) implementation model offered by us. It consists of two units, namely unit of situations and scenarios recognition and unit of voice transactions generating.

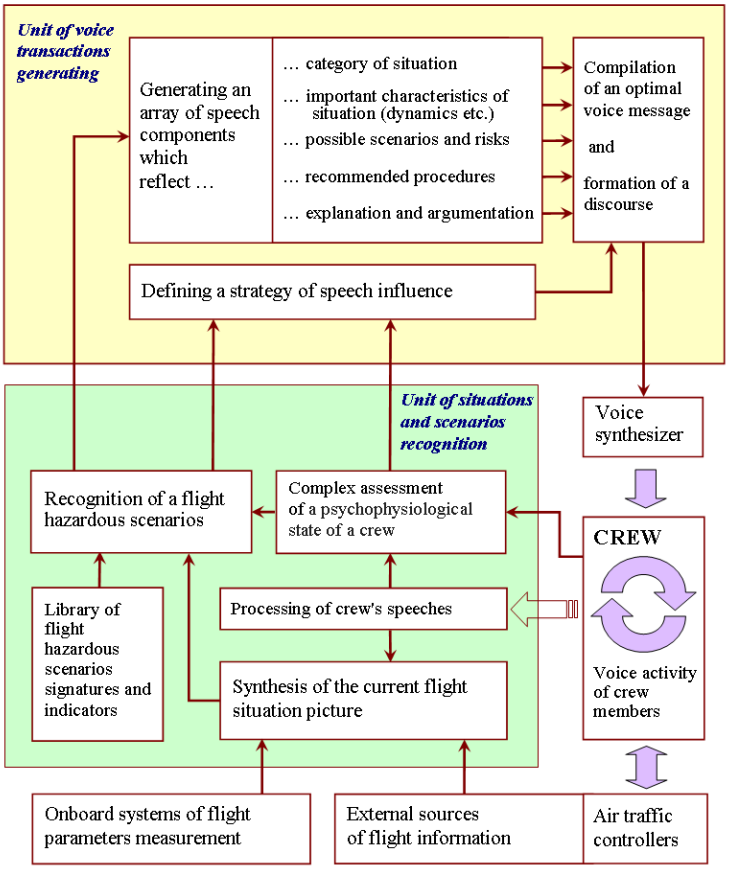


Figure 1. The flowchart of Intelligent Voice Support System (Petrenko, 2014-2015)

The main task of this system is to prevent choosing dangerous and unreasonable strategies for the crew performance, as under the conditions of various situations and circumstances they can provoke rise of nonadequate attitudes and stereotypes, faulty assessment of its abilities by the crew, etc., and lead to violation of the set principles, rules and standard operational procedures endangering the safety.

For the purpose of setting the requirements to the construction of the on-board voice support system messages which are able to help crews to recognize and overcome the flight hazardous scenarios, at the present time we launched an empirical investigation involving in it a sample group of pilots of different flight performance experience. It is expected that this investigation results would allow to take a step closer to understanding of the main principles of the voice warning and recommending transactions generating.

Conclusion

In the context of aircraft cockpit intellectualization prerequisites for the voice information reporting system functions transformation appear. IT state-of-the-art allows on-board voice information reporting systems to use speech specific properties connected with high information capacity, polysemous notional units representation possibility and highly generalized information representation. Development of real systems based on the proposed concept of IVSS requires a depth psychological researches to clarify principles of formation and patterns of understanding machine speech messages including in the context of flight crew activities.

We note finally that the focus on the complex concepts operated by the professional in their activities means eventually the focus on their professional weltanschauung, understanding of their personal role and limits, personal ambitions, etc. This affords ground to consider the intelligent voice support system construction as a forerunner of a human factor control new ideology formation in aviation. The idea is focused on the personality of the professional, not on the individual with a number of cognitive abilities. Though it is easy to say, hard to do, we hope "hard" involves its being of great interest.

References

Dorneich, M.C., Passinger, B., Beekhuyzen, M., Hamblin, C., Keinrath, C., Whitlow, S., & Vašek, J. (2011). The Crew Workload Manager: An Open-loop Adaptive System Design for Next Generation Flight Decks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Las Vegas, NV, September 19-23 (pp. 16-20).

Eschen-Léguedé, S., Knappe, K., Keye, D. (2011). Aspects of personality in highly automated Human-Maschine-Teams - Development of a questionaire. In: *Reflexionen und Visionen der Mensch-Maschine-Interaktion* - Aus der Vergangenheit lernen, Zukunft gestalten Reihe 22. Mensch-Maschine-Systeme, 33. ZMMS (pp. 459-464).

Intelligent cockpit: support to anticipation. Retrieved from: http://astute-project.eu/content/intelligent-cockpit-support-anticipation

Karapetian, G.S., Mikhailik, N.F., Pichko S.P., Prokof'ev, A.I. (1989). Preventing unintended actions of the crew in flight. - Moscow, Transport (172 p.).

Levenson, N.G. (2013) Applying systems thinking to aviation psychology. *Proceedings of the 17-th International Symposium on Aviation Psychology*, Right State University, Dayton, OH, (pp. 1-7).

Miller, G.A., Galanter, E, Pribram, K.H. (1960). Plans and the structure of behavior. – Holt, Rinehart and Winston, Inc. (p. 38).

Mouthaan, Q., Ehlert, P., L. Rothkrantz, L. (2003). Situation recognition as a step to an intelligent situation-aware crew assistant system. Retrieved from: http://www.kbs.twi.tudelft.nl/Publications/Conference/2003/Mouthaan.Q.M-BNAIC2003.html

Petrenko, O. (2013). Man-machine symbiosis in aviation: new risks and capabilities in view of information technology expansion. *Proceedings of the 17-th International Symposium on Aviation Psychology*, Right State University, Dayton, OH, (pp. 116-121).

TOWARD AN INTEGRATED ECOLOGICAL PLAN VIEW DISPLAY FOR AIR TRAFFIC CONTROLLERS

Bram Beernink, Clark Borst, Joost Ellerbroek, René van Paassen, Max Mulder

Faculty of Aerospace Engineering, Control & Simulation Division,
Delft University of Technology, Delft, The Netherlands

To cope with increasing demand on the air traffic management system, this paper proposes a novel user interface which supports air traffic controllers in conflict detection and resolution. The concept is based on previous work on 3D solution space displays for air traffic control, but then aimed at improving the interaction by better integrating speed, heading, and altitude constraints on the Plan View Display. A preliminary, subjective human-in-the-loop evaluation study was performed using five participants with various experience levels in real-life air traffic control. Results from the evaluation indicate that both successful use of the interface as well as the perceived support from the user interface depend on the participant's experience in air traffic control. The most experienced controller relied much less on the interface and found it the least useful. Future work is recommended for improving the user interface to better suit a controller's tasks.

The main responsibility of air traffic control (ATC) is the safe separation of aircraft. To resolve separation conflicts, controllers provide aircraft with heading, speed, and/or altitude clearances. Currently, the primary source of information to formulate such conflict-avoiding clearances is the Plan View Display (PVD). The PVD, however, is far from ideal as it requires controllers to integrate lower order aircraft state information, presented in flight labels, to action-relevant information in terms of what can and cannot be done to resolve a conflict. With the expected availability of high-quality digital datalinks between airborne and ground systems, the next generation PVD would allow for better information integration and improved visual representions that allow controllers to think more productively about the control problem they need to solve.

In previous research, the Delft University of Technology has designed a constraint-based interface for controllers to support aircraft separation by applying principles from Ecological Interface Design. The first version of this interface, called the Solution Space Diagram (SSD), provided conflict detection and resolution support in the horizontal plane by showing constraints on heading and speed of the aircraft to ensure separation from other traffic (Mercado Velasco, Mulder, & van Paassen, 2009). More recently, the SSD interface has been adapted by extending it to include the vertical dimension into its representation, allowing controllers to issue speed, heading, and/or altitude clearances (Lodder, Comans, Van Paassen, & Mulder, 2011). Despite a successful initial human-in-the-loop simulator trial, several issues came to light that made clear that the extended SSD could be improved by better integrating the available information.

First of all, the SSD display was a separate interface and thus not well integrated into the PVD. This made it difficult for controllers to divide their attention between the SSD and the PVD, especially in high traffic densities. Second, the controllers could not (quickly) relate the constraints rendered in the SSD with the aircraft shown on the PVD. This is especially a disadvantage for air traffic controllers, as they are responsible for multiple aircraft. Controllers

would benefit from a more exocentric perspective that makes the relationships between aircraft salient, for example to solve a conflict cooperatively. Third, the SSD required cumbersome controller inputs to preview constraints on various flight levels before actually implementing an altitude clearance. For upper area control, where altitude clearances are usually the preferred means to solve conflicts due to small speed margins, such cumbersome interactions showed to be counterproductive, especially under high traffic densities.

This paper will describe the design and a preliminary subjective evaluation of an enhanced PVD that relies on the concepts of the original SSD for controllers, but is augmented with the goal to resolve aforementioned issues.

Previous Work

Solution Space Diagram

The solution space presents the constraints imposed on an aircraft's velocity envelope by the horizontal part of the conflict zone in a velocity diagram as shown in Figure 1. It shows which combinations of speed and heading will eventually lead to a loss of separation. The diagram is constructed by first calculating the velocities that will lead to a conflict with each surrounding aircraft, called the intruding aircraft. The calculated conflict zones (i.e., Forbidden Beam Zones (FBZs)) are then clipped by an annular section that has an internal radius equal to the minimum velocity, V_{min}, and an outer radius equal to the maximum velocity, V_{max}, of the controlled aircraft.

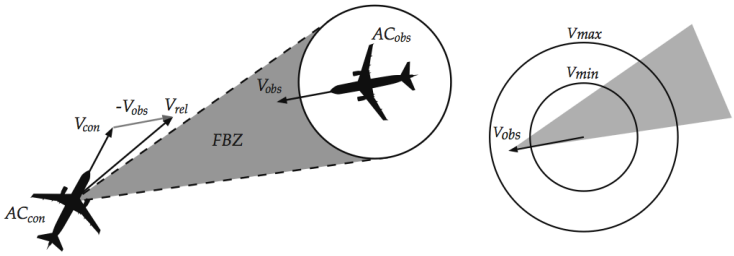


Figure 1. Conflict geometry and the resulting Solution Space Diagram (SSD) for the controlled aircraft AC_{con}.

A drawback of the solution space is that it is only presenting conflicts in the horizontal plane. Flying, on the other hand, is a three dimensional activity. This vertical component becomes especially important in climb and descent maneuvers. When only aircraft on the same altitude are shown on the display, it can not be used during climb and descent maneuvers.

Altitude-extended Solution Space Diagram

Lodder et al. (2011) showed how the FBZs from the SSD can be filtered when taking into account that aircraft can also be separated in altitude. When aircraft are separated in altitude and they stick to their altitude, the SSD of a selected aircraft just shows the conflict zones of all aircraft that are on the same flight level. However, when an aircraft is cleared to climb or descend to a different flight level, the conflict zones induced by aircraft on intermediate and the target flight level need to be taken into account. To this end, Lodder et al. (2011) proposed a

method called Altitude Based Filtering to truncate the conflict zones of aircraft on intermediate flight levels by the time the cleared aircraft is located on a particular flight level. To add a safety margin to the truncated conflict zones, they took into account the fastest time and slowest time it takes for an aircraft to reach a particular flight level. These times not only included the aircraft climb and descend performances, but also fast and slow response times of the flight crew to react to an altitude clearance. For air traffic control, Lodder et al. (2011) designed a user interface based on this filtering of FBZs (Figure 2). The interface showed the PVD and next to it the conventional SSD for a selected aircraft. The controller could press 'FL-' and 'FL+' buttons to inspect for a target altitude of this selected aircraft, and the SSD would show the FBZ filtered for that target altitude.

(a) Plan View Display (b) Altitude-Extended SSD

Figure 2. The altitude-extended SSD interface as proposed by Lodder.

The disadvantage of this interface, is that it requires the controller to use these 'FL-' and 'FL+' buttons to find a resolution in altitude. The user will therefore have to mentally integrate these different altitude-filtered SSDs for different altitudes to build an image in three dimensions. Looking at the experimental results from the altitude-extended SSD, it seems that this caused problems for participants. Ideally, all constraints at all flight levels are visualized simultaneously to better plan an altitude clearance. Second, participants also had problems linking the conflict zones to the aircraft shown on the PVD. Third, the SSD was a separate interface next to the PVD, causing controllers to devide their attention between the PVD and SSD.

Proposed Improved Concept: Integrated Solution-Space Diagram (iSSD)

This section will introduce a new user interface for conflict detection and resolution: the Integrated and Interactive Solution Space Diagram (iSSD). The iSSD aims to provide the air traffic controller with better insight into inherent constraints and relations of the ATC work domain compared to the altitude-extended SSD. The iSSD mainly consists of the Speed-Heading Diagram (SHD) and the Altitude-Heading Diagram (AHD) (Figure 3). It is assumed that the iSSD, consisting of the AHD linked to the SHD, will give the air traffic controller a more integrated three-dimensional image of the work domain constraints and relations, so that the air

traffic controller can more easily use altitude, heading and speed clearances as a means for separation. By also making the links between aircraft the conflict zones more salient, the iSSD should have the potential to overcome the disadvantages of the altitude-extended SSD.

The first noticable improvement is that the iSSD is now directly portrayed on the PVD instead of a separate interface. When selecting (i.e., clicking on) an aircraft on the PVD, the iSSD of the selected aircraft is opened. The diagram between the inner dashed circle and the outer thick circle represents the SHD, which is in fact the altitude-extended SSD, as was previously described. In this diagram it is possible to click on the conflict zone to highlight the aircraft that is responsible for the conflict zone. The Altitude-Heading Diagram (AHD) is a new visualisation, which shows safe combinations of altitude and heading to maintain separation for a selected aircraft. This diagram directly shows all altitude constraints on different flight levels. The AHD is a visualisation which uses a polar coordinate system. The radial angle is equal to aircraft heading. An increase in radius equals an increase in altitude. Since screen estate is limited, it is impossible to always display all possible altitudes with this coordinate system. Considering upper area control, there are reasonable limits that can be put on the range of altitude that is depicted. To address the issue of low display resolution in altitude, it was chosen to locally magnify altitude based on the position of the mouse cursor on the AHD. As an example, Figure 3 shows that the selected aircraft (on FL 330) has a conflict with another aircraft on the same flight level. This conflict is shown in the SHD. To resolve the conflict, a heading and/or speed change in the horizontal plane can be considered. Alternatively, a flight level change can be commanded using the AHD. The AHD in Figure 3 shows that at the current flight level and heading (i.e., green marker) a conflict is present, and also that by increasing the flight level another conflict will be triggered with the aircraft on FL 340. Because there are no conflict zones a flight level below, the selected aircraft can safely be cleared to FL 320 and continue to fly straight. Another possible solution visible within the iSSD is a heading change to the left with an altitude increase.

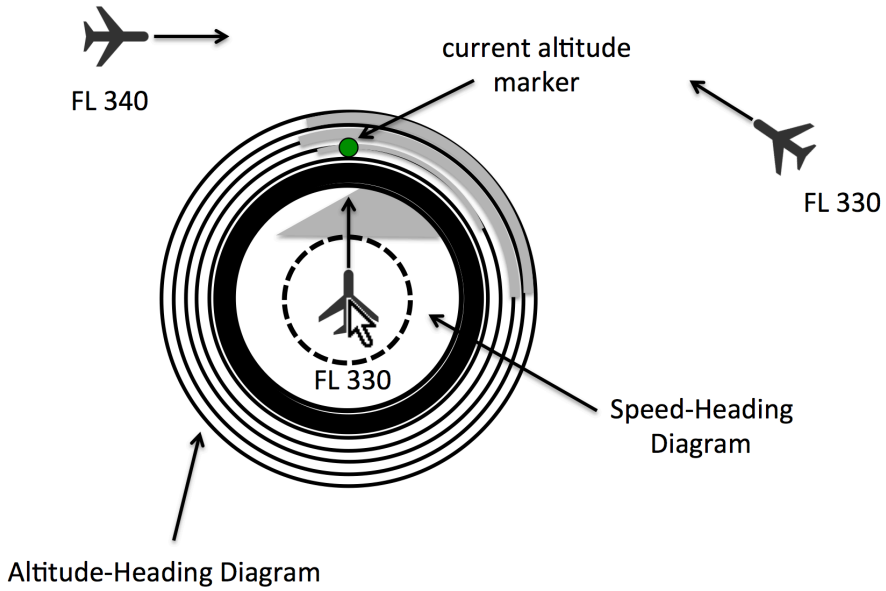


Figure 3. The iSSD interface consisting of the SHD and AHD centered around a selected aircraft.

Subjective Evaluation Study

Goal and Tasks

An initial human-in-the-loop evaluation was conducted to verify whether the iSSD was indeed able to provide support in conflict detection and resolution in various traffic scenarios (Figure 4). It was not expected that in the short available amount of time (i.e., one day per participant), participants would be able to use all the information from the iSSD. This would require extensive training. The goal, therefore, was to get initial understanding on the user interaction, get feedback on the interface and verify some of the expected support that the iSSD should provide, with particular regard to the novel interface elements of the iSSD.

An important criterion for the design of the evaluation was that participants would use the iSSD in such a way that it would reflect real-world behaviour in a demanding situation. It was therefore decided to let the participants control traffic in a sector, in which the traffic was simulated without speed- up, so that participants had some time to plan and consider options. During the simulation, the goals for the participant were threefold. In order of importance they were:

1) Maintain separation, either five nautical miles horizontally or 1,000 feet in altitude.
2) Ensure that aircraft exit the sector at the required exit altitude and within five nautical miles from the required exit point.
3) Let aircraft fly towards the required exit point at the required altitude as soon as possible.

The secondary goal reflects a restriction that occurs in real-world settings as a hand-over criterion between sectors. The tertiary goal can be seen as a more immediate request for a specific aircraft state, which for example occurs when an aircraft wants to fly at an optimal flight level. Together, the secondary and tertiary goals were introduced to allow for more control of the evaluation, in order to see whether the iSSD supported participants to meet such a specific demand for aircraft state.

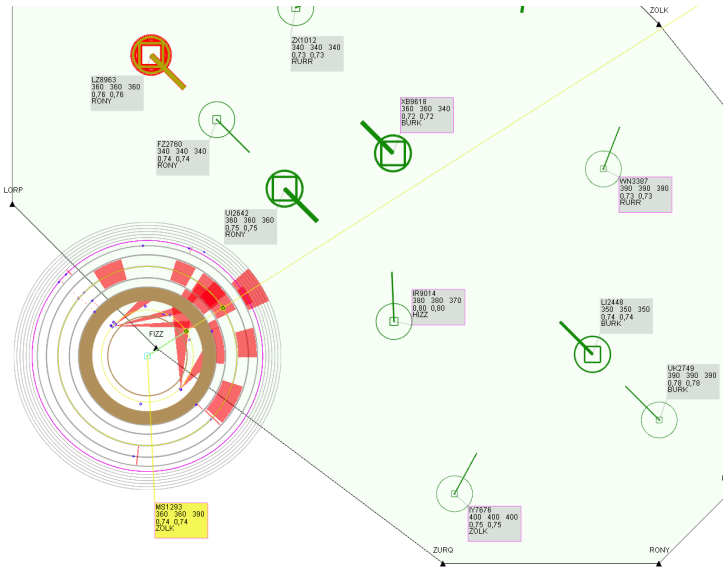


Figure 4. The iSSD prototype as integrated on the PVD, showing one of the four traffic scenarios used in the evaluation.

Participants

The sample of participants were all familiar with the aviation domain. Participants were chosen such that a wide mix existed in the level of experience that participants had with the original SSD and their experience in air traffic control. This resulted in participants with no experience with the SSD to participants who were experts with the original SSD. Regarding previous ATC experience, the lowest level of experience consisted of some hours of experience with simulated arrival management, while the most experienced participant had many years of active service at different control stations.

Results

The results of the questionnaires and the observed interaction with the iSSD showed large differences between participants. All participants without real-world ATC experience seemed to appreciate and heavily rely on novel elements of the iSSD. Without the iSSD, they indicated that controlling the various traffic scenarios would have been very difficult. However, the more experience a participant had in ATC, the less the participant made use of the iSSD to control traffic. The most experienced controller reported an increase in workload due to the iSSD and mainly used the iSSD as a "verification tool" to confirm the solution he already invented by simply using the information from the flight labels. As such, the proposed iSSD was in his opinion too wild to be of any use in a real operational setting. Finally, it was also found that the iSSD provided too much information and that the diagrams occasionally obscured the traffic pircture, resulting in clutter under high traffic densities.

Conclusion

The newly proposed interface, called the iSSD, was designed to support air traffic controllers in maintaining safe separations between aircraft by integrating low-level state information in a way that allows them to deduce action-revelant information (i.e., heading, speed, and/or altitude clearances). A brief subjective evaluation of the interface revealed that the level of ATC experience played an important role in finding the interface useful and how it was used. Whereas inexperienced participants relied most on the iSSD to control traffic, the most experienced controller mainly used the iSSD to verify his own solutions and found the iSSD to increase the workload and hinder the traffic picture on the PVD. Future work will look into integrating heading, speed, and altitude constraints in a better and perhaps simplified way to reduce clutter.

References

Mercado-Velasco, G., Mulder, M., & Van Paassen, M. (2010). Analysis of Air Traffic Controller Workload Reduction Based on the Solution Space for the Merging Task. In *AIAA Guidance, Navigation, and Control Conference*.

Lodder, J., Comans, J., Van Paassen, M.M., & Mulder, M. (2011). Altitude Extended Solution Space Diagram for Air Traffic Controllers. Proceedings of the 16th International Symposium on Aviation Psychology, Dayton (OH).

THE SMART COCKPIT INITIATIVE

Kevin M. Smith
Captain USN (Ret.)
Mesquite, Nevada, USA
Stephane Larrieu
Emirates Airlines
Dubai, UAE

Flight deck displays that automatically adapt themselves to changing operational conditions are referred to as mission adaptive displays, or smart cockpits. Most smart-enabling technology is already available in modern aircraft. To be operationally effective, however, mission adaptive displays should:

- Present mission critical information when it is most urgently needed.
- Be capable of responding to all mission critical events—single and multiple occurrences.
- Depict a rising risk profile based upon risk defining criteria.
- Utilize abstract clusters to reduce workload in high stress situations.
- Contain, whenever appropriate, performance aids so that precise maneuver execution is assured.

Mission performance aids (MPAs) possess what we call super-function attributes and, importantly, directly contribute to the precise execution of all known critical flight maneuvers. They should, as a priority, provide meaningful content for all known escape maneuvers. This escape display feature should receive urgent attention by the aviation industry.

A major supplier of software products, SAP, asserts that complexity is the most intractable problem of our age. They refer to this as the quagmire of complexity. This intractable quagmire is the result of a proliferation of technological silos across the industrial landscape—and unfortunately in our modern flight decks as well. This vertical orientation of discrete operating units makes cross-communication almost impossible. Arguably, it is the cause of much flight crew performance error detected by the aviation schoolhouse and professional instructors.

Human-caused design errors typically fall into three categories: (1) faulty, possibly unsafe design where a safety recall may be necessary, (2) poorly functioning design but not unsafe, and (3) design that appears to function, but resulting in operator confusion and workload often makes mission success questionable.

In his book *The Logic of Failure,* Dietrich Dorner (1996) says performance collapse in humans is due primarily to the application of a plausible but ineffective problem solving approach. This approach considers problems in isolation, ignoring their interactive properties, and thus losing the big picture. People have the greatest difficulty dealing with complex situations, focusing on *the what instead of the how.*

Similar studies by Daniel Kahnamen in his book *Thinking Fast and Slow* reveal the same: cognitive errors associated with judgment and choice typically fall into one or more of about twenty error categories. Significantly, when faced with a rather complex problem, most revert to rapid, ad hoc problem solving strategies that often prove ineffective.

Dr. Atul Gawande, in his book *The Checklist Manifesto*, examines the problem of complexity relating to medical surgery and asserts that we have reached the limits of reductionism, where everything is reduced to a singularity. Instead we must now enter “the century of the system.” Gawande claims we need a new emphasis, one that consider the system, its organizing principles, important interactions, and what is necessary for mission success.

Aviation has been challenged over the course of its history by complexity and its related effects. These are often referred to as degraded situation awareness (loss of the big picture), high workload, and poor judgment. Since the introduction of the B-17 Flying Fortress, checklists have been constantly in use and have become a major feature of the aviation profession, as a way to help avoid human error. Since that time complexity issues have intensified, encouraging the development of the automatic checklist for current generation airliners. Furthermore, Wickens (1984), Endsley (1995a, 1995b), Helmreich (1995), Smith (1993), Reising (1995), and others have addressed, over

many years, various aspects of human performance in complex environments. These performance issues included refocused automation, individual and team performance, operational decision making, signal detection and perception, and advanced instructional methods. All of these initiatives attempted to address the problem of complexity.

After a series of high-profile accidents, crew resource management (CRM) was introduced into the aviation operational community. These accidents revealed that perfectly sound planes were crashing because of serious performance errors by the crew. Thus programs were put in place to improve the integration of crew member functions across the full spectrum of activities that relate to mission success, thus improving the operator's management of complexity.

Ultimately, what all this means is that how we think is how we build and how we build is how we perform. We have built this reality of functional singularities or silos in the cockpit because of how we think. Our thinking has largely been focused on discrete events or functions with little or no attention to the overall system. How we build is how things work—either well or poorly. Now it is time to question our thinking (meta-cognition) and build a new kind of system which incorporates much of what we have learned concerning human and system performance.

We must first consider how we think. Once we begin to embrace effective reasoning, then we can consider how to deal effectively with complexity. Dealing with complexity challenges us to develop an understanding of the organizing principles that take a loose collection of components and create a system. CRM helps us understand how crew members can organize themselves to improve mission success. Using this as a starting point, we can then adopt many of the principles of critical thinking, and begin to work on what will usher in this century of the system—namely the specification and design of the super-function.

In order to create super-functions that are horizontally aligned, possess advanced reasoning capabilities, and communicate seamlessly with one another, we need to adopt a system of thought that will permit such an objective. For super-functions in large-scale dynamic systems to work together, the primary objective is risk management and mission completion. Risk management, as well as resource management, trajectory management, and energy management, are key performance areas. As we are beginning to appreciate in aviation, it is not enough to know the current state of the vehicle, but its projected state as well. One of the most important features of a smart cockpit therefore is kinematics.

Challenges: Operational Engineering

Operational engineering is distinctly different than technology engineering, which is currently in use today. Technology engineering assumes the value of the technology so its functional utility is not analyzed. This approach adds to the proliferation of technological silos. In many cases it does not perform as intended and thus operator-developed workarounds become the norm.

Operational engineering, as an alternative, is informed by technology and, importantly, by what constitutes mission success. Mission critical events, therefore, become the major design drivers. Operational engineering recognizes risk can continue to rise up to a point beyond which catastrophic mission failure occurs. Such a point is called the critical event horizon. Knowing key points where mission failure can occur is some of the bedrock knowledge of a super-function.

Advanced Reasoning

A new type of intelligent system architecture is needed in order for the smart cockpit to possess advanced reasoning capabilities so that intelligent units communicate with one another and risk can be effectively managed. This leads us to the creation of the super-function. The first step is to engineer a system algorithm that detects, processes, and classifies mission critical events originating from both onboard and off-board sources. The dynamic aspects of the vehicle also need to be a continuous part of the information stream.

How do we inform the design process of operational engineering? If we can answer this, then we will be able to provide for the incorporation of the super-function in modern-day cockpits.

A smart cockpit with advanced reasoning features (super-functions) must keep track of all events during the course of the mission. An event is either mission critical or not. If an event is mission critical it is placed in the

analysis pipeline. For events that are not critical, they are noted but not acted upon. Mission critical events are then classified by risk posture. For each risk posture there is a corresponding response category. So for example, if the risk posture is extreme, the event is designated as mission critical and is placed in the appropriate section that deals with extreme events. There is a direct correspondence between extreme risk and the escape response category. Similarly, there is a direct correspondence between each risk category and a response protocol. Here this must be a conscious effort; otherwise we are faced with a formless collection of data around random aspects of reality (Dorner, 1996). Selecting the proper response protocol with respect to a particular risk posture is crucial and is one of the most important features of a smart cockpit.

Challenges: Operation in Adverse Conditions

Operators in high-risk domains such as aviation often need to make decisions under time pressure and uncertainty. Time constraint reveals two dimensions; the first is the actual amount of time available to react to an event and the second is the perceived amount of time available to act. In the latter dimension, there is a "hurry up" syndrome well known among pilots, when a pilot's performance is degraded by a perceived or actual need to hurry or rush tasks or duties for any reason. Some situations are so dire and time critical that all energy and attention must be given to controlling and landing the airplane with few resources to spare. This time pressure is drastically increased with the possible catastrophic consequence of (unpaired) actions.

Responding to Adverse Conditions

Many events a smart cockpit needs to respond to are generated by adverse conditions. Meteorological conditions are not the only category of problems crews have to face. In the complex cockpit environment, a lot of information is sent to the flight deck, some of which is contradictory or even ambiguous. Human limitations and team work are also essential issues in such a time-constrained environment. These factors, whether solely or combined, are adverse conditions with which pilots are confronted. These adverse conditions are threats, hazards, and factors that generate a level of risk pilots have to manage and mitigate.

Defining Risk in the Aviation Industry

Over the years the aviation industry has developed strategies, procedures, and improved systems to make air transportation safer. However, accidents still occur. According to the Flight Safety Foundation, "The global accident rate [in 2013] was 2.8 accidents per million departures" (Jackman, 2014). Despite technical developments or improvements, there are still hazards and risks present in nature. Risk is the probability of damage, injury, death, or any negative occurrence that is caused by external or internal vulnerabilities that may be avoided through preemptive actions. A risk is qualified by its likelihood of occurrence and severity of consequence.

Not all risks can be removed, nor are all possible risk mitigation measures economically practical. It is acceptable to society that there will be some risk of harm to people, property, or the environment in order for airplanes to fly. As risk managers, pilots play a vital role in addressing the risk in practical terms. It requires a coherent and consistent process of objective analysis, in particular for evaluating the operational risks.

Risk management consists of hazard identification, risk assessment, and risk mitigation. Hazard identification is identifying adverse events that can lead to a hazard and analyzing mechanisms by which these events may occur and cause harm. Hazards are ranked in order of their risk-bearing potential. If the risk is considered acceptable, operation continues without any intervention. If the risk is not acceptable, a risk mitigation process is engaged such that control measures are taken to fortify and increase the level of defenses or to avoid or remove the risk.

Weather Conditions

The obvious adverse events are bad weather conditions. Passengers consider price, comfort, and punctuality when planning their journey. As a consequence, airlines fly almost anytime whatever the weather conditions are. However, landing in fog or on a snowy airfield contain a level of risk managed by the aviation industry.

Today's crew preparation is mainly devoted to weather report analysis and the availability of airport facilities. For example, some aircraft may be certified for CAT-2/3 approaches, which means they can land with very low minimum visibility in foggy conditions, but some airports are eligible for CAT-2/3 approaches and some are not. The level of risk is different when analyzing the weather report for different aircraft and airports. Each preparation is unique and is context dependent. The crew weather reports should not be considered as just a list of weather

conditions, but as potential mission critical events encountered in a real environment.

The main challenge with meteorological events is for the pilot to detect their existence and to have sufficient cognitive abilities to react to their cumulative effects. Indeed, landing on a wet and short runway is a demanding activity and requires accurate flying skills. It could become critical if an additional event, such as a tail wind, has not been taken into account by flying personnel, because then the workload becomes too high.

Human Factor Limitations

Flying a large jet aircraft with high-technology automated flight and guidance systems is a complex task, so much so that two crew members are still required. These crew members' activities are highly interdependent: monitoring systems and each other's actions, entering data into flight management computers, updating weather and airport information, communicating with Air Traffic Control, company operations, working with cabin crew, dealing with passengers, and anticipating future events in case of abnormal or emergency situations. When an emergency occurs, the crew must understand what the problem is, judge the level of risk and the time pressure, plan for contingencies, and decide and implement a course of action. These tasks increase the crew workload.

In high workload situations, crew errors and non-optimal responses can be linked directly to inherent limitations in human cognitive processes. Studies of human performance reveal that cognitive performance is significantly compromised under stress. When experiencing stress, human performance narrows. This phenomenon is called tunneling (Bundesen, 1990). When someone is tunneling they focus narrowly on what are perceived to be the most salient or threatening cues (Wickens, 1984). A pilot may focus on a single cockpit indicator and not notice other indications also relevant to their situation.

Toward a Solution: The Smart Cockpit Initiative

The primary (terminal) mission objective for all air transport operations is to plan and execute a mission that is at once operationally sound (safe), flown within established parameters, and executed with precision.

The Smart Cockpit

The basic idea behind the smart cockpit is to optimize mission success despite adverse conditions, extreme complexity, and increased time compression. This requires, as a top priority, the ability to optimize operational decision making (ODM), through the employment of the super-function. Decisions are tough things to deal with. As Kahnamen (2011) and others have pointed out, humans typically do not do this very well. This points to the need to develop means and measures to support and improve this difficult but necessary activity. With respect to ODM, this is only the beginning. Much work needs to be done in the area of decision support systems (DSS) in a whole range of operational arenas.

We have focused on reducing the complexity in large-scale dynamic systems—thus the introduction of the mission critical event and its management by the smart cockpit. Once we understand clearly the operational impact of the mission critical event, we can then formulate effective responses.

An important aspect of optimizing system performance (leading to mission success) is to provide the operators with the right information at the right time. Often this informational package needs to be delivered when a mission critical event has occurred and something needs to be done without delay. The timely delivery of an informational package containing mission critical information with easy opening features is our aim. Furthermore, we believe that in many cases, if not most, the informational packages should open automatically, thus ushering in truly intelligent systems (like the smart cockpit). Such an intelligent system knows (1) where it is, (2) what is happening, (3) when and how to respond, and (4) when response is complete. Notice we have made no distinction between man and machine.

A major feature of a smart cockpit is mission adaptive displays. These can automatically adapt themselves to changing operational conditions and thus can be called intelligent systems. Software giant SAP, the authors, and others such as Dr. Gawande believe we can build such systems and teach them to be intelligent. This machine intelligence not only deals with internal functions, but includes intelligent communications between significant numbers of intelligent agents. These intelligent agents can be human operators, performance aids, decision support functions, performance monitoring, kinematic prediction, and so forth. All of these comprise a set of super-functions.

Our technology can be much smarter than it is now, ushering in the century of the system. Given that the aircraft's intelligent systems (human and avionics) know that the aircraft is approaching the final approach course, then a super-function called the intelligent kinematic adviser, for example, can alert the crew when and where the onset of instability is likely to occur, and thus early crew-directed intervention can occur. Here we can see the immense value of the meaningful communications between intelligent systems. Having a kinematic adviser working with the flight crew when preparing to fly the final approach segment of the mission will significantly improve mission success and reduce complexity and workload in a high-stress, time-compressed portion of the mission. Currently, this task is done manually without aids, relying exclusively on memory and often fraught with errors.

Takeoff Operations with a Smart Cockpit

In this case the flight crew has made their final preparations for takeoff. Included in this preparation is the contingency briefing for an engine failure on takeoff—the terrain escape maneuver. The smart cockpit provides a pull-up feature for the crew to aid in the briefing exercise.

During takeoff roll a representation of the runway is automatically displayed on the center mode flight display (MFD). This display contains the exact location in which the three primary takeoff speeds (V1, VR, and V2) occur on the runway and the delayed VR speed, which is being used during this flight because of potential wind shear. If an engine failure occurs during takeoff roll at or above the takeoff refusal speed (V1), the center MFD display will automatically provide a swerve correction target to aid the crew in trajectory management. Additional display information is presented on the navigation display (ND) during takeoff roll and extending until completion of second segment climb. This pertains to the engine failure on takeoff route information. It is represented as a picture-in-picture (PIP) in the upper right hand corner of the ND and is made available for reference purposes.

The smart cockpit provides another significant feature during takeoff operations. The following critical flight maneuvers have been initialized and are ready to be automatically activated at the triggering event. These are displayed on the lower MFD and are shown in the table below.

Table 1.
Critical Flight Maneuvers

Escape Maneuver	Operational Concern
SE terrain escape	Terrain critical airport
V2 engine failure	Always possible
Wind shear recovery	LLWS advisories in effect
Upset recovery	Proceeding "heavy"

The single engine (SE) terrain escape maneuver and procedure are listed because the departure airport is terrain critical. In this case the maneuver requires a turn to 030 degrees shortly after takeoff in order to avoid a range of mountains to the west. Engine failure on takeoff, during second segment climb (V2 engine failure) is always in the takeoff and departure queue. This procedure is the most difficult maneuver to perform and is a situation judged to be the most dangerous and unforgiving of mistakes. *It is the authors' contention that all aircraft should be retrofitted with this performance aid as soon as practicable.*

Terrain Escape Maneuver

This maneuver, in the smart cockpit era, is aided by an onboard performance aid. This mission performance aid (MPA), a super-function, is automatically activated and displayed to the flight crew. This particular performance aid consists of two components. The first is the ordered representation of all components of the maneuver. The second is the route display.

The maneuver component portion of the MPA consists of five distinct maneuver elements. These are represented in Table 2. The first maneuver element depicts a turn point at 1.6 nautical miles from the VOR location. At this point a right turn must be performed to a heading of 020 degrees. The second element is triggered at an altitude of 3,200 feet. After passing this altitude, the right turn should continue to 060 degrees. This heading represents an intercept heading for the 031 degree radial of the Las Vegas VOR. Approaching the 031 degree radial, turn left and track outbound on the 031 degree radial. Track outbound on the 031 degree radial until 20 nautical miles. If the aircraft is not above 6,000 feet at this point it must enter a holding pattern until 6,000 feet is reached. Once the aircraft is above 6,000 feet and beyond 20 nautical miles, then it may proceed on course or activate and fly

another route segment.

Table 2.
Terrain Escape Maneuver

WP	L-NAV	V-NAV	PWR	Configuration
@ 1.6 DME	RT 020 degrees	C—800' AGL	MAX	F-5
(Turn)	To—020 degrees	Accel—CMS	MAX	F-1; F-U
@ 3200'	RT 060 degrees	Climb @ CMS	MCT	F-U
+ 031 R	TR—031 R	Climb to 6000'	MCT	F-U
031/20	Hold	Climb to 6000'	MCT	F-U

Commentary on the Maneuver

Taken by itself, the maneuver appears to be flyable as a horizontal procedure, although it clearly is complex. However, when one considers that there is an important vertical element to the maneuver, its complexity becomes not only recognizable, but concerning. The greatest challenge comes in the area of 1.6 nautical miles where a turn must commence. This point is where the acceleration altitude will be reached and the pitch attitude must be lowered to the acceleration target pitch. This point of the maneuver, where a turn and a pitch change must occur simultaneously, is where operational error is likely to happen. Other portions of the maneuver are similar and it is strongly suggested that a smart cockpit equipped with a terrain escape maneuver performance aid be developed as soon as practical.

Conclusion

Mission performance aids (MPAs) are a major part of the smart cockpit. They are defined as a set of flight deck attributes and crew station display features that possess the built-in capability to automatically adjust themselves to changing operational conditions. These functional attributes have been referred to as super-functions that cross traditional boundaries and thus provide for a more comprehensive response to mission critical events. The overall purpose is to provide the flight crew with mission essential information packages at the time that they are most needed. The term *information packages* used in this context is that which is required to accomplish an activity set. These mission essential information packages are derived from a comprehensive analysis of an operational environment. The overall objective of this packaging and timely delivery of mission essential information is to achieve the dual goals of mission success and excellence in flight operations.

References

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97,* 523-574.

Dorner, D. (1996). *The Logic of Failure: Recognizing and Avoiding Error In Complex Situations*. Cambridge, MA: Perseus Books.

Gawande, A. (2011). *The Checklist Manifesto*. New York, NY: Picador.

Endsley, M.R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65–84.

Endsley, M.R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64.

Helmreich, R.L. (1995). Interpersonal human factors in the operating theater. Paper presented at the Danish Anesthesia Simulator conference. Copenhagen, Denmark.

Kahnaman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Reising, J.M. 1995. Performance measures and situational awareness: How strong the link? In *Proceedings of International Conference on Experimental Analysis and Measurement of Situational Awareness.* Daytona, FL: University of Dayton.

SAP. (n.d.). WHAT THE WORLD NEEDS NOW IS SIMPLE. Retrieved February 10, 2015, from http://global.sap.com/campaigns/digitalhub-runsimple/manifesto/index.html

Smith, K. M., & Hastie, R. (1992). Airworthiness as a design strategy. In *Proceedings of the FSI Air Safety Symposium.* San Diego, CA.

Wickens, C. D. (1984). Processing resources in attentions. In *Varieties of Attention*, Parasuraman, R., & Davies, D. R. (Eds.). New York: Academic Press, 63-101.

THE EFFECTS OF WORKLOAD AND STRESS ON TEAMWORK IN A HIGH FIDELITY SIMULATION

Andrea M. Georgiou
Middle Tennessee State University
Murfreesboro, TN

With a unique high fidelity simulation lab, participants completed 3 hour work shifts to a run a simulated regional airline. The experimental design consisted of three teams randomly assigned to either a minimal, moderate, or maximum level of difficulty. Increases in workload and stress were implemented with various triggers and the participants had to quickly develop solutions to mitigate the problems. After the simulation, the participants completed the CATME (Comprehensive Assessment of Team Member Effectiveness) online survey for evaluations of their performance for five variables. (Ohland et al., 2012). Based on one-way analysis of variance (ANOVA), the results suggest only two components of teamwork were affected by workload and stress, expectation of quality and having relevant KSA's. This leads to the conclusion that generally a team will perform based on their level of team cognition and efficient group behaviors, not necessarily based on the demands of workload and stress.

Multi-team, safety critical professions such as the aviation industry, military operations, and the medical field utilize realistic Simulation-Based Training (SBT) to allow people to refine their technical and nontechnical skills in a non-consequential environment (Lazzara et al., 2010; Shapiro et al., 2008). Effective teamwork is vital for any organization to meet their performance goals, yet communication and coordination within teams rarely comes easy for people. To help improve team performance, people are often placed into various simulated settings to practice improving their teamwork skills so those skills can transfer into the workforce. Simulation can be an effective measurement tool as it allows for people to make mistakes without actual consequences. While it is important to be able to complete technical tasks, such as flying an aircraft or operating in a surgery room, there are also "non-technical" skills that must be perfected to obtain the necessary knowledge, skills, and abilities (KSA's) for high performing teams (Alinier, Hunt, Gordon, & Hardwood, 2006; Beaubien & Baker, 2004; Lazzara et al., 2010; Shapiro et al., 2008). Taking into account varying workload demands and stresses are inevitable throughout one's career, examining how these factors affect teamwork during simulations can help broaden the knowledge on group dynamics and team cognition research. Without a doubt, simulation training has the ability to revolutionize team psychological research.

Like a well-harmonized symphony, it takes effort and dedication from all team members for a team to operate efficiently. The negative effects of poor team performance can be catastrophic if operating in a high-risk field in which people's lives depend on the team communicating and coordinating as one unit. With increasing technology and teams co-located in different places across the globe, it has become even more important for psychology to continue to examine teamwork in various settings and to integrate simulation as an evaluative tool. Not only does teamwork research increase understanding of human interactions in team settings, but also the results from such studies can be used to improve training in educational and professional environments.

Review of Team Literature

The guiding foundations for the design and implementation of this study were the merits of two psychological theories: group dynamics and team cognition. The synthesis of these two theories

created a single, complete theory in which to objectively describe the often subjective concept of effective teamwork. On one hand, teamwork can be simply defined as the workings of a team; on the other hand, the teamwork construct can become complex when trying to objectively measure team performance. Many areas of research and safety-critical industries use simulation as a catalyst for examining the dynamics of group behavior and thinking and improving team effectiveness. While there are some potential negative uses of simulation that need to be avoided, Simulation-Based Training (SBT) is a highly effective tool to practice the technical and non-technical skills that must be mastered for high performing teams.

Synthesis of Group Dynamics and Team Cognition

In order to acccurately address and measure teamwork, the evolution of a single, cohesive theory surfaced from an exploration of group dynamics and team cognition theories. To simplify the complexities of team theories, team cognition explains the relevant aspects of cognitive functioning within a team and group dynamics addresses the behaviors. Together, the manner in which a team thinks and behaves determines the level of team effectiveness. The only way to truly determine if a team performs well is to examine the internal and external factors surrounding the dynamics of a team. The unification of both team theories is evident considering most research that measures teamwork includes some component of cognitive and group dynamics (Burtscher et al.,2011; Ellis & Pearsall, 2011; Waller, Gupta, & Giambatista, 2004). For example, Waller, Gupta, and Giambatista (2004) combined principles from both theories by examining adaptive behaviors and shared mental models to explain team performance. Interestingly, few differences were found to exist in team adaptability between low and high performing teams. The greatest impact on team performance was the way in which information was collected and processed, along with the accuracy of the shared mental models.

Part of the focus in group dynamics and team cognition research is to provide effective training to individuals that work in dynamic, technological environments. Often times, individuals must be flexible and be able to adapt to an ever-changing team environment (Resick et al., 2010). An excellent example of the necessity of adaptability is evident with the professional pilot career. On a weekly basis, there are changes in the times a pilot must report to duty, their route of flights, type of airplane, and the flight crewmembers they fly with in the cockpit. Regardless of the composition of a flight crew, pilots must be able to fly the aircraft and adapt to unforeseen weather conditions, system malfunctions, passenger issues and more. This requires effective communication, good decision-making, the use of all available resources, and, above all, working together as a team, all which are key components of Crew Resource Management (CRM) training (FAA, 2004).

Mandatory for airline pilots, aircraft dispatchers, flight attendants, and many other positions within airline operations, the purpose of CRM training is to reduce human error and improve performance. While it is impossible to completely eliminate human error, CRM training focuses on the continued education of vital aspects of effective teamwork so as to reduce the number of mistakes that occur from poor communication and coordination. By holistically viewing teamwork from the internal and external dyamics of a team, it fosters the ability to develop effective training modules that addresses how a team can operate in an efficient and safe manner.

Usefulness of Simulation-Based Training (SBT)

A growing body of literature supports the validity and reliability of using simulations to help improve the technical and non-technical aspects of effective team performance (Arora et al., 2010; Gettman et al., 2009). When the design of a study includes simulation training and suitable measures, this can be an excellent evaluative tool for team performance. The merits of simulation to improve team performance are evident in the ability to measure and evaluate the knowledge, skills, and abilities (KSA's) of individuals interacting in a team environment (Alinier, Hunt, Gordon, & Harwood, 2006; Beaubien & Baker, 2004; Lazzara et al., 2010; Shapiro et al., 2008). The ultimate goal of integrating

simulation is to foster a learning environment which affords the unique opportunity to practice skills and develop strategies for effective teamwork (Arora et al., 2010; Bond et al., 2007). Errors made during a simulation should be viewed as an opportunity to learn how to improve one's level of knowledge and capabilities so as to not make the same mistakes when performing in real-world teams.

Methodology

The purpose of this study was to gain a better understanding of the effects of the difficulty of high fidelity simulation on teamwork. Understanding the role of simulation difficulty is important so as to avoid the misuse of simulation which could lead to negative learning of how to effectively communicate and coordinate in a team environment (Bond et al., 2007; Salas, Bowers, & Rhodenizer, 1998). The simulation lab closely mirrors a regional airline operations center with various positions that must interact and coordinate with one another to run the airline.. The research design followed a quantitative, experimental approach with 3 experimental groups. The degree of difficulty in simulation was the independent variable (IV) that was manipulated in the experiment and was divided into minimal, moderate, and maximum levels of difficulty. The dependent variable was teamwork based on 5 dependent team effectiveness variables. The 5 team effectiveness variables were based on the Comprehensive Assessment of Team Member Effectiveness (CATME) online self and peer evaluation tool. Those categories were contributing to teamwork, interacting with the team, keeping the team on track, expecting quality, and having the needed knowledge, skills, and abilities. The emphasis was to determine how the difficulty of a simulation affects the team and their performance.

The sample type was senior-level undergraduate aviation students at a large southeastern university from their capstone course. It was important that the participants had a strong foundation of knowledge, skills, and abilities (KSA's) in order to understand and contribute during the simulation exercises; therefore, the participants were selected from a senior capstone course in which all concentrations were required to take this course their last semester. Within the capstone course, the students were randomly assigned into teams and then based on their concentration assigned to positions in the lab. Once the participants were selected and placed into teams, the teams were randomly assigned into one of the experimental groups (minimal, moderate, and maximum level of difficulty).

As the three teams were part of a larger class, the entire class received an onboarding training session which included the purpose of the simulation lab, the various functions and roles within the lab, basic operational information of the simulated airline, and the informed consent forms were handed out. After the completion of the initial training, the teams participated in the simulation exercise and were given scenarios that mimicked real-world problems and disruptions that required quick resolutions with minimal impact to the airline operations. The subject matter experts (SME's) implemented various "triggers" to the team and observe how they react to solve the issues. They are called triggers because they have the potential to cause compounding downstream impacts if not handled correctly. The level of simulation difficulty was based on the implementation of various triggers and the potential for disruption of normal operations.

Upon completion of the simulation, the participants completed the CATME (Comprehensive Assessment of Team Member Effectiveness) web-based self and peer-evaluation survey in a computer lab. After a brief explanation of the survey, the participants completed the CATME online survey and the results were accessed with a faculty login account on the CATME website. The information was delineated so it was not possible to identify the data with participants.

Results

A one-way analysis of variance (ANOVA) was used to answer the following research question: What is the difference in teamwork between groups with minimal, moderate, and maximum levels of simulation difficulty? As teamwork was measured on five categorires, there were five alternative and null hypothesis. Results from ANOVA analysis found there were significant differences in two of the five teamwork categories between the minimal, moderate, and maximum levels of simulation difficulty. As illustrated in Table 1, there were differences amongst the groups in for having relevant knowledge, skills, and abilities (KSA's), F (2, 27) = 5.765, p= .008, $p < .05$, and the expectation of quality, F (2, 27) = 6.13, $p = .006$, $p < .05$. These two null hypotheses were rejected. Specifically, the differences in means for KSA's were between the minimal and maximum level of simulation difficulty and the minimal and moderate levels. There were not significant differences in KSA's between the moderate and maximum groups. For the expectation of quality, the post hoc results showed the differences existed between the minimal and maximum groups only.

Regarding the other three teamwork variables, results from the one-way ANOVA found the differences in teamwork amongst the experimental groups not to be significant at the $p < .05$ level. Significant differences were not found in regards to the contribution to the team (F (2, 27) = .326, p= .725), interaction with the team (F (2, 27) = .147, p= .864), and keeping the team on track (F (2, 27) = .150, p= .861). As a result, he null hypotheses relating to the aforementioned teamwork variables were not rejected.

Table 1**.**
Results From the One-Way ANOVA

		Sum of Squares	Df	Mean Square	F	Sig.
Contribution to Team	Between Groups	.061	2	.030	.326	.725
	Within Groups	2.513	27	.093		
	Total	2.574	29			
Interaction with Team	Between Groups	.013	2	.006	.147	.864
	Within Groups	1.161	27	.043		
	Total	1.174	29			
Keeping on Track	Between Groups	.035	2	.017	.150	.861
	Within Groups	3.120	27	.116		
	Total	3.155	29			
Expecting Quality	Between Groups	.893	2	.446	6.130	.006
	Within Groups	1.966	27	.073		
	Total	2.859	29			
Having KSAs	Between Groups	1.173	2	.586	5.765	.008
	Within Groups	2.746	27	.102		
	Total	3.919	29			

Pair-wise comparisons between the three experimental groups of simulation difficulty and the teamwork categories were made using the Scheffe's test. Based on this test (using $\alpha = .05$), it was found that differences in expectation of quality were present between the minimal and maximum groups. The means from the Scheffe's test revealed the maximum level of group had a higher

expectation of quality (M = 4.25) from themselves and their team members as compared to the minimal level of group (M = 3.83). In regards to for having KSA's, there was not a significant difference in having KSA's between the moderate and maximum level of difficulty. On the other hand, there was a significant difference between the minimal versus moderate and the minimal versus maximum groups. For the KSA's mean scores, the minimal group had the lowest scores, M = 3.83, followed by the maximum group, M = 4.21, and the moderate group scored slightly higher than the other groups (M = 4.28). Overall, for the areas in which significant differences in teamwork were found, the minimal level of difficulty group produced the lowest scores for having relevant knowledge, skills, and abilities (KSA's) and their expectation of quality.

Discussion of Findings

With two teamwork variables being significantly different between the experimental groups, this translates to the idea that some components of teamwork are affected by the design of the simulation while others are not as susceptible to its effects. The ANOVA calculations yielded results that suggest the degree of relevant knowledge, skills, and abilities of a team and the expectation of quality are affected by the difficulty of the simulation. From the overall group perspective, the degree of simulation significantly effected having KSA's between the minimal and moderate groups, as well as the minimal and maximum group. There were not significant differences between the maximum and moderate group and this may be because there is not as great of a leap from moderate to maximum simulation as from the lowest level of minimal difficulty to a higher level. For the expecting quality variable, only the minimal versus maximum group yielded significant differences due to the level of simulation difficulty.

Taking into consideration only two out of the five teamwork variables were significantly different due to differences in workload and stress, this leads to the conclusion that generally a team will perform based on their level of team cognition and efficient group behaviors, not necessarily based on the degree of difficulty presented during a simulation. While the decrease or increase of workload and stress integrated into simulation-based training (SBT) may play a role in team performance, it is not a primary determinant in how well a team coordinates and communicates.

Recommendations for Future Research

I have three recommendations for improvements in future research. First, use a larger sample size to ensure the assumptions of sphericity and equal variances are met. This in turn leads to stronger results, improved research design, and ability to generalize from the sample to a larger representative population. Secondly, I recommend combining the self and peer CATME tool with ratings from subject matter experts that observe individual and group performance in the experimental groups. Pairing the CATME results with SME observer ratings will open the opportunity to compare the two ratings to determine if the participants are performing at similar levels rated by the experts. The third and final suggestion is to give the CATME survey to participants after a debriefing session as opposed to after completion of the simulation. The debriefing sessions would provide immediate feedback as to their positive and negative outcomes and ensure the particpants know how their team performed. Without this debriefing session, particpants lack the feedback from the SME's and are basing their performance soley on their perceptions.

References

Alinier, G., Hunt, B., Gordon, R. & Hardwood, C. (2006). Effectiveness of intermediate-fidelity simulation training technology in undergraduate nursing education. *Journal of Advanced Nursing 54*(3), 359–369.

Arora, S., Lamb, B., Undre, S., Kneebone, R., Darzi, A., & Sevdalis, N. (2010). Framework for incorporating simulation into urology training. *BJU International, 107*(5), 806- 810. doi: 10.1111/j.1464-410X.2010.09563.x.

Beaubien, J.M., & Baker, D.P. (2004). The use of simulation for training teamwork skills in healthcare: How low can you go? *Quality Safety Health Care, 13*, i51- i56.

Bond, W.F., Lammers, R.L., Spillane, L.L., Smith-Coggnins, R., Fernandez, R., Reznek, M.A., Vozenilek, J.A., & Gordon, J.A. (2007). The use of simulation in emergency medicine: A research agenda. *Academic Emergency Medicine, 14*, 353-364. doi:10.1197/j.aem.2006.11.021.

Burtscher, M.J., Kolbe, M., Wacker, J., & Manser, T. (2011). Interactions of team mental models and monitoring behaviors predict team performance in simulated anesthesia inductions. *Journal of Experimental Psychology: Applied, 17* (3), 257-269.

Ellis, A.P.J., & Pearsall, M.J. (2011). Reducing the negative effects of stress in teams through cross-training: A job-demands resources model. *Group Dynamics: Theory, Research, and Practice, 15*(1), 16-31. doi: DOI: 10.1037/a0021070.

Federal Aviation Administration [FAA]. (2004). *Advisory circular 120-5E Crew resource Management*. Retrieved on May 22, 2012, from http://rgl.faa.gov/Regulatory_and_Guidance_Library/rgAdvisoryCircular.nsf/list/AC%20 120-51E/$FILE/AC120-51e.pdf

Gettman, M.T., Pereira, C.W., Lipsky, K., Wilson, T., Arnold, J.J., Leibovich, B.C.,…Dong, Y. (2009). Use of high fidelity operating room simulation to assess and teach communication, teamwork and laparoscopic skills: Initial experience. *The Journal of Urology, 181*(3), 1289-1296.

Lazarra, E.H., Weaver, S.J., Diazgranados, D.B., Rosen, M.A., Salas, E., Wu, T.S., . . . King, H. (2010). Team Medss: A tool for designing medical simulation scenarios. *Ergonomics in Design: The Quarterly of Human Factors Applications, 18*(11), 11-17. doi: 10.1518/106480410X12658371678435.

Ohland, M.W., Loughry, M.L., Woehr, D.J., Finelli, C.J., Bullard, L.G., Felder, R.M., . . . Schmucker, D.G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self and peer evaluation. *Academy of Management Learning & Education, 11*(4), 609-630.

Resick, C. J., Murase, T., Bedwell, W.L., Sanz, E., Jimenez, M., & DeChurch, L. A. (2010). Mental model metrics and team adaptability: A multi-facet multi-method examination. *Group Dynamics: Theory, Research, and Practice, 14*(4), 332-349. doi: 10.1037/a0018822.

Salas, E., Bowers, C.A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology, 8*(3), 197-208.

Shapiro, M.J., Gardner, R., Godwin, S.A., Jay, G.D., Lindquist, D.G., Salisbury, M.L., & Salas, E. (2008). Defining team performance for simulation-based training: Methodology, metrics, and opportunities for emergency medicine. *Academic Emergency Medicine, 15*, 1088-1097. doi: 10.1111/j.1553-2712.2008.00251.x.

Waller, M.J., Gupta, N., & Giambatista, R.C. (2004). Effects of adaptive behaviors and shared mental models on control crew performance. *Management Science, 50*(11), 1534-1544.

AFTER-ACTION REVIEWS: BEST PRACTICES AND APPLICATION TO AEROSPACE EDUCATION

Richard G. Moffett III, Michael B. Hein, and Jessie M. McClure
Middle Tennessee State University

This study describes an approach to after–action reviews (AARs) used in a university capstone course that uses a high-fidelity team simulation of a flight operations center for a regional airlines. The specific methods used in the AARs are discussed in the context of comparing them to possible best practices for conducting AARs.

After-action reviews (AARs) are an approach to performance improvement and/or training that use systematic reviews of a recent performance or training event (Morrison & Meliza, 1999). After-action reviews have been used by the U.S. Army since the mid-1970's (Morrison & Meliza, 1999) and have been applied to other settings including education (Ellis, Ganzach, Castle, & Sekely, 2010) and aviation (Dismukes, McDonnell, & Jobe, 2000). They are also known as after-event reviews (Ellis & Davidi, 2005) and debriefings (Tannenbaum and Ceralsoli, 2013). Most of these approaches are characterized by four central processes that include incorporating active self-learning, having a developmental intent, focusing on specific events, and using multiple sources of information (Tannenbaum and Ceralsoli, 2013). For our discussion we will use the term after-action review (AAR) and will concentrate on their application in team situations.

Comparison of Our AAR Approach with Best Practices

A recent meta-analysis by Tannebaum and Ceralsoli (2013) found a 25% improvement in team performance when AARs were used. These authors suggested that AARs are more effective when the process is structured and uses a facilitator. They also identified that AARs are characterized by four central features: incorporating active self-learning, having a developmental intent, focusing on specific events, and using multiple sources of information. These suggestions can be viewed as best practices and will be discussed in the context of applying them to aerospace education.

The educational context for our approach to AARs is a university capstone course that uses a high-fidelity team simulation of a flight operations center for a regional airlines. Teams comprised of senior aerospace students are given various scenarios (or triggers) that require the students to work together to resolve issues quickly and effectively During the 2.5 hour simulation, the students collectively work to operate the simulated airline. The airline is a regional carrier with a fleet of 30 aircraft, two regional hubs, and 14 additional airports. Approximately 60 flight events (takeoffs and landings) occur. Much of the activity involves routine handling of flights and requires communication and teamwork. In addition, unexpected events, (severe weather; triggers such as maintenance issues or other problems requiring attention) occur and further increase the need for information transfer, coordinated action, and adaptation.

Teams typically participate in three simulations in a semester. Each simulation is followed by a facilitated AAR in which participants discuss both positive and negative performance

events and the team behaviors associated with those events. Finally, the team members derive lessons learned from the discussion of the team performance and behaviors.

Structured and Facilitated

Our approach to conducting AARs adapts some of the procedures used in the nominal-group technique (Delbecq & Van de Ven, 1971), a structured approach to brainstorming. Upon completion of a simulation, team members immediately receive a two-page AAR handout (see Figure 1) and examples of a completed observation form. These documents are reviewed with the team before they leave the simulation room. A researcher provides an overview of the AAR that includes specifying that the focus of the AAR is to help them learn how to improve their team performance, summarizing the overall process (an individual observation phase and a group discussion phase as described in the AAR handout), and encouraging all team members to complete the AAR observation form as soon as possible while the memory of the simulation experience is still recent. Team members are informed that they must bring the completed form to the AAR session as this is a required assignment of their course.

The individual observation generation phase of the AAR requires team members to use the AAR handout to capture at least two positive outcomes and two negative outcomes. They also are to list any team behaviors that contributed to each outcome.

The second phase of the AAR is a group discussion phase. This is a facilitated and structured group discussion of the individual observations the team members developed during the previous phase. Also, this phase involves having the team members develop lessons learned from the discussion of the simulation. The details of the process below reveal the structure and facilitation used in our approach to AARs.

> One week after the simulation, a 45-60 minute AAR is held in a conference room, which offers privacy and removes external distractions. Members of the team are seated around a large conference table. Also in attendance are the AAR facilitator (a researcher experienced in group facilitation) and the scribe. The scribe uses a laptop computer to collect the team members' observations/ideas, which are simultaneously projected on a screen for all team members to see. AAR ground rules are posted next to the screen. Team performance data (e.g., penalties and delay loss in dollars, average delay loss in dollars from teams in previous semesters) are posted on a whiteboard at the rear of the room.
>
> The procedural steps used in the AAR are:
>
> 1. The facilitator reminds team members of the purpose, ground rules, and procedures used during the group discussion phase of the AAR.
> 2. The facilitator reviews the team performance data and answers any questions.
> 3. The facilitator uses a round-robin procedure to allow each team member to present at least one positive outcome and the team behaviors they believe contributed to that outcome. Another round-robin procedure is used to solicit from each team member a negative outcome and the associated team behaviors. The facilitator ensures that team members focus on team behaviors

and follow the AAR ground rules. During this part of the AAR, the scribe types the ideas of each team members and simultaneously projects them on a screen for all team members to see.

4. After both round-robin procedures are completed, the facilitator provides the supplemental information developed by researchers.
5. The facilitator instructs the team members to review the notes projected by the scribe and extract what can lessons can be learned. A "feed-forward" focus is encouraged so team members will concentrate on the team behaviors to be continued in the next simulation and the team behaviors that should be improved and how. Team members share their ideas with the team, and the scribe types the ideas as they are presented and simultaneously projects them on a screen for all team members to see.
6. After the AAR is completed, team members turn in their AAR handouts with their written comments. The facilitator informs the team a) the scribe will review all of the documents and add any comments that were not presented during the AAR, b) the scribe will email each of them a copy of the summary of the results from the AAR, and c) they should use the information to help improve their team processes for the next simulation.

Active Self-Learning

Tannebaum and Ceralsoli (2013) suggested that AARs must involve active self-learning rather just receiving feedback and performance improvement suggestions from a supervisor, coach, or external observer. We promote active self-learning in our approach to AARs in a number of ways. First, team members are prompted to self-reflect on their simulation experience in the individual observation generation phase. Second, during the group discussion phase team members are actively sharing their observations with fellow team members and are discussing the impact of team behaviors on team performance. Third, team members develop ideas for improving the performance of the team in the next simulation when they create their lessons learned at the end of the AAR. Although the facilitator provides supplemental information developed by researchers based on their group discussion of observations of the team during the simulation, the majority of the information discussed in the AAR is provided by the team members.

Developmental Intent

Another characteristic suggested by Tannebaum and Ceralsoli was that AARs should focus being developmental in nature rather than being a mechanism to evaluate or judge the performance of a team. In our approach to AARs, team members are told in the initial briefing that the AARs are a part of their educational experience in the simulation and that the focus is on learning. This is reiterated in the AAR handout which describes part of the purpose of the AAR is to focus on improvement. The facilitator also reinforces this developmental intent during the AAR suggesting that team members embrace a learning goal orientation rather than a performance outcome goal orientation (Seijts & Latham, 2005); that is, focus on improving team processes rather than focusing on reducing the dollars lost in delays or penalties.

Focus on Specific Events

What differentiates AARs from other types of interventions is their focus on specific events rather than general or overall team performance (Tannebaum & Ceralsoli, 2013). According to those researchers, focusing on a specific event allows for analyzing specific behaviors and developing specific plans for improving team processes. In our approach to AARs, we help develop this focus by conducting AARs after each simulation rather than waiting until the end of the semester and conducting one AAR that would combine team performance across multiple simulations. Additionally, we have team members recall specific instances of positive and negative outcomes and the specific team behaviors associated with each outcome and write them on the AAR handout. The lessons learned created by the team members at the end of each AAR are drawn directly from the specific events and team behaviors.

Multiple Information Sources

The final characteristic suggested by Tannenbaum and Ceralsoli is that AARs should include information from more than one source. They suggest that this increases the likelihood that a breath of team behaviors will be taken into account and increases the credibility of the feedback. In our approach we use multiple sources of information. In the AAR all team members are contributing their individual observations and perceptions. We also provide team performance data from the simulation that includes monetary penalties and delay loss in dollars, and provide the average delay loss in dollars from teams in previous semesters. Finally, we also include observations from researchers. In the time between the simulation and the AAR, the researchers (approximately nine in number) meet to share their observations of the team's performance and behaviors exhibited during the simulation. The researchers identify the positive and negative performance outcomes demonstrated by the team in the simulation, and identify team behaviors that contributed to these outcomes. These observations are summarized and presented to the team as supplemental information by the facilitator during the AAR.

Conclusions

The results of the meta-analysis by Tannebaum and Ceralsoli (2013) provide a number of ideas for conducting AARs. These ideas can be seen as possible best practices for designing and conducting AARs. Our approach to conducting AARs appears to be well aligned with some of the suggestions provided by Tannebaum and Ceralsoli. Possible areas of improvement in our AAR process include implementing systematic training of facilitators to ensure consistent application of the facilitation process and reducing the amount of elapsed time between a simulation and its associated AAR. The application of our approach to other settings, (e.g., non-simulated workplaces) should be examined on an individual basis as our context, a high-fidelity team simulation used in aerospace education, may have certain advantages and resources unavailable in other contexts.

Figure 1. After Action Review handout provided to team members.

References

Baird, L., Holland, P., & Deacon, S. (1999). Learning from action: Imbedding more learning into the performance fast enough to make a difference. *Organizational Dynamics, 27,* 19–32.

Delbecq, A. L., & Van De Ven, A. H. (1971). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science, 7,* 466-492.

Dismukes, R. K., McDonnell, L. K., & Jobe, K. K. (2000). Facilitating LOFT debriefings: Instructor techniques and crew participation. *International Journal of Aviation Psychology, 10,* 35-57.

Ellis, S., & Davidi, I. (2005). After-event reviews: Drawing lessons from failed and successful events. *Journal of Applied Psychology, 90,* 857–871.

Ellis, S., Ganzach, Y., Castle, E., & Sekely, G. (2010). The effect of filmed versus personal after-event reviews on task performance: The mediating and moderating role of self-efficacy. *Journal of Applied Psychology, 95,* 122–131.

Morrison, J. E., & Meliza, L. L. (1999). *Foundations of the after action review process* (Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Seijts, G. H., & Latham, G. P. (2012). Knowing when to set learning versus performance goals. *Organizational Dynamics, 41,* 1-6.

Tannebaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 55,* 231-245.

DEVELOPMENT OF AN ALTERNATIVE METHODOLOGY FOR IMPLEMENTATION OF SAGAT DURING TASK PERFORMANCE

Durant C. Bridges
Middle Tennessee State University
Murfreesboro, TN

Situation awareness (SA) has been linked to performance in a variety of disciplines to date, but originated in the aviation arena. Situation awareness derives from attention and working memory being used toward acquiring and interpreting information from the environment (Endsley, 1995). The most revered objective method of measuring situation awareness is the Situation Awareness Global Assessment Technique (SAGAT). This technique employs random freezes of participant interfaces during simulations to query participants and assess the level of knowledge of what is happening at the time of the freezes. A discussion of applications using an alternative approach to SAGAT for operators in a high-fidelity simulation of a regional airline control room will be discussed.

An aviation department within a midsize university in the southeastern United States has built an airline operations control simulation. The simulation uses students as the operators inside the simulation. The students employ coordinated problem solving efforts toward disruption management and schedule optimization of a regional airline. Students from the department's 5 aviation specializations interactively complete a simulated work shift fulfilling the responsibilities of dispatchers, pilots, ramp controllers, crew schedulers, weather briefers, and aircraft maintenance coordinators. While it any participation in a simulation can be beneficial, their experiences may be enhanced by improvement to simulation design. Systems designed to enhance situation awareness could improve student performance as operators within the simulation. From this, they may realize the richness and full potential of participation in such a high fidelity simulation.

The focus of this paper is to establish situation awareness measurement as an integral process in the instructional design cycle of the airline control room simulation. By taking an in-depth look at student situation awareness within a high fidelity aviation simulation, we can begin to understand student perceptions of their environment as they participate as operators in human-machine systems. We can also learn what may be necessary in order to make adjustments to training parameters and system design to exploit all available resources for their benefit. As a part of an ongoing analysis cycle, we can revisit each element of the design process and make improvements that increase operator situation awareness (SA). Although the issues of situation awareness are applicable to virtually any complex system or vehicle, I will discuss the domain of airline simulation control room operators because that is the focus of this study.

Research Design

The purpose of this quantitative experimental study was to examine situation awareness levels of junior and senior-level collegiate aviation students at a mid-sized Southeastern university, as they participate in a high fidelity airline control room simulation playing the role of

aircraft maintenance coordinators. A quantitative methodology was used for this study. Endsley (1995) established that the

> Situation Awareness Global Assessment Technique (SAGAT) was developed to assess SA based on operator SA requirements. As a global measure, SAGAT queries about SA requirements, including level 1, 2, and 3 components and considers the system functioning in status and relevant features of the external environment (p. 70).

The SAGAT was used to collect data regarding the participant's levels of situation awareness and varying points during simulations.

The SAGAT developed for the study was a 15-item variable question type set of queries. It was based from an analysis of situational awareness goals and requirements determined by the aircraft maintenance coordinator position 1 and 2 responsibilities. The SAGAT was administered twice during each simulation intervention. Known as a one-group pretest-posttest design, this study investigated whether or not lab simulation exposure affected situation awareness levels in subsequent lab simulations.

Selection of Participants

The participants came from a convenience sample of a 60-student and 30-student roster in two collegiate aviation capstone upper-division level courses at a mid-sized university in the Southeast. Data was collected across two semesters, with 6 teams participating during the first semester and 3 teams participating during the second semester. The students were subdivided into 9 teams of 10, and a total of 9 students took situation awareness measures. Student-participants were representative of the sample and population. There were a total of 90 student participants at the facility throughout the study, 9 of which were participants available for the study. Nine teams each consisted of 2 pilots, 3 dispatchers, 2 aircraft maintenance coordinators, and 1 of each ramp, crew scheduling and weather personnel. The two aircraft Maintenance Coordinator positions 1 and 2 were referred to as "Aircraft Maintenance Control" and "Aircraft Maintenance Planning and Scheduling", respectively.

The participants were all within their final 4 semesters of study in the undergraduate Aerospace curriculum. While all of the airline simulation positions required a general understanding of how an airline functions, some of the lab positions required certain specific skills that students must have learned prior to entering the lab training. Pilots and dispatchers were the two positions that required additional training for their specific tasks and roles. Beyond these certain skill requirements, students were randomly assigned to teams. If the random assignments did not meet the basic needs of the skill positions, adjustments were made to accommodate for the need. Of the 10 student-participants, those assigned to the Aircraft Maintenance Coordinator 1 or "Aircraft Maintenance Control" position were tested for situation awareness measures.

Situational Awareness Requirements: Goals and Decisions

Specific goals of the Aircraft Maintenance Control position are shown in Table 1. They are based upon those of the aircraft maintenance technicians in the team situational awareness study but limited by the role in the simulation from performing physical tasks on the aircraft

(Endsley & Robertson 2000). These benchmarks may be used to determine individual SA levels, as well as the position's ability to contribute to team SA. Here, it becomes more evident the position's reliability upon relayed information pertaining to aircraft status in terms of troubleshooting progress and system functionality.

Table 1.
Maintenance Control Goals (Based upon Endsley & Robertson 2000).

1.0 Aircraft safety
1.1 Deliver aircraft in airworthy, safe condition
1.1.1 Assess reported potential problems
1.1.2 Solve problems
1.1.3 Schedule repairs
1.1.3.1 Determine maintenance issue eligibility
1.1.3.2 Placard problem
1.1.4 Schedule aircraft servicing
1.1.5 Provide quality records
2.0 Deliver aircraft on time
Prioritize tasks

Training

Prior to conducting the study, all participants were trained on use of the voice over IP communications software, aircraft maintenance activities tracking computer software, the minimum equipment list (MEL) for the CRJ-200 aircraft, as well as how to interpret the master schedule. The maintenance coordinators were instructed to utilize all available resources, including any other department to acquire the information needed to make informed decisions toward performing job functions.

As part of position training within the simulation, the participant in the role of the Aircraft Maintenance Coordinator 1 was presented with a set of questions that was to be used as the SAGAT query. They were asked to review each question and make inquiries on any questions they felt needed further clarification. All inquiries were addressed until the participant felt as though they were completely clear and understand every question with no ambiguity.

Data Collection

Before each simulation began, participants read the instructions from the Aircraft Maintenance Control Job Aid, which provided participants with specific guidance on how to perform their job functions. An additional job aid addressed how to access and manipulate program software as a part of job function. All job aids were accessible to the participants throughout all simulations and training.

During each of the simulations, a Subject Matter Expert (SME) with access to all of the information that related to the situational awareness goals and requirements completed a set of SAGAT queries from the perspective of having perfect situation awareness. Immediately after

completion by the subject matter expert, the Aircraft Maintenance Coordinator 1 participant was removed from the simulation temporarily in order to complete the SAGAT query. This process was repeated later in the simulation, for a total of 2 times per simulation. At no time was the participant removed for more than 5 minutes, which is consistent with the amount of time pilots were shown to be able to report relevant SA information following freezes in aircraft simulations without working memory decay (Endsley & Garland, 2000).

The SAGAT queries were contained within password protected Google Drive cloud storage software that could be accessed through any computer or mobile device. The queries were administered through both of these media, depending on availability and wireless internet signal strength. This method of data collection was administered to both the subject matter expert by whom the participant was compared, as well as by the participant.

Freezing

One of the most important features of data collection utilizing the objective measurement technique known as Situation Awareness Global Assessment Technique is simulation freezing. Specifically, the freezing technique was developed to overcome the limitations of memory decay in data collection upon completion of simulations (Endsley, 1988). Endsley (1995) stated, it may be possible to use the (SAGAT) technique during actual task performance if multiple operators are present to ensure safety. In a high-fidelity simulation that relies upon interdepartmental coordination of multiple teams for overall success, it may be difficult to utilize the SAGAT technique during actual task performance. The periodic "freezes" of the simulation may disrupt the performance of those participants whose situation awareness levels are not being tested. Originally developed for and practiced using fighter pilot simulations, an alternative method of querying allowed in two-pilot arrangements for one pilot to be verbally queried while the other controlled the aircraft (Endsley, 1995). Therefore, if multiple systems operators are participating in a single department with redundant responsibilities, this allows the possibility to query one operator while another temporarily assumes the position responsibilities. It is also recommended that interruption times for each SAGAT administration is delivered randomly, so that participants may not prepare in advance for the querying sessions (Endsley, 1988).

Research Question

What is the effect of exposure to a high fidelity airline operations simulation on situation awareness levels while participating in subsequent simulations in collegiate aviation students at a university in the Southeastern United States?

Hypothesis Sets

(T1 is session 1, T2 is session 2, and T3 is session 3)

Null hypothesis 1: There will be no statistically significant changes in SA Total Level SAGAT between T1 and T2.

Null hypothesis 2: There will be no statistically significant changes in SA Total Level SAGAT scores between T2 and T3.

Null hypothesis 3: There will be no statistically significant changes in SA Total Level SAGAT scores between T1 and T3.

Null hypothesis 4: There will be no statistically significant changes in SA Level 1 SAGAT scores between T1 and T2.

Null hypothesis 5: There will be no statistically significant changes in SA Level 2 SAGAT scores between T1 and T2.

Null hypothesis 6: There will be no statistically significant changes in SA Level 3 SAGAT scores between T1 and T2.

Null hypothesis 7: There will be no statistically significant changes in SA Level 1 SAGAT scores between T2 and T3.

Null hypothesis 8: There will be no statistically significant changes in SA Level 2 SAGAT scores between T2 and T3.

Null hypothesis 9: There will be no statistically significant changes in SA Level 3 SAGAT scores between T2 and T3.

Null hypothesis 10: There will be no statistically significant changes in SA Level 1 SAGAT scores between T1 and T3.

Null hypothesis 11: There will be no statistically significant changes in SA Level 2 SAGAT scores between T1 and T3.

Null hypothesis 12: There will be no statistically significant changes in SA Level 3 SAGAT scores between T1 and T3.

Measurement of SA

According to Endsley and Garland (2000), the three main reasons for measuring SA are design evaluation, evaluation of training techniques, and investigation of the SA construct. It is necessary to determine the degree to which new technologies, design concepts, and training parameters actually improve or degrade operator SA. Through understanding these processes and products, we may further understand and improve upon instructional designs used in high-fidelity simulations.

Acknowledgements

This research was supported in part by a contract (NNX09AAU52G) from NASA awarded to the Middle Tennessee State University Center for Research on Aviation Training. We are grateful to Paul A. Craig at the Department of Aerospace and the Center for Research on Aviation Training at Middle Tennessee State University for his helpful advice. Correspondence concerning this article should be addressed to Durant C. Bridges Department of Aerospace,

MTSU Box 67, Middle Tennessee State University, Murfreesboro, TN 37132. E-mail: bridges@mtsu.edu

References

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 32, pp. 97–101). SAGE Publications. Retrieved from http://pro.sagepub.com/content/32/2/97.short

Endsley, M. R. (1990). Predictive utility of an objective measure of situation awareness (Vol. 34, pp. 41–45). Presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications.

Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 65–84. doi:10.1518/001872095779049499

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. *Situation Awareness Analysis and Measurement, 10.*

Endsley, M. R., & Robertson, M. M. (2000). Training for situation awareness in individuals and teams. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement.* (pp. 349–365). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

HIGH FIDELITY SIMULATION AND AVIATION TRAINING TO IMPROVE PROBLEM SOLVING SKILLS AND COORDINATION

Evan M. Lester
Middle Tennessee State University
Murfreesboro, Tennessee
Paul A. Craig
Middle Tennessee State University
Murfreesboro, Tennessee

The Flight Operations Center – Unified Simulation (FOCUS) lab was created in 2010 to break down the "barrier" between aerospace concentrations at Middle Tennessee State University (MTSU) and address several aspects of teamwork in aviation. While participating in the FOCUS lab's high-fidelity simulations, teams of senior undergraduate aerospace students work together to solve complex and real-world scenarios, which helps improve each student's coordination, teamwork, problem-solving, and communication skills. Each team's performance in the FOCUS lab's simulations is evaluated by the FOCUS lab research team and discussed at the team's After Action Review (AAR) to determine how the team can improve its performance in the next simulation. Overall, the FOCUS lab helps prepare all undergraduate aerospace students at MTSU for working in the aviation industry.

In the last several decades, a high number of accidents and incidents have occurred in the aviation industry due to complications within teams comprised of members from different areas of aviation (Burke, Donnelly, Priest, & Salas, 2004; Hamman, 2004). These complications exist because each area of aviation focuses on teaching the skills and techniques that are vital for completing its tasks and goals. This method excludes teaching those skills and techniques that are needed to interact with the other areas of aviation (Baker, Day, & Salas, 2006). However, colleges have noticed that this method of teaching is not effective. The most effective teaching method is the use of active learning techniques (DeNeve & Heppner, 1997). Therefore, the use of simulations has been implemented in order to break down the "barrier" between aviation concentrations and to help aviation college students refine their teamwork skills in order to become exceptional performers in the aviation industry (Bowers, Rhodenizer, & Salas, 2009; Gordon, Issenberg, McGaghie, Petrusa, & Scalese, 2005).

A simulation is a device that allows teams to work in a safe environment and is similar to the actual aviation environment (Baker & Beaubien, 2004; Bond et al., 2007). In addition, it allows researchers to embed scenarios with various levels of difficulty and interject triggers into a simulation to generate a teamwork response (Burke & Salas, 2002; Gardner et al., 2008). The team can evaluate its responses to a scenario or trigger to determine the positive and negative consequences from the decisions that are made. The team is then able to prevent similar negative decisions from being made in the next simulation or the real world (Breuer & Tennyson, 2002; Burke & Salas, 2002; Hunt, Nelson, Shilkofski, & Stavroudis, 2007). Not only can the use of simulations help teams reduce the number of errors that occur in a simulation and the real world, but it can also help members from each aviation concentration to develop effective communication (Hall & Kuehster, 2010).

Simulations help members from each aviation concentration develop effective communication by practicing closed-loop communication (Hunt et al., 2007). During closed-loop communication, a team member sends a message, the receiver receives, interprets, and acknowledges the receipt of the message, and the sender follows up to ensure that the message was interpreted correctly (Burke et al., 2004; Hunt et al., 2007). The use of closed-loop communication has been found to increase team performance, enhance each team member's decision-making skills, increase team knowledge, and help team members understand the vital importance of each team member's communication and input in a team setting (Fiore & Salas, 2004; Hall & Kuehster, 2010; Krivonos, 2007).

Based on the literature, it is important to bring members from all aviation concentrations together through simulations to enhance communication, increase safety, and reduce the number of errors that occur in a team setting. One simulation that allows members from all aviation concentrations to come together is the Flight Operations Center – Unified Simulation (FOCUS) lab. The FOCUS lab is a high-fidelity simulation that replicates a regional airline operations center. Senior undergraduate aerospace students from all concentrations at Middle Tennessee State

University (MTSU) are placed into teams to work a "shift" in a position that is directly related to their concentration, such as the maintenance controller, pilot, flight operations coordinator, ramp controller, and crew scheduler, in the FOCUS lab's virtual airline called Universal E-Lines. This virtual airline allows teams to operate and manage 30 aircraft, specifically Canadair Regional Jets (CRJ) – 200s. During their three-hour simulation, a team handles approximately 80 flights that fly on 16 designated flight routes throughout the southeastern United States. A team also has to manage and solve routine and non-routine events that occur during the team's simulation by effectively communicating and coordinating with each team member while adhering to federal regulations and standard operating procedures. In addition, research is continuously conducted during the FOCUS lab's simulations by the FOCUS lab research team to discover the best practices for teamwork, information utilization, and communication in an aviation team setting.

FOCUS Lab Background and Concept

Before the creation of the FOCUS lab, MTSU undergraduate aerospace students were taught in educational "silos," which means that each aerospace concentration taught its students the information and skills that were only needed for its concentration. For example, flight dispatchers trained only with other flight dispatchers in MTSU's aerospace program to learn how to safely and effectively dispatch aircraft. The issue with this educational technique was that undergraduate aerospace students were not able to interact with other undergraduate aerospace students from every aerospace concentration. This situation caused students to not fully understand how members from all areas of aviation work together in teams to perform various tasks in an aviation team setting. Also, many experts in the aviation industry have reported that it takes up to 10 years for newly hired aviation professionals to fully understand how an airline operates and how their performance and decisions impact an airline. Therefore, Dr. Paul A. Craig, an aerospace professor at MTSU, received a NASA grant in 2010 to create the FOCUS lab in order to decrease the number of years it takes for a newly hired aviation professional to fully understand the big picture of an airline and break down the educational "silos" in order to give senior undergraduate aerospace students the opportunity to work together in teams to enhance their teamwork skills that are vital for working in the aviation industry.

Each team that participates in the FOCUS lab's simulation is composed of 10 to 12 senior undergraduate aerospace students that are placed in a specific position that is most related to his or her aerospace concentration. However, the students do not work together in a single location. There are four distinct locations that are utilized during a simulation. The FOCUS lab houses Universal E-Lines' operations center. The positions that are located in the operations center include the flight operations coordinator, flight operations data, flight tracking and scheduling, weather and forecasting, crew scheduling, maintenance control, and maintenance planning and scheduling. In a room adjacent to the FOCUS lab, the ramp tower position manages all aircraft arriving and departing from one of the airports that is used by Universal E-Lines, specifically Nashville International Airport (KBNA). In an office across from the FOCUS lab, the pseudo pilot position plays the role of the pilot for each Universal E-Lines' simulated aircraft, except for one. The one aircraft that is not controlled by the pseudo pilot is controlled by two students in the position known as the Canadair Regional Jet (CRJ) – 200 simulator flight crew. At the Murfreesboro Municipal Airport (KMBT) there is a CRJ – 200 flight training device (FTD) that two students fly for Universal E-Lines during a simulation.

Throughout a simulation, the FOCUS lab research team, which consists of undergraduate students, graduate students, and professors from MTSU's Aerospace and Industrial and Organizational Psychology Departments, implements complex and real-world scenarios into a simulation to give students the opportunity to apply their knowledge that they have gained from the classroom to solve the scenarios. Once a team determines and executes a solution for a particular scenario, a team quickly learns how its solution affected Universal E-Lines through immediate feedback and real-time performance analysis, which includes simulated financial data. In addition, various measures are conducted by the FOCUS lab research team during a simulation to give effective feedback to a team at its After Action Review (AAR). An AAR is a debriefing process where a team identifies ways to improve its performance the next time it participates in a FOCUS lab simulation.

Implementation of Scenario Triggers

The FOCUS lab research team implements real-world scenarios, or triggers, that vary in difficulty into a FOCUS lab simulation. These scenarios must be resolved by a team in a safe, effective, efficient, and quick manner

that has a minimal to no impact to Universal E-Lines. Also, the team's solution for each scenario must comply with federal regulations and standard operating procedures. However, each scenario has the potential to cause downstream implications if it is not correctly handled by a team. Therefore, once a team determines and executes a solution to a specific scenario, the FOCUS lab research team will evaluate the team's solution and determine the direction in which the scenario will go based on the team's solution. For example, if a team dispatches an aircraft overweight, then there will be both simulated legal and financial ramifications that the team will face. Ultimately, this realistic method helps each student on a team understand how his or her performance and decisions impact the virtual airline's safety and economics.

High-Fidelity Components of the Simulation

To make the FOCUS lab's operations center and simulations realistic, the FOCUS lab relies on various types of technology that are both commercially available and specifically developed for the FOCUS lab.

At each position's station in the FOCUS lab, there are desktop computers with dual monitors that give each team member space to organize and display multiples sources of information and special programs that he or she needs to perform the tasks associated with his or her position. Headsets are also connected to each position's desktop computer that can be used for verbally communicating with any team member. In addition, three large LCD television screens on each sidewall in the FOCUS lab display information that is commonly used by each position. Specifically, the television screens display real-time weather maps, the flight tracking radar, and the flight status board. Adjacent to the FOCUS lab, the ramp tower room houses three large LCD television screens, 12 computers, and several control stations. These pieces of equipment operate the software that controls the movement of Universal E-Lines' simulated aircraft on the flight tracking radar along 16 designated flight routes in the southeastern United States. Also, the television screens in the ramp tower room display a 150-degree view of Concourse C at Nashville International Airport (KBNA), which is one of 16 airports utilized by Universal E-Lines. This view allows the ramp tower position to monitor and manage the movement of Universal E-Lines' simulated aircraft at the airport.

Each position in the FOCUS lab utilizes an interactive Microsoft Excel document that is specifically made to help students retrieve the data they need to perform their position and tasks. The data in the Excel documents can also be manipulated by students to gather the information that needs to be given to the other positions. In addition, each Excel document consists of a detailed flight status board that displays the flight number, departure airport's International Civil Aviation Organization (ICAO) identifier code, departure time, arrival airport's ICAO identifier code, and arrival time for every simulated flight. The flight status board also utilizes status lights for every flight that automatically update based on Greenwich Mean Time (GMT). Also, the flight status board displays and calculates the total time of delays, average departure performance time, daily revenue, and financial delay loss based on a team's performance in a simulation.

There are two computer applications that are used by both students at each position and the FOCUS lab research team to effectively communicate and manage information. One of the computer applications used is Skype. By using Skype, each position in the FOCUS Lab, ramp tower room, and pseudo pilot room can communicate via voice or text with any team member to conduct tasks as a team during a simulation. The FOCUS lab research team uses Skype to communicate via voice or text with each position during a simulation to respond to a team's solution for each scenario that is implemented into the simulation. The second computer application used is an internet application called "join.me." This application gives the FOCUS lab research team the capability of observing the computer screens at each position on an internet-enabled device, such as a smartphone, tablet, or a computer, to assess each student's performance during a simulation.

At the Murfreesboro Municipal Airport (KMBT), a Federal Aviation Administration (FAA) – certified Canadair Regional Jet (CRJ) - 200 flight training device (FTD), or simulator, is used during the FOCUS lab's simulations. Due to the network connections that connect the CRJ - 200 simulator to the FOCUS lab, every position in the FOCUS lab can track the flight path of the CRJ - 200 simulator on the flight tracking radar screen while the CRJ – 200 simulator flight crew flies the CRJ - 200 simulator. The CRJ – 200 simulator flight crew can also communicate with the flight operations coordinator, ramp tower, weather and forecasting, and maintenance controller positions in the FOCUS lab through various communication protocols to gather essential information during all phases of flight.

Recently, 17 documents were made to help the FOCUS lab research team determine whether or not each student was performing his or her tasks correctly. Also, these documents keep track of simulated financial penalties that a team receives during a simulation for negatively affecting the safety of Universal E-Lines, not adhering to federal regulations, or not following standard operating procedures. After a simulation, these documents are used during the After Action Review (AAR) to give each student concrete feedback on his or her performance. Ultimately, these documents help students realize that their performance and decisions actually affect the virtual airline

The After Action Review (AAR)

The primary source of learning in the simulation process is the debriefing process (Holtschneider, 2007). A debrief is defined as reviewing a simulation after its completion to determine a team's positive and negative actions during a simulation (Bonacum, Graham, & Leonard, 2004; Hunt et al., 2007). During a debrief, a professional and experienced facilitator guides the debriefing process by allowing each team member to discuss what he or she believes contributed to the team's success and what contributed to undesirable outcomes (Bond et al., 2007; Hall & Kuehster, 2010). Once each team member discusses his or her views on the simulation, the facilitator reviews both positive and negative aspects of the team's performance (Burke et al., 2004). As a result of a simulation debrief, the team learns how to focus on repeating the types of performances that were successful, avoiding the types of performances that were detrimental to the team, reducing the number of errors that were made, and increasing the level of safety in the decision-making process in both a simulation and the real world (Burtscher, Kolbe, Manser, & Wacker, 2011; Gardner et al., 2008; Hall & Kuehster, 2010).

Throughout a simulation, the FOCUS lab research team monitors and evaluates each team's performance. Once the evaluations are completed, MTSU Industrial and Organizational (I/O) Psychology graduate students and professors perform and facilitate an After Action Review (AAR), which gives a team feedback on its performance during its simulation. Also, the AAR gives each team member the opportunity to discuss aspects of the team's performance that were successful and unsuccessful in order to learn from the team's mistakes. This reinforces a team's positive performances and helps a team build new strategies and goals to prevent similar mistakes from being made in the next simulation. In addition, the MTSU I/O Psychology graduate students and professors discuss decisions that were made by the team that violated standard operating procedures or federal regulations, if any, to ensure those decisions are not made in the actual aviation industry. By participating in an AAR, students ultimately improve their teamwork, problem-solving, and coordination skills, develop strategies to combat their weaknesses, and enhance their strengths to help them become exceptional aviation professionals.

Summary

In conclusion, the FOCUS lab is an invaluable tool for an aerospace student's collegiate training, researchers, and the aviation industry. By participating in the FOCUS lab's high-fidelity simulations of a regional airline operations center and After Action Reviews (AAR), senior undergraduate aerospace students are capable of refining their teamwork, coordination, communication, problem-solving, and decision making skills that are needed to become a successful aviation professional. Also, their participation helps reduce the amount of time they need to fully understand how an airline operates and how their performance and decisions ultimately impact an airline. The FOCUS lab also gives the FOCUS lab research team the opportunity to conduct research during the lab's simulations in order to address several key areas of teamwork in aviation, including communication, information utilization, coordination, and team performance. With the continuous advancements being made to the FOCUS lab's high-fidelity simulations, the FOCUS lab will continue to be an invaluable tool for all MTSU undergraduate aerospace students, researchers, and the aviation industry.

Acknowledgements

This research was supported in part by a contract (NNX09AAU52G) from NASA awarded to the Middle Tennessee State University Center for Research on Aviation Training. We are grateful to Dr. Paul A. Craig at the Department of Aerospace and the Center for Research on Aviation Training at Middle Tennessee State University.
Correspondence concerning this article should be addressed to Evan M. Lester, 1301 East Main Street, MTSU Box 6582, Murfreesboro, TN, 37132. Email: eml3d@mtmail.mtsu.edu

References

Baker, D. P., & Beaubien, J. M. (2004). The use of simulation for training teamwork skills in health care: How low can you go?. *Quality and Safety in Health Care, 13,* 51-56. doi: 10.1136/qshc.2004.009845

Baker, D. P., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research, 41*(4), 1576-1598. doi: 10.1111/j.1475-6773.2006.00566.x

Bond, W. F., Coggins, R. S., Fernandez, R., Gordon, J. A., Lammers, R. L., Reznek, M. A., … Vozenilek, J. A. (2007). The use of simulation in emergency medicine: A research agenda. *Academic Emergency Medicine, 14*, 353-364. doi: 10.1197/j.aem.2006.11.021

Bowers, C. A., Rhodenizer, L., & Salas, E. (2009). It is now how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology, 8*(3), 197-208. doi: 10.1207/s15327108ijap0803_2

Breuer, K., & Tennyson, R. D. (2002). Improving problem solving and creativity through use of complex-dynamic simulations. *Computers in Human Behavior, 18*, 650-668. doi: 10.1016/S0747-5632(02)00022-5

Burke, C. S., Donnelly, K. W., Priest, H., & Salas, E. (2004). How to turn a team of experts into an expert medical team: Guidance from the aviation and military communities. *Quality and Safety in Health Care, 13*, 96-104. doi: 10.1136/qshc.2004.009829

Burke, C. S., & Salas E. (2002). Simulation for training is effective when …. *Quality and Safety in Health Care, 11*, 119-120. doi: 10.1136/qhc.11.2.119

DeNeve, K. M., & Heppner, M. J. (1997). Role play simulations: The assessment of an active learning technique and comparisons with traditional lectures. *Innovative Higher Education, 21*(3), 231-246. doi: 10.1007/BF01243718

Fiore, S. M., & Salas, E. (2004). Why we need team cognition. In *Team cognition: Understanding the factors that drive process and performance* (pp. 235-248). Retrieved from http://tpl.ucf.edu/summit/ss_research/theory/Fiore_and_Salas_(2004)_TeamCog.pdf

Gardner, R., Godwin, S. A., Jay, G. D., Lindquist, D. G., Salas, E., Salisbury, M. L., & Shapiro, M. J. (2008). Defining team performance for simulation-based training: Methodology, metrics, and opportunities for emergency medicine. *Academic Emergency Medicine, 15*, 1088-1097. doi: 10.1111/j.1553-2712.2008.00251.x

Gordon, D. L., Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27*(1), 10-28. doi: 10.1080/01421590500046924

Hall, C. D., & Kuehster, C. R. (2010). Simulation: Learning from mistakes while building communication and teamwork. *Journal for Nurses in Staff Development, 26*(3), 123-127. doi: 10.1097/NND.0b013e3181993a95

Hamman, W. R. (2004). The complexity of team training: What we have learned from aviation and its applications to medicine. *Quality and Safety in Health Care, 13*, 72-79. doi: 10.1136/qshc.2004.009910

Hunt, E. C., Nelson, K. L., Shilkofski, N. A., & Stavroudis, T. A. (2007). Simulation: Translation to improved team performance. *Anesthesiology Clinics, 25*, 301-319. doi: 10.1016/j.anclin.2007.03.004

Krivonos, P. D. (2007, June). *Communication in aviation safety: Lessons learned and lessons required*. Paper presented at the meeting of Australia and New Zealand Societies of Air Safety Investigators, Wellington, Australia.

TOWARD A HUMAN PERFORMANCE STANDARD OF EXCELLENCE IN AIR TRAFFIC MANAGEMENT

Paul Krois, Federal Aviation Administration, Washington, D.C.

Damien Armenis, Airservices Australia, Brisbane, Australia

Rémi Joly, NAV CANADA, Ottawa, Canada

Barry Kirwan, EUROCONTROL, Bretigny-sur-Orge, France

Claire Marrison, Airservices Australia, Canberra, Australia

Neil May, NATS, London, United Kingdom

Dino Piccione, Federal Aviation Administration, Washington, D.C.

Michaela Schwarz, Austro Control GmbH, Vienna, Austria

Air Traffic Management (ATM) is a 24/7 industry that strongly depends on people and needs its frontline staff to be on top performance to maintain safety and efficiency of the air transport system. However, Air Navigation Service Providers (ANSPs) too often downplay the integration of human performance against higher priority operational and business issues. At the same time, human factors experts are sometimes challenged in communicating their tools and methods in ways that are seen as pertinent to ANSP issues. In order to bridge these organizational stove pipes, an international approach is being harmonized for ANSPs to gauge their maturity for how human performance is integrated across ATM system design, development and operation. A Human Performance Standard of Excellence (HPSoE) frames a business case to invest in human performance using three axes and associated assessment scales: Business Vision (appreciation of the role of human performance in the safe delivery of service), Human Performance (focusing on all job-related factors at individual, group, and organizational levels), and Human Factors (applying scientific knowledge to optimize human - system performance).

Introduction

Air Navigation Service Providers (ANSPs) around the world place high priority on ensuring and delivering safe and efficient Air Traffic Management (ATM) including Air Traffic Control (ATC) to the flying public. Because ATM is real-time, and incident evolution is typically measured in minutes, frontline staff (e.g., air traffic controllers, traffic management specialists, system technicians, maintainers, supervisors, and managers) need to exhibit peak performance around the clock. Yet the integration of human factors (HF) – the scientific discipline whose sole purpose is to enhance human performance – into ATM, when viewed globally, seems weak and patchy compared to other high-risk, high performance industries such as nuclear power or the defense domain. While some ANSPs have a strong HF capability supporting human performance, most do not and instead deal with human performance issues in other ways. But as ATM becomes ever-busier, more complex, and more inter-connected across different ANSPs, it is timely to consider how human performance is best optimized to continue ATM's reputation for smooth, efficient and safe handling of traffic.

In 2013, the Federal Aviation Administration (FAA), EUROCONTROL, and several ANSPs joined through Action Plan 15 to address and harmonize how the need to optimize safety and human performance can be supported by aligning and leveraging research advancements. One such effort is the development of a Human Performance Standard of Excellence (HPSoE) to assess and understand the

maturity of ANSPs for integrating human performance in ATM systems.[1] This paper describes the need for the HPSoE, identifies examples of industry capability maturity models (CMMs), and explains the approach being taken for the development of the HPSoE.

Why the Need for the HPSoE

Human performance refers to "the performance of jobs, tasks, and activities by operational personnel – individually and together" (EUROCONTROL/FAA, 2010). Human performance is important to ANSPs because people ensure that ATM service is kept safe and efficient for the flying public. In particular, ANSPs can use this HPSoE to establish a baseline upon which improvements can be identified in order to better manage operational safety risks and improve efficiency and resilience. The business case includes managing costs (e.g., insurance, borrowing) that can increase as incidents and accidents occur globally.

Why Human Factors Is Not Used

While human performance can be enhanced by applying human factors (HF) science, usage of HF experts in the ATM industry has generally been less than optimal. Operations and maintenance managers, as well as acquisition program managers, may want and plan for human performance excellence but they may find it necessary to limit how HF science is used, or find ways other than HF to try to reach their goals. This occurs when managers are under time or cost pressure or because they do not have the human performance intelligence to recognize when HF expertise is needed. This may also be due to ATM getting more complex and more dependent on automation, which brings with it tricky challenges like changes in job roles and balancing workload associated with use of new decision support tools that can introduce new sources of operational drift such as automation that is not used or used in ways not intended.

How to Avoid Having Human Factors Lost in Translation

A common complaint in many discussions between HF experts and decision makers is that the former do not speak to the issues of the latter, and the latter do not speak the HF lingo. Decision makers want to know things such as whether operations are safe and cost efficient, has training been effective, and will a new system being installed deliver better performance. In contrast, HF experts talk about human performance assurance through tools and methods like "training needs analysis" and "human centered automation guidelines" that present a different language for decision makers. The HPSoE recognizes that HF comprises a systems discipline and that HF experts need to connect with different parts of the organization. In fact, HF experts often view their efforts as a catalyst for interactions between organizational stove pipes because their expertise and tools contribute to teams in different departments.

How to Consistently Demonstrate the Benefits of Human Factors Integration

It is recognized that HF is by no means the whole answer but, without HF, ANSPs will be challenged to reach the most efficient and reliable levels of safe operations. An investment in HF is an investment in safety as well as efficiency. To accomplish this integration the ANSP needs a vision and a pathway to build that vision in order to achieve sustained improvement in human performance. Of course,

[1] Action Plan 15 (AP15) on Safety and Human Performance falls under the umbrella of the FAA-EUROCONTROL Memorandum of Cooperation, and is one of more than twenty different Action Plans. Since 2003, AP15 has focused on enhancing understanding of systemic safety issues, ranging from safety toolkits, to safety culture and resilience, to system-wide risk pictures and models, to HF. Like all Action Plans, the AP15 Terms of Reference are revised every three years.

the vision and capability to improve human performance must be tailored and proportionate to the size and complexity of each ANSP.

Capability Maturity Models

CMMs provide a framework for describing and assessing maturity of organizations and their use is well established in many industries. Within ATM, there already exists a CMM. A Standard of Excellence (SoE) for Safety Management Systems (SMS) was developed by the Civil Air Navigation Services Organization (CANSO) and provides an industry standard for gauging SMS maturity of ANSPs on five different levels. Guidance is provided on how improvements can be made to the SMS (CANSO, 2014). The HPSoE is designed to sit alongside and to complement the existing SoE for SMS. As with some CMMs, the SoE for SMS uses five levels to relate maturity and effectiveness, as shown in Figure 1.

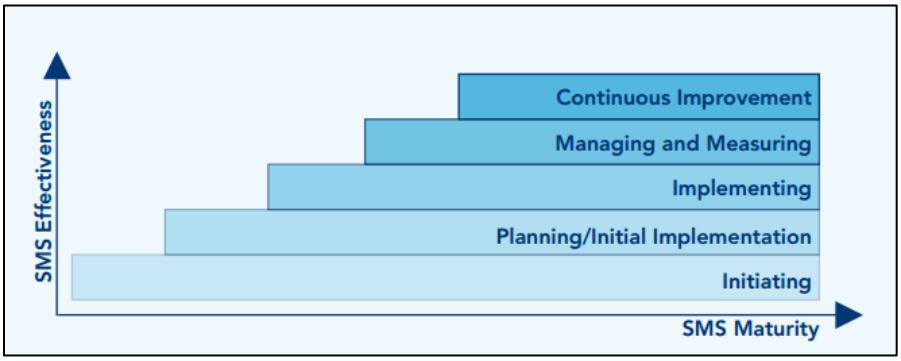


Figure 1. CANSO SoE for SMS Maturity Pathway.

In other industries, the People Capability Maturity Model (People CMM) developed by the Carnegie Mellon University is used by organizations to address their critical human capital issues (Curtis, Hefley & Miller, 2009). The People CMM adapts well-established maturity models for software development capability and uses a process maturity framework to align best practices for managing and developing an organization's workforce. The structure of the People CMM is summarized in Table 1.

Table 1.
Summary of the People CMM.

Level	Maturity	Characteristics
Five – Optimizing	Change Management	Continual improvements made in workforce practices and adoption of innovative technologies
Four – Predictable	Capability Management	Measures used to establish process performance baselines and assess priorities for improvements
Three – Defined	Competency Management	Workforce competencies tied to current and future business objectives as critical enablers
Two – Managed	People Management	Focus at unit level to overcome uneven skills, work overload, and poor communication
One – Initial	Inconsistent Management	Ad hoc practices, oriented toward administration rather than managing people

In the People CMM, an organization can transition from Levels One to Two by focusing on development of repeatable processes; from Two to Three by developing competency based practices; from Three to Four by using measured and delegated practices; and from Four to Five by continuously improving practices.

A CMM developed for the offshore oil and gas industry in the United Kingdom is the Human Factors Assessment Model (HFAM) (McLeod, 2004). HFAM intends to provide a practical and easy-to-use method to assess the maturity by which good practices are used for human issues associated with the design of new or changes in equipment and processes. HFAM uses various HF elements (e.g., roles and responsibilities, user involvement) with examples of possible evidence and weightings that are summed up and translated into an easy to interpret percentage score (see Table 2.).

Table 2.
Summary of HFAM.

Level	HFAM Score	Maturity
Five	91% or more	Best practice
Four	76-90%	Good practice achieved, towards best practice
Three	66-75%	Good practice
Two	46-65%	Some elements of good practice achieved, but not enough to be confident that it will be applied consistently (reasonable good practice)
One	45% or less	Definitely not following good practice

Human Factors Integration Strategy

Following review of CMMs from several industries it became apparent that a common HF integration strategy needed to be developed that would reach across the diversity of ANSPs. The strategy is depicted in Figure 2. The strategy leverages where and how improvements can be gained relative to both quick wins and for long haul efforts. The strategy enables each ANSP to assess how its capability aligns with its organizational vision and resources. Fortunately, the SoE for SMS in ATM (CANSO, 2014) provided a foundation of best practice with ANSPs. This included using multiple elements to construct the Standard with a phased approach to enable step wise implementation by ANSPs.

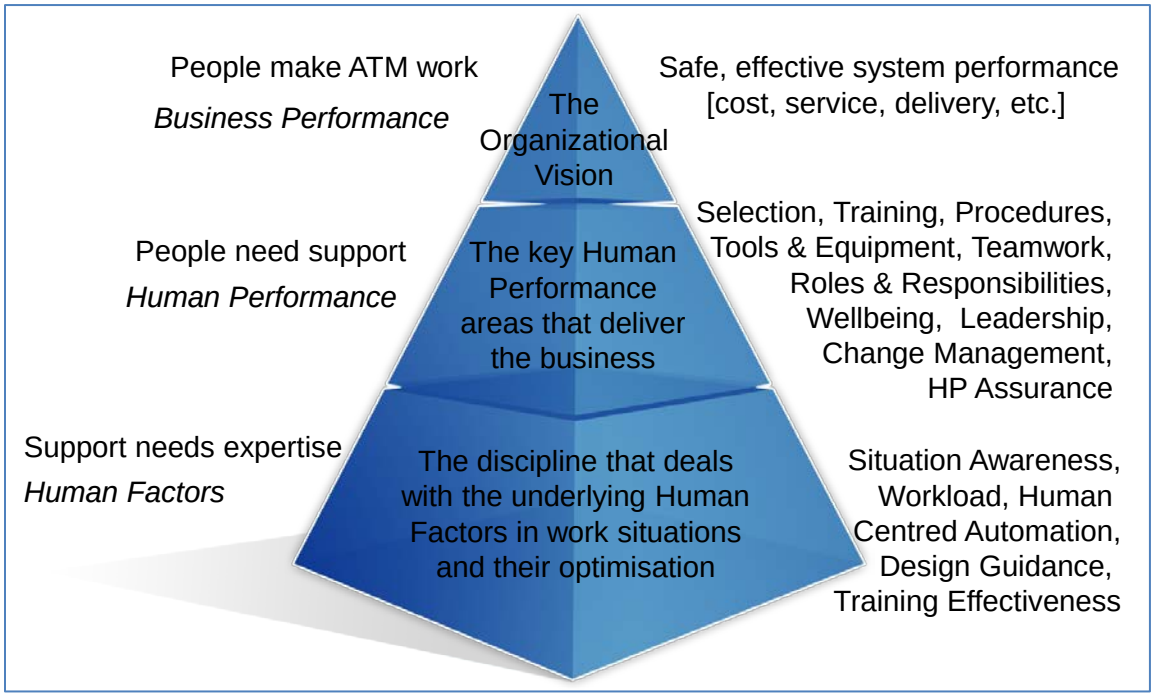


Figure 2.
Human Factors Integration Strategy as Foundation for the Development of the HPSoE.

Development of the HPSoE started with recognizing how HF experts currently contribute to ANSPs. This includes operating across existing organizational structures by being an important part of teams involved in operations, design, safety, training, engineering, and human resources. Through this perspective, three principal Axes emerged that characterize the contributions that HF make to these teams. Finally, these Axes were decomposed into thirteen Elements with assessment scales to gauge levels of maturity for how HF contribute to these teams and ANSPs. Together these Axes and Elements provide the framework for how ANSPs can rate themselves for maturity. At this point in time in development of the HPSoE, the three principal Axes with their thirteen Elements are shown in Figure 3.

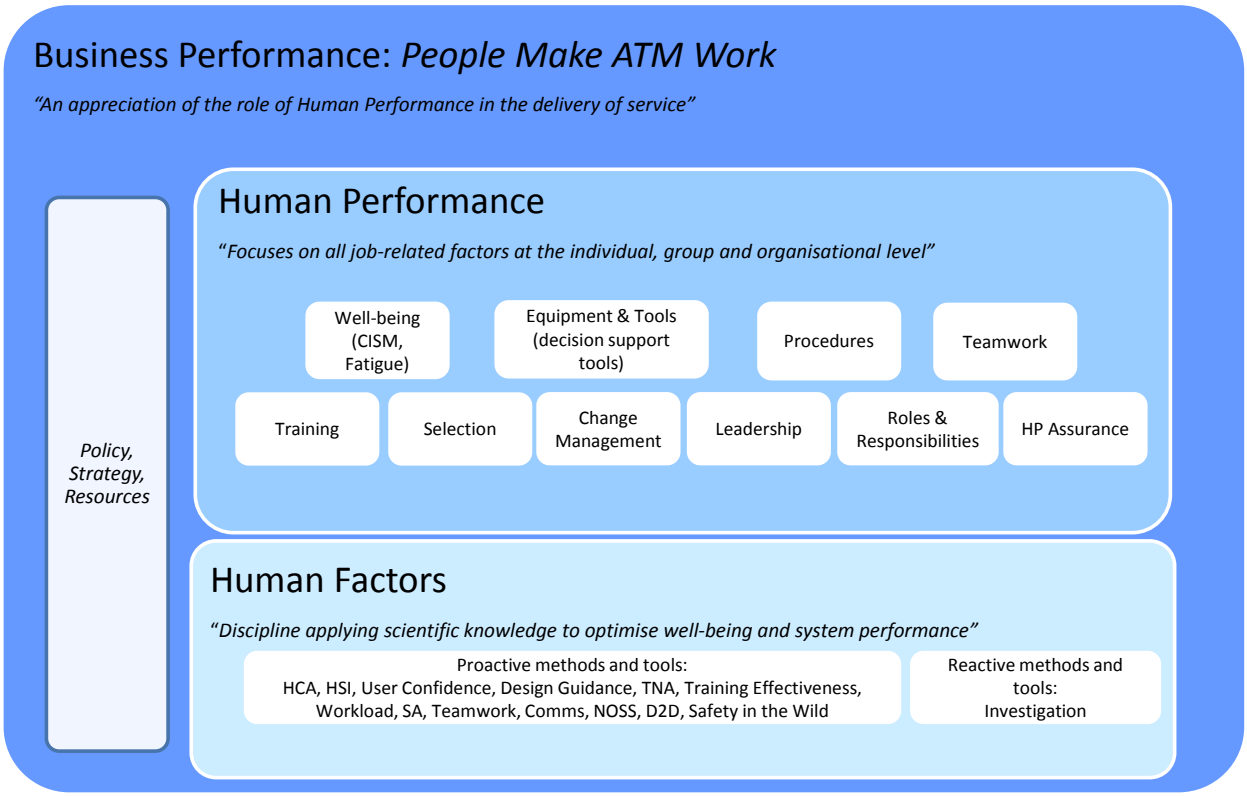


Figure 3.
Framework of Axes and Elements

An example of the assessment scales for Elements developed to assess ANSP maturity is shown in Table 3, for Policy, Strategy, and Resources.

Way Forward

Development of the HPSoE is accompanied with questions still needing to be addressed. This includes, for example, how does the HPSoE scale up or down with different sizes and complexities of ANSPs? What requirements or guidance is used as evidence of maturity assessments? What different paths can ANSPs use to step up to the next level of maturity? Does reaching a certain level of maturity infer comparability across ANSPs rated as having that level of maturity?

Conclusions and Outlook

The HPSoE can provide a benchmarking system to facilitate ANSPs seeking to improve Human Performance and so leverage the integration of HF in ATM system design, development and operation. It

provides a vision for how ANSPs can raise their maturity to better leverage the human contributions to operations, acquisitions, and maintenance of ATM systems.

Table 3.
Example Rating Scale for Element of Policy, Strategy, and Resources.

Objective	Initiating	Planning/Initial Implementation	Implementing	Managing & Measuring	Continuous Improvement
Provide a consistent and reliable level of Human Performance which ensures a safe and high quality level of service.	There is no recognition of the importance of the role that people play in delivering a safe and high quality level of service. The ANSP meets the minimum regulatory standards in respect of licensing, training, reporting, and so forth.	There is some recognition of the value that improving Human Performance can bring. The company has functions responsible for areas such as training, occupational health, and investigations. Initial planning is in place to improve Human Performance.	Human Performance is being actively improved. There is recognition of the value that HF expertise can bring. A person is identified with a clear remit and budget for addressing HF issues and they are embedded within a division of the organization.	Key Performance Indicators are in place to measure Human Performance and to identify priorities for improvement. The HF capability available is tailored and proportionate to the maturity and complexity of the ANSP. HF experts are operating within several divisions of the organization.	The role of the human is recognized as being integral to the success of the organization. A strategic vision is built around continuously improving the capability and performance of its people. The ANSP supports and uses HF R&D as a means of gaining intelligence on how to improve Human Performance.

Acknowledgement

The contents of this document reflect the views of the authors and do not necessarily reflect the views of their organizations.

References

Civil Air Navigation Services Organization (2014). CANSO Standard of Excellence in Safety Management Systems. Transpolis Schiphol Airport, The Netherlands.

Curtis, B., Hefley, B., & Miller, S. (2009). People Capability Maturity Model (P-CMM) Version 2.0, Second Edition. Technical Report CMU/SEI-2009-TR-003. Pittsburgh, PA: Carnegie Mellon University.

EUROCONTROL/FAA Action Plan 15 Safety (2010). Human Performance in Air Traffic Management Safety: A White Paper. Paris, France: EUROCONTROL.

McLeod, R. (2004). Human Factors Assessment Model Validation Study. Glasgow, United Kingdom: Health and Safety Executive, Offshore Safety Division.

IMPACT OF NEXTGEN ON NATIONAL AIRSPACE ACTORS

Kelley J. Krokos
American Institutes for Research
Washington, DC

Michael W. Sawyer & Katherine A. Berry
Fort Hill Group LLC
Washington, DC

The Federal Aviation Administration (FAA) is executing a transformation of the National Airspace System (NAS) through the implementation of the Next Generation Air Transportation System (NextGen). This paper presents two research efforts related to understanding and analyzing the effects of planned NextGen changes across NAS actors. American Institutes for Research is completing a Strategic Job Analysis and Strategic Training Needs Analysis of two NAS actors. The results are intended to provide recommendations to selection and training requirements necessary to support NextGen implementation. Fort Hill Group is building Human-System Interaction Models (HSIMs) that identify the human-system interactions affected by planned changes for individual and aggregated NextGen changes. The results are primarily being used to identify and mitigate safety risks as concepts are developed and implemented. These projects will provide the FAA with a comprehensive view of the impact of NextGen on NAS actors.

Strategic Job and Training Needs Analyses

A Strategic Job Analysis (SJA) is a future-oriented methodology designed to define a job as it will exist in the future. This methodology is often used to evaluate the impact of changes that are proposed to occur to a job. A Strategic Training Needs Analysis (STNA) is also a form of future-oriented evaluation designed to determine what training will be required to support employees working in that future job. In contrast to analyses designed to describe a current job that typically rely on current job incumbents for information, no employees are performing the future job; no incumbents exist. As a result, strategic analyses rely heavily on the input of experts who are involved in planning, designing, or deploying the proposed changes. The results of these types of analyses are estimates to the extent that they depend on stated plans for the future that are susceptible to changes in technology, funding, stakeholder priorities, and other disruptions.

However, these methodologies are extremely useful to organizations. First, the analyses often result in summaries of planned changes that may not have been available previously, such as when changes are being managed by different groups within a large organization. Second, the summaries may uncover new information about the changes such as risks and interdependencies. Finally, the information that results about the future job and the required training is useful for planning purposes. For example, if an SJA indicates that substantial changes will be required to the human abilities required to perform a job, the organization can plan for the changes that may be needed to the relevant human resource (HR) processes (e.g., recruitment, pre-employment selection test). Given the time and resources required to build and validate most HR systems, having this information well in advance of the changes is critical. The advance notice is also especially important in jobs where the consequence of error is high or whether the training pipeline is long. It was with these benefits in mind that the FAA funded the American Institutes for Research (AIR) to perform a series of strategic analyses to evaluate the impact of NextGen.

SJA and STNA for Controllers

Beginning in 2009, AIR conducted an SJA to evaluate the impact of NextGen on the job of Air Traffic Control Specialists, or controllers, and an STNA to estimate the training required to support the new job by the NextGen mid-term, which at that time was defined as 2018. For the SJA, AIR updated the current job analysis for controllers (i.e., tasks performed; knowledge, skills, abilities, and other personal characteristics (KSAOs) required of the people who perform the job; and the tools and equipment used). Next, AIR identified and described the NextGen technologies, automation, and procedures that the FAA plans to implement by 2018—or Drivers—and evaluated the impact of those Drivers on the controller job. For the STNA, AIR identified all the employee groups that would need to be trained on each Driver; the number of hours of training required; the proposed administration method (e.g., instructor-led training; simulation); and algorithms that can be used to estimate the resources required for each phase of the FAA's training process (i.e., design, development, implementation, evaluation, and maintenance).

The results of that research suggest that what controllers do will not change significantly by 2018, but that how they perform the job will change. As a result, AIR recommended that significant changes would not be needed to the FAA's current pre-employment selection test battery (the AT-SAT). However, significant changes were proposed to be required to the training program for controllers. AIR provided estimates of the training and the significant resources that would be required to support it. This research has been captured in myriad reports and publications (c.f., Baumann, Krokos, & Hendrickson, 2014). However, many aviation professionals contribute to the culture of safety that the traveling public enjoys today. That is, knowing the impact on controllers is critical but it is not enough; the FAA also needs information about the impact of NextGen on other FAA-employed professionals. Furthermore, they need this information in time to build or validate the human capital systems that support this workforce. Of particular interest are the estimated 6000 Airway Transportation Systems Specialists—or technicians—who maintain NAS systems (e.g., Aids to Navigation). Consequently, the FAA funded AIR to conduct similar future-oriented analyses on the job of technician.

SJA and STNA for Technicians

Technicians have a direct and critical responsibility for ensuring the safety of the traveling public. Like controllers, the consequence for error on this job is potentially catastrophic loss of life or property, and the training pipeline is long. Furthermore, the NextGen Drivers identified by AIR in its research on controllers suggested that technicians will also be influenced by NextGen. Consequently, the FAA funded AIR to conduct an SJA and an STNA to evaluate the impact of NextGen on field (i.e., bargaining unit) technicians by 2020 (current mid-term).

The process for conducting the SJA and STNA for technicians has been largely the same, to date, as for controllers. Although this research is in process, AIR has identified the NextGen Drivers that are proposed to affect technicians by 2020. AIR is currently updating the job analysis for how the technician job is currently performed. These two results will be synthesized and evaluated to determine the impact on the job and training of technicians by 2020. Despite the similarities in the SJA process, there are some noticeable differences. For example, although there is significant overlap in the lists of Drivers that will affect controllers and technicians in the mid-term, the lists do differ. For example, technicians install and maintain much of the hardware and software that controllers use. Some Drivers, such as 4-Dimensional Weather, include many hardware and software components that technicians must service, but fewer components with which controllers actually interface. In this case, technicians have more Drivers than controllers. On the other hand, some Drivers are procedures that do not require FAA-owned hardware or software. For example, controllers have to be taught new Performance-Based Navigation routes, but technicians have no role to play in the implementation of new routes.

Once the update to the job analysis of the current technician job is complete, AIR will complete the SJA by evaluating the impact of the identified Drivers on the current job. Then, those results will be used as the foundation for conducting the STNA. Collectively, the results will describe the future job as it is proposed to exist in 2020, and will provide information about the training required to prepare technicians to perform that job by 2020.

Update the SJA for Controllers

NextGen is an evolving initiative; changes in technology, funding, and priorities have had significant effects on various NextGen programs since AIR's first NextGen SJA and STNA were completed. Consequently, the FAA funded AIR to begin an update to its original controller SJA. Although the results are not final, preliminary results show that some previously-identified Drivers have been eliminated completely, while others have appeared on the list. For example, Flexible Airspace and High Altitude Airspace were identified as Drivers in AIR's previous research. However, these concepts are not currently being considered. Similarly, Unmanned Aircraft Systems (UAS) were not identified as Drivers in AIR's previous research but UAS has now been added as a Driver potentially having an impact on controllers by 2020. In addition to changes in the Drivers since the original research, the preliminary results also suggest that what controllers do in 2020 will not be significantly different than today. However, additional research will need to be conducted to evaluate the impact on how controllers perform their jobs by 2020.

Acknowledgements

AIR gratefully acknowledges the contractual, technical, and financial support of the Federal Aviation Administration's Human Factors Research and Engineering Division (ANG-C1). In addition, this research, and other projects like it, would not be possible without the significant contribution of technical expertise by NextGen experts, controllers, and technicians.

References

Baumann, E., Krokos, K. J., & Hendrickson, C. (2014). Building algorithms to estimate training resource requirements. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58,* 2335-2339. doi:10.1177/1541931214581486

NextGen Integrated Human-System Interaction Models

As new NextGen concepts are planned and developed, concept designers will need to consider the impact of these changes to the larger-scale interactions between new systems, procedures, and human operators. This is especially true given the concurrent nature of NextGen concept development and implementation. While many concepts consider the system dependencies and relationships necessary to ensure successful implementation, the impact to human-system interactions and relationships can be overlooked. The Human Systems Integration Roadmap provides a high-level view of the relationship between NextGen infrastructure deployment and National Airspace System (NAS) actors. A more in-depth analysis of individual planned changes is needed to develop specific recommendations for concept developers.

Planned NextGen changes primarily take the form of Operational Improvements (OIs). Each OI includes a description of the planned change, along with additional information on system relationships. While the OI descriptions provide a summary of each individual planned change, there is no direct way to demonstrate the cumulative impact of these changes on actors in the NAS. As such, Human System Interaction Models (HSIMs) have been developed to provide a consistent and scalable depiction of human-system interactions for proposed NextGen changes.

HSIM Development

NextGen HSIMs provides a baseline for graphically representing the NextGen impact to human-system interactions across NAS actors and systems. Each HSIM graphically depicts and describes the actor-actor or actor-system interactions associated with a proposed change. Each HSIM is composed of boxes representing different NAS actors and high-level NAS systems. Arrows represent the interactions between actor boxes and describe the interaction between those boxes. Figure 1 provides an overview of the HSIM data elements.

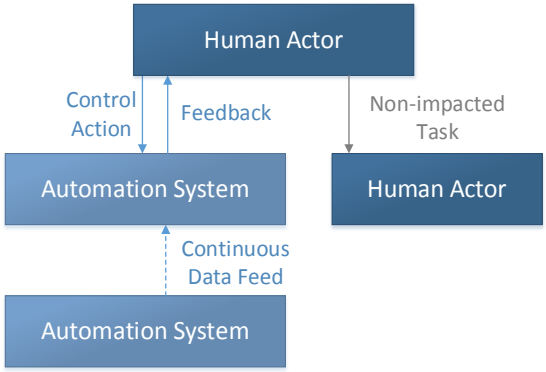


Figure 1. Human System Interaction Model Example

HSIMs are developed based on the text description of a NextGen OI and any additional information available on related systems. A team of air traffic control, commercial aviation, and human factors subject matter experts reviews each OI to first identify the controller and pilot interactions affected by the proposed change. Identified interactions are then used to develop HSIMs for each OI. Multiple HSIMs can then be combined to provide an integrated view of multiple proposed changes.

For example, the HSIM for a proposed radar conformance monitor would depict the en route automation system receiving surveillance and flight plan information and then display the alert to the en route sector controller. Following this, the model would show the controller identifying the alert on their

automation system and issuing instructions to the flight crew to return to their flight path. By utilizing a structured framework for describing human-system interactions, the individual models can be aggregated to show cumulative impacts across multiple proposed changes. Figure 2 provides excerpts from an individual OI HSIM and an integrated HSIM representing four OIs.

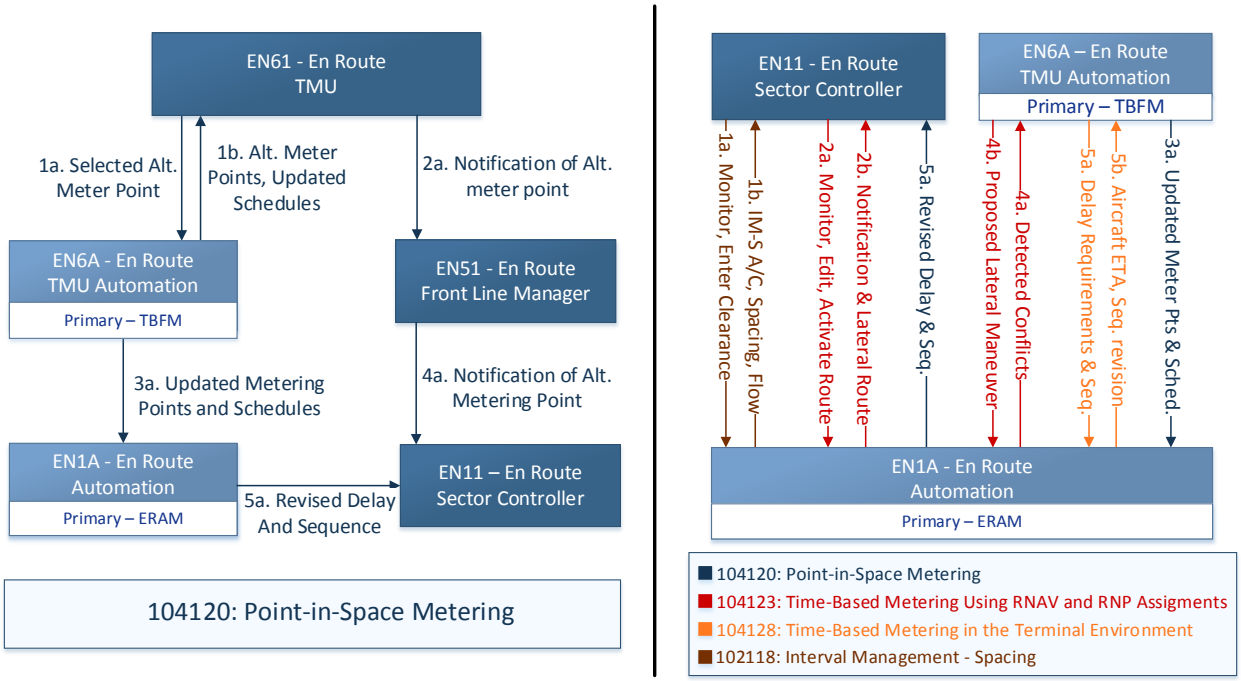


Figure 2. Individual HSIM (left) and Excerpt From Integrated HSIM (Right)

Current HSIM Status

NextGen HSIMs have been created for all planned NextGen changes scheduled for implementation between 2016 and 2020 where human-system interactions will be directly affected. These HSIMs serve as a foundational resource for understanding the impact of planned changes in terms of concept interactions, workload modeling, requirement elicitation, and integrated safety assessment.

Acknowledgements

We would like to acknowledge the Federal Aviation Administration's Human Factors Research and Engineering Division (ANG-C1) for funding this project and similar work. Additionally, we would like to acknowledge the air traffic control and human factors subject matter experts who provided the valuable insight necessary to develop these results.

References

Adams, J. A. (1979). On the evaluation of training devices. *Human Factors, 21*, 711-720.

Broach, D., Schroeder, D., & Joseph, K. (2000b). *Pilot age and accident rates report 4: An analysis of professional ATP and Commercial Pilot accident rates by age*. Oklahoma City, OK: FAA Civil Aerospace Medical Institute. Retrieved from http://www.cami.jccbi.gov/aam-400A/AGE60/age60_4.pdf

PERFORMANCE ASSESSMENT METHODS TO EVALUATE DISCRETIONARY ATC SAFETY STANDARDS

Julia Pounds
Federal Aviation Administration
Washington, DC

Lisa Galoci
Federal Aviation Administration
Washington, DC

The US Federal Aviation Administration (FAA) established the Air Traffic Safety Oversight Service (AOV) in 2005. Its mission is to provide independent oversight of the US air traffic service provider, Air Traffic Organization (ATO). AOV monitors ATO compliance with safety standards and safety management systems (SMS) including the ATO's policies, procedures, and practices. AOV based its surveillance activities on systematic auditing methods to monitor ATO compliance with required safety controls. However, audits could only be used to monitor required safety controls having objective (strong) evidence of performance available. AOV lacked methods to monitor discretionary (weak) safety controls or safety controls without objective evidence of performance (weak). By adapting data collection methods from other domains of human performance assessment, we developed methods for monitoring discretionary controls and all controls without objective evidence of performance. The systematic assessment method complements AOV's audit process so that more extensive oversight activities can be conducted. 150

The US Federal Aviation Administration (FAA) is responsible for civil aviation safety and provides for the safe and efficient use of the national airspace. The FAA established the Air Traffic Safety Oversight Service (AOV) in 2005 to be responsible for independent oversight of the US air traffic service provider, the Air Traffic Organization (ATO). To accomplish its mission AOV monitors ATO compliance with safety standards and safety management systems (SMS) including the ATO's safety controls executed through its policies, procedures, and practices. These are documented in ATO orders and other guidance documents and are accepted controls for known system hazards or potential hazards. SMS dictates continuous improvement and feedback. To support this, AOV regularly and systematically conducts surveillance of the ATO's performance by collecting safety-related data.

Air traffic control (ATC) is a highly proceduralized environment. Initially, AOV modeled its oversight activities after the FAA's regulation of the aviation industry (e.g., commercial passenger carriers) and airmen (i.e., pilots). It developed its procedures and materials for conducting compliance audits based on those from other regulatory organizations, such as the US General Accounting Office (GAO) and the various Offices of Inspectors General (OIGs). Ensuring acceptable performance is the responsibility of the ATO. Although the controls provided herein as examples use the performance of an air traffic controller, safety

controls are in place for performance at all levels of the ATO organization. AOV monitors the ATO's execution of this responsibility at all organizational levels.

To accomplish this, AOV uses systematic auditing methods to monitor ATO compliance with required safety controls. Audit results identify potential safety impacts of ATO policies, procedures, practices, and any changes proposed to them. Required controls are signaled by direct words like "must" and "will" such as, *Ground control must notify local control when a departing aircraft has been taxied to a runway other than one previously designated as active.*

AOV needed methods to conduct oversight of discretionary controls. Discretionary controls are easily identified because they are signaled by words like "may," "should," and "can" that give the performer more than one option in the situation. An example of a discretionary control is *Once the pilot informs you action is being taken to resolve the situation, you may discontinue the issuance of further alerts.*

Moreover, performance of both required and optional controls is sometimes modified by explanations with words like "timely" and "controller judgment." AOV needed unbiased methods to collect and report subjective performance data.

In sum, while AOV had the methodology for oversight of ATO's required safety controls, audit methods permitted only monitoring of compliance for required safety controls and only those with objective evidence of performance. AOV lacked methods to monitor (1) discretionary safety controls and (2) all safety controls where evidence of performance was subjective.

Method

Based on our general experience in research methods and in conducting requirements audits, we identified ATO directive documents containing the types of controls and performance that we could use to develop a basic assessment methodology.

Control Types

Using FAA Order 7110.65 *Air Traffic Control* as the example ATO directive for this activity, we examined how safety controls were written. We identified two types of safety controls in the ATO document:

(1) those required to be performed (these we labeled "strong" - as in strongly written so as to be mandatory) and

(2) those performed at one's discretion (these we labeled "weak" - as in weakly written so as to be optional).

Data Types

We identified two types of data to reflect performance: (1) objective evidence that a control was performed (e.g., Yes, it was completed; No, it was not completed) or (2) subjective evidence (e.g., Was it timely? Was good judgment used?).

To organize subjective data, descriptive survey methods were adopted to construct response scales, etc. (e.g., Babbie, 2007; Brace, 2004; Kumar, 1999). Similar to constructing a census when the count and distribution of a population is of interest, appropriate data collection instruments can collect frequencies and variability of performance.

These scales are constructed by the assessment team to reduce likelihood of a biased response tool. Data patterns can then be reported without the risk of interjecting personal opinion or judgmental bias during data collection. For example, "time" amounts can be measured without also decreeing a judgment about whether a response was "timely" or not.

Results

Methods for the two types of controls (strong control, weak control) were listed individually in the row and column labels. The methods for the two types of controls (strong, weak) were then crossed with methods for the two data types (strong data, weak data). Each cell then has an assigned (1) control type and (2) data collection method (Yes/No/NA or scaled). The fourth (cell A1) represents the existing compliance audit methodology. The resulting four quadrants and their associated methods in Table 1 define three new oversight methods shown. Cells A2, B1, and B2 give instructions for the type of control and data to be assessed. The matrix provides a tool to help assessment teams build their materials to conduct an assessment and suggested methods to report the results.

Table 1.
Matrix of Methods

		START HERE > The safety control of interest is …	
		A: STRONG CONTROL Compliant performance is mandatory or prohibited (e.g., must, must not) • Use checklist questions restating the requirement(s) for compliance.	**B: WEAK CONTROL** Performance is optional (e.g., may, need not). • Use survey items pre-defined by AOV to describe outcomes.
Then > Id data type.	**1: STRONG DATA** are clear objective, countable, measurable • Use methods for objective data (i.e., Yes, No) • Report results using count and percent. • May also include data range and distribution. (STRONG Evidence)	**A1: Conduct an AUDIT** Compliant performance is mandatory or prohibited. • **Materials**: Use checklist questions using requirement(s) to be audited. • **Data Collection**: Use methods for objective data (i.e., Yes, No; was/wasn't done, etc.). • **Results**: Report count, percent, range, and distribution.	**B1: Conduct an ASSESSMENT** Compliant performance is optional or described in vague terms. • **Materials**: Use survey items pre-defined by AOV to describe outcomes. • **Data Collection**: Use methods for objective data (i.e., Yes, No; was/wasn't done, etc.) • **Results**: Report count, percent, range and distribution.
	2: WEAK DATA are unclear subjective ambiguous, vague, unobservable • Use methods for unclear data (i.e., see AOV Job Aid for pre-defining variables and data). • Report results using count, percent, range, and distribution. (WEAK Evidence)	**A2: Conduct an ASSESSMENT** Compliant performance is defined as mandatory or prohibited. • **Materials**: Use checklist questions using requirement being assessed. • **Data Collection**: Use methods for unclear data with AOV pre-defined variables. • **Results**: Report count, percent, range and distribution.	**B2: Conduct an ASSESSMENT** Compliant performance is defined as optional or in vague terms. • **Materials**: Use survey items pre-defined by AOV to describe outcomes. • **Data Collection**: Use methods for unclear data with AOV pre-defined variables. • **Results**: Report count, percent, range and distribution.

Conclusions

By identifying two types of controls and adapting data collection methods from other domains of human performance assessment, we developed methods for monitoring discretionary controls and all controls without objective evidence of performance. The systematic assessment method complements AOV's audit process so that more extensive oversight activities can be conducted. Results of assessments may be useful to identify controls which need to be (a) strengthened, (b) added when a gap in hazard control is identified, or (c) removed because they add an unnecessary complication to the safe provision of ATC services. Distributions of times and performance variance (for example) across the same situations can lead to productive discussions about safety and risk in the provision of services.

References

Babbie, E. R. 2007. *Basics of Social Research* (4th ed.). Belmont, CA: Wadsworth.
Brace, I. 2004. *Questionnaire Design*. Sterling, VA: Kogan.
Kumar, R. 1999. *Research Methodology*. Thousand Oaks, CA: Sage.

DEVELOPING QUANTITATIVE AIR TRAFFIC RISK-BENEFIT PATHWAYS FOR CLASS DELTA AIRPORTS: IMPROVING SMALL TOWER OPERATIONS

Katherine A. Berry, Michael W. Sawyer, and Jordan Hinson
Fort Hill Group, LLC
Washington, DC

The primary responsibility of an Airport Traffic Control Tower (ATCT) controller is to prevent collisions between aircraft and other hazards on the surface and in the immediate vicinity. The safety service provided by controllers at towers with larger operations greatly exceeds the costs of establishing those towers. As the number of operations decreases, the costs of operating the tower may begin to outweigh the benefits of staffing the tower. Safety event reports describing instances where an ATCT controller provided a service that reduced the consequences of the event were collected. The reports were classified to identify latent factors, causal factors, and positive safety benefits. The adverse causal factors and positive safety benefits were then utilized to determined statistically significant risk-benefit pathways describing the safety benefits that controllers provide at airports in Class Delta (D) airspace. This paper presents the dynamic risk-benefit pathway, one of the three pathways for Class D ATCT.

ATCTs and the controllers that staff them provide both efficiency and safety services to the aviation industry. The primary responsibility of an ATCT controller is to prevent collisions between aircraft and other hazards (e.g., terrain, ground vehicles) on the airport surface and in the immediate vicinity of the airport (FAA, 2012). Set in 1990, the Office of Policy and Plans (APO) developed criteria for the establishment and discontinuance of ATCT (FAA-APO-90-7) (FAA, 1990). However, operations in the National Airspace System (NAS) have and are continuing to transition to support Next Generation Air Transportation System (NextGen) initiatives and other enhancements to the NAS. The Federal Aviation Administration's (FAA's) APO is reviewing and potentially updating the cost, safety benefit, and efficiency benefit criteria outlined in the 1990 policy for ATCT establishment; the focus of the review is on low volume tower operations, such as airports in Class D airspace. In examining the safety benefit of ATCT controllers, the safety service provided by tower controllers at towers with larger operations, such as the Core 30 airports, greatly exceeds the costs of establishing those towers. Controllers at larger operation towers are necessary to efficiently and safely manage air traffic. However, as the number of operations at a tower decreases, the costs of operating the tower may begin to outweigh the benefits.

Prior internal research of Class D airports identified hazards and classified those hazards for towered airports in Class D airspace. The impact those airport characteristics have on operations and controller performance has yet to be fully examined. With the focus on visual air traffic services (VATS), the purpose of this study is to assess the operational safety benefit provided by tower controllers in Class D airspace and to determine the potential safety benefit that a controller could have provided during safety events in non-towered operations. As part of the

larger project (Berry, Sawyer, & Hinson, 2014), this paper presents the safety benefits and associated risks with the previously identified hazards representing dynamic hazards.

Methodology

For the safety benefits assessment of VATS operations, a sample of 35 FAA towered airports in Class D airspace was identified. Utilizing a previous FAA study, the airport characteristics were identified for each of the airports in the sample set. Narrative safety data for the airport sample set was gathered from the FAA's Air Traffic Safety Action Program (ATSAP). ATSAP is a voluntary, non-punitive reporting system for air traffic controllers. ATSAP reports submitted by controllers at the sample airports for the calendar years of 2011, 2012, and 2013 time period were queried, resulting in 792 reports and safety event narratives. The focus of the ATSAP program is to provide the air traffic community an outlet for reporting a safety event that might otherwise have gone unknown. The purpose of this analysis is to examine the safety benefits that controllers provide in the control tower environment. The 792 ATSAP reports were filtered to identify those reports describing a safety event where the controller provided a safety benefit. The question examined in the filtering exercise was, "Did the controller provide a service that reduced the severity or consequences of the safety event described in the report?" Each of the 792 ATSAP reports were examined with the question by at least two human factors subject matter experts (SMEs), resulting in 175 ATSAP reports identified as describing a safety event where a controller provided a safety benefit.

Classification of Benefits and Risks

The filtered 175 ATSAP reports were classified with the Air Traffic Analysis and Classification System (AirTracs) utilizing the consensus method, which required a consensus or agreement on the causal factors contributing to the report by a panel. The panel members included human factors experts, retired air traffic controllers, and flight deck experts. AirTracs provides a framework for systematically and thoroughly examining the impact of human performance on air traffic accidents and incidents. The framework of the AirTracs causal category model is based on the Department of Defense (DoD) Human Factors Analysis and Classification System (HFACS) model (DoD, 2005), while the detailed causal factors incorporate factors from Human Error in ATM (HERA) and JANUS (Isaac et al., 2003). The AirTracs framework promotes the identification of causal trends by allowing factors ranging from the immediate operator context to agency-wide influences to be traced to individual events. The causal category model is displayed in Figure 1. For more information on the AirTracs causal factor categories see Berry, Sawyer, & Austrian, 2012.

To determine the risks or latent factors present, each report was evaluated across all levels of the AirTracs framework, and the presence or absence of each AirTracs causal category was recorded. It is important to note that the AirTracs categories are not mutually exclusive. For example, an individual report can include both an execution act and a decision act. To determine the safety benefits present, each safety benefit was classified with the FAA's strategic job analysis for the tower domain (AIR, 2011). In order to identify risk-benefit pathways, associations among AirTracs factors and safety benefit tasks were measured. Starting at the highest AirTracs tier and continuing to the lowest AirTracs tier, the relationship among the factors within the tier, the various factors at lower tiers, the strategic job tasks, and airport characteristics were examined using a Pearson's chi-square test to measure the statistical strength of the association. In the instances where the assumptions of the Pearson's chi-square test were

not met, a Fisher's Exact Test was conducted (Sheskin, 2011). If the relationship resulted in a significant association identified through the Pearson's chi-square test or Fisher's Exact Test ($p<0.05$), the odds ratio value was calculated for that particular association (Sheskin, 2011).

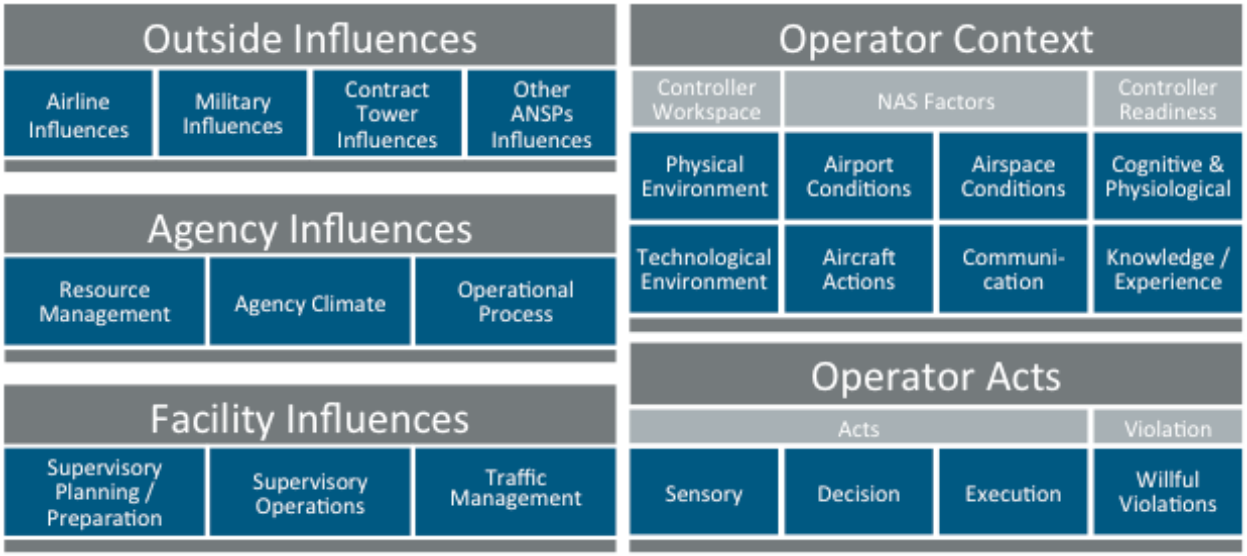


Figure 1. AirTracs Framework

Results and Discussion

When examining the safety benefits that tower controllers provide at the sample set of FAA staffed towered airports in Class D airspace, the three following human factors safety-benefit pathways emerged: Dynamic Risk-Benefit Pathway, Static Risk-Benefit Pathway, and Communication Risk-Benefit Pathway. The human factors safety-benefit pathways represent key associations among AirTracs factors, safety-critical tasks, and airport characteristics. This paper will present the findings for the Dynamic Risk-Benefit Pathway. The first human factors-safety risk-benefit pathway incorporates how a controller at a Class D towered airport provided a safety-benefit service to mitigate a dynamic risk and can be found in Figure 2.

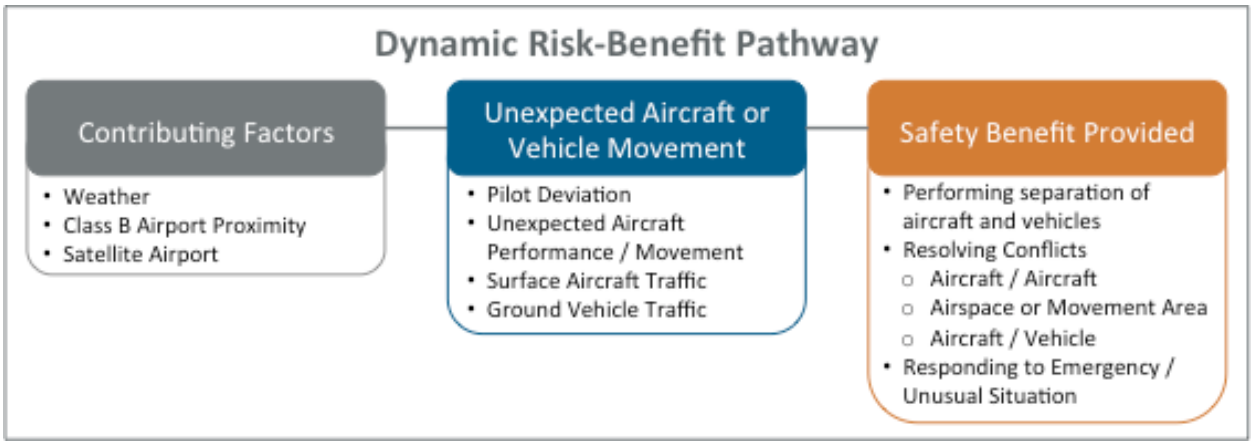


Figure 2. Dynamic Risk-Benefit Pathway

The central blue box in the risk-benefit pathway graphic depicts the AirTracs factors that presented a key risk to operations in the ATSAP reports. For this pathway, the factors were

dynamic in nature as they were a result of human actions and were not consistently present at all airports in every situation. Those dynamic risk factors were found to be pilot deviations, unexpected aircraft performance/movement, airport surface aircraft traffic, and ground vehicle traffic. Table 1 shows the level of classification for each risk factor. The values in Table 1 can be interpreted in the following way: in 62.26% of the ATSAP reports classified, there was a pilot deviation. In most cases, risk factors represent active pilot or driver errors or failures, and it is necessary to examine the latent factors associated with those risk factors to better understand why those risk factors may occur.

Table 1. *Risk Factor Classification Level – Dynamic Risk-Benefit Pathway*

Risk Factor	Percentage of Classified Reports
Pilot Deviation	62.26% of ATSAP Reports
Unexpected Aircraft Performance/Movement	25.47% of ATSAP Reports
Airport Surface Aircraft Traffic	10.38% of ATSAP Reports
Ground Vehicle Traffic	6.60% of ATSAP Reports

The left gray box in the risk-benefit pathway graphic depicts the contributing factors associated with the risk factors. The contributing factors represent a combination of the airport characteristics previously identified (e.g., Class B Airport Proximity) and contributing factors from the application of AirTracs (e.g., weather). Those contributing factors found to be associated with the dynamic risk factors were weather, Class B airport proximity, and satellite airports. Table 2 shows the level of classification for each contributing factor. The values in Table 2 are represented in one of two manners: 1) For airport characteristics, 34.29% of the sampled towered airports are in proximity to a Class B airport; 2) For AirTracs factors, in 7.55% of the ATSAP reports classified, weather was a contributing factor.

Table 2. *Contributing Factor Classification Level – Dynamic Risk-Benefit Pathway*

Contributing Factor	Percentage of Classified Reports or Airports
Satellite Airports	42.86% of the Sampled Towered Airports
Class B Airport Proximity	34.29% of the Sampled Towered Airports
Weather	7.55% of ATSAP Reports

In order for the contributing factor to be included in the pathway, at least one of the contributing factors had to have a statistical association with at least one of the risk factors. Table 3 depicts the associations and their odds ratios. For those pairings with odds ratios, the pairing was first found to be statistically significant via the Pearson's Chi Square test or Fisher's Exact Test ($p < 0.05$). Upon being found significant, the odds ratio for the pairing was determined. The odds ratio can be interpreted in the following way: when a report was found to include weather as a contributing factor, the odds of the report also including unexpected aircraft performance/ movement were 5.758 times greater than those reports that did not indicate weather as a factor.

Table 3. *Contributing Factors – Risk Factors Associations Odds Ratios – Dynamic Risk-Benefit Pathway*

	Risk Factors	
Contributing Factors	Pilot Deviation	Unexpected Aircraft Performance/Movement
Satellite Airports	2.285	2.526
Class B Airport Proximity		2.526
Weather		5.758

The right orange box in the risk-benefit pathway graphic depicts the safety benefits provided by a controller through safety-critical tasks. These safety-critical tasks depict how a controller identified, responded to, and recovered from the dynamic risks. For the dynamic risk-benefit pathway, the safety benefits provided by tower controllers include performing separation of aircraft and vehicles, resolving conflicts, and responding to emergencies/unusual situations. Table 4 showing the level of classification for each benefit. The values in Table 4 can be interpreted in the following way: in 37.74% of the ATSAP reports classified, a controller performed safety-critical tasks related to resolving aircraft to aircraft conflicts.

Table 4. *Contributing Factor Classification Level – Dynamic Risk-Benefit Pathway*

Benefit Factor	Percentage of Reports
Resolving Conflicts – Airspace or Movement Area	40.57% of ATSAP Reports
Resolving Conflicts – Aircraft/Aircraft	37.74% of ATSAP Reports
Responding to Emergencies/Unusual Situations	17.92% of ATSAP Reports
Resolving Conflicts – Aircraft/Vehicle	12.26% of ATSAP Reports
Performing Separation of Aircraft and Vehicles	8.49 % of ATSAP Reports

In order for the safety benefit to be included in the pathway, at least one of the risk factors had to have a statically significant association with at least one of the safety benefits. Table 5 depicts the associations and their odds ratios. For those pairings with odds ratios, the pairing was first found to be statistically significant via the Pearson's Chi Square test or Fisher's Exact Test ($p < 0.05$). Upon being found significant, the odds ratio for the pairing was determined. The odds ratio can be interpreted in the following way: when a report was found to include a pilot deviation as a risk factor, the odds of the report including the safety-benefit tasks associated with resolving aircraft to aircraft conflicts were 2.50 times greater than those reports not including a pilot deviation.

Table 5. *Risk Factors Safety Benefit Associations Odds Ratios – Dynamic Risk-Benefit Pathway*

	Risk Factor			
Safety Benefit	Pilot Deviation	Unexpected Aircraft Performance/ Movement	Airport Surface Aircraft Traffic	Ground Vehicle Traffic
Performing Separation of Aircraft and Vehicles			5.56	
Resolving Conflicts	9.05	3.19		
Resolving Conflicts – Aircraft/Aircraft	2.50		21.67	
Resolving Conflicts – Airspace or Movement Area	5.66	3.08		
Resolving Conflicts – Aircraft/Vehicle				78.86
Responding to Emergencies/Unusual Situations	14.00	6.10		

Acknowledgements

We would like to acknowledge the FAA's Human Factors Research and Engineering Division (ANG-C1) for funding this project and similar work. Additionally, we would like to acknowledge the air traffic control and human factors subject matter experts who provided the valuable insight necessary to develop these results. The results presented herein represent the results of this research project and do not necessarily represent the view of the FAA.

References

American Institutes for Research. (2011). *Job Description for the NextGen Mid-Term ARTCC Controller*. Retrieved 2011 from https://www2.hf.faa.gov/HFPortalNew/.

Berry, K., Sawyer, M., & Austrian, E. (2012). AirTracs: The development and application of an air traffic safety taxonomy for trends analysis. In *Proceedings of the 1st Annual Conference on Interdisciplinary Science for Air Traffic Management*, Daytona Beach, FL.

Berry, K., Sawyer, M., & Hinson, J. (2014). *Controller Safety Benefits in Low Volume Tower Operations: A Human Factors-Safety Assessment*. Retrieved 2014 from https://www.hf.faa.gov/hfportalnew/admin/FAAAJP61/Controller%20Safety%20Benefit%20in%20Low%20Volume%20Tower%20Operations.pdf

Department of Defense (2005). DoD HFACS: a mishap investigation and data analysis tool. Retrieved 2011 from http://www.public.navy.mil/navsafecen/Documents/aviation/aeromedical/DoD_hfacs.pdf

Isaac, A., Shorrock, S.T., Kennedy, R., Kirwan, B., Anderson, H., & Bove, T. (2003). *The human error in ATM technique (HERA-JANUS).* (EUROCONTROL Doc HRS/HSP-002-REP-03).

FAA. (1990). *Establishment and Discontinuance Criteria for Airport Traffic Control Towers.* (FAA-APO-90-7).

FAA. (2012). *Air Traffic Control* (JO 7110.65U). Retrieved 2014 from http://www.faa.gov/documentLibrary/media/Order/ATC.pdf

FAA. (2013). *Safety Benefits of Air Traffic Control Towers: Part 1. An Evaluation of Hazards at Towered Airports in Class D Airspace.*

FAA. (2014). *Visual Air Traffic Control Towers: A Review of Fatal Accidents, 2003 – 2012*.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton: Chapman & Hall.

THE ROLE OF PERSONNEL SELECTION IN REMOTELY PILOTED AIRCRAFT HUMAN SYSTEM INTEGRATION

Thomas R. Carretta
Air Force Research Laboratory
Wright-Patterson AFB, OH

Raymond E. King
FAA Civil Aerospace Medical Institute
Oklahoma City, OK

Effective human-system integration (HSI) incorporates several domains: manpower, personnel, and training, human factors, environment, safety, occupational health, habitability, survivability, logistics, intelligence, mobility, and command and control. These domains are interdependent and must be considered in terms of their interrelationships. Human factors engineers typically focus on system design with little attention to the skills, abilities, and other characteristics needed by the human operator. Personnel selection is seldom considered during the HSI process. Complex systems require careful selection of the individuals who will interact with the system. Selection is a two-stage process: *Select-in* procedures determine who has the aptitude to profit from a training program and, thus, represents the best investment. *Select-out* procedures focus on medical qualification and disqualification. Generally, less expensive screenings methods should be used first and more expensive filters used in later stages of the selection process. Personnel selection has a vital role to play in human-system integration.

Achieving high levels of effectiveness for complex systems such as remotely piloted aircraft (RPA) cannot be done solely through technological advances. Systems such as RPA consist of hardware, software, and human personnel which must effectively work together to achieve organizational objectives. Human-systems integration (HSI) is a comprehensive management and technical approach to address the role of human operators in system development and acquisition. HSI incorporates several domains including manpower, personnel, and training, human factors, environment, safety, occupational health, habitability, survivability, logistics, intelligence, mobility, and command and control (United States Air Force, 2014). These domains are interdependent and their interrelationships must be considered. HSI must be considered early in the system development and acquisition process to be effective. It is difficult and costly, if not impossible, to "fix" a poorly designed complex system after it has been built and implemented. This paper discussed the role of personnel measurement and selection for HSI, the development of Undergraduate RPA training (URT) selection standards, other important considerations in personnel selection, and expected changes in selection requirements as RPAs evolve.

Role of Personnel Measurement and Selection for HSI

Those responsible for human-system integration should be aware of the relations between selection, training, and human-system design and how they interact to affect overall system effectiveness. Poor personnel measurement and selection will result in higher training attrition and training costs, increased human-system integration costs, and lower levels of job performance. Poor training will require higher quality applicants and improved human-systems design to mitigate its effects. Poor human factors (i.e., clumsy automation, operator-vehicle interface design) will increase operator cognitive demands and workload, resulting in increased selection and training requirements. Effective selection (Carretta & Ree, 2003) and training (Patrick, 2003; Smallwood & Fraser, 1995) methods and human-automation interaction (Paraduraman & Byrne, 2003) can help reduce life cycle costs and contribute to improving organizational effectiveness.

The Development of Undergraduate RPA Training (URT) Selection Standards

US Air Force (USAF) RPA Pilot Selection Methods

In the USAF, early efforts to field RPA systems focused on technology development. Personnel selection, training, and human-interface design were given little attention, as the RPA system manning approach was to retrain manned aircraft pilots to operate RPAs. Although this approach was mostly effective, as demand for the capabilities provided by RPAs increased, it became too costly and unsustainable. In 2009, an Undergraduate RPA training

(URT) program was established to train personnel with no prior flying experience to operate RPAs. URT curricula were developed and selection requirements based on those for manned aircraft pilot training were established.

URT selection methods involve both select-in-and select-out procedures and are very similar to those for manned aircraft pilot training. Aptitude testing and Medical Flight Screening are two important factors. Aptitude testing includes the Air Force Officer Qualifying Test (AFOQT; Drasgow, Nye, Carretta, & Ree, 2010), Test of Basic Aviation Skills (TBAS; Carretta, 2005), and Pilot Candidate Selection Method (PCSM; Carretta, 2011). Aptitude requirements for URT qualification are identical to those for manned aircraft pilot training. Medical Flight Screening (MFS) includes successful completion of a FAA Class III Medical Certificate and an USAF Flying Class IIU Medical Examination (United States Air Force, 2011), review of medical records, psychological testing, and an interview. Results from the MFS psychological testing and interview are not used as part of a select-out process with strict minimum qualifying scores. Rather, a licensed psychologist uses clinical judgment to assess the psychological disposition of URT applicants to determine whether there is an aeromedically disqualifying condition in accordance with Air Force guidelines (United States Air Force, 2011). Results of two recent USAF predictive validation studies (Carretta, 2013; Rose, Barron, Carretta, Arnold, & Howse, 2014) have demonstrated similar levels of validity for the AFOQT Pilot and PCSM composites to those observed for manned aircraft pilot training.

Results for studies examining the utility of personality for URT are less consistent (Chappelle, McDonald, Heaton, Thompson, & Hanes, 2012; Rose et al., 2014). Chappelle et al. examined the predictive validity of the AFOQT Pilot composite, Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992), and a neuropsychological battery, the MicroCog (Powell, Kaplan, Whitla, Weintraub, Catlin, & Funkenstein, 2004) versus URT completion. The best-weighted regression composite for predicting URT completion included the AFOQT Pilot composite, several NEO-PI-R scales, and the MicroCog Reaction Time subtest. Discriminant analyses showed that personality scores improved classification accuracy (identification of true positives and true negatives) beyond that provided by cognitive ability and prior flight time. Classification accuracy improved from 57.1% to 75.2% when personality scores were included; however, these results likely capitalized on chance given the large number of NEO-PI-R scales relative to the small sample size.

Rose et al. (2014) examined the extent to which scores from a Big Five measure of personality could improve prediction of URT completion and training grades beyond the AFOQT Pilot and PCSM composites. Regression analyses showed no incremental validity for personality scores when used in combination with the AFOQT Pilot or PCSM composite scores for predicting URT completion. However, the Openness score demonstrated small, but statistically significant incremental validity for predicting initial RPA qualification training grade.

RPA System Job/Task Analyses

Despite the predictive validity of current RPA pilot training selection methods, several studies have been conducted to determine whether there are any unique job-related skills, abilities, and other characteristics (SAOCs) not adequately measured by current selection methods (for a summary, see Carretta, Rose, & Bruskiewicz, in press; Williams, Carretta, Kirkendall, Barron, Stewart, & Rose, 2014). In the Williams et al. study, Air Force, Army, and Navy subject matter experts in personnel measurement, selection, and testing identified and assigned importance ratings to 115 SAOCS that appeared in one or more military RPA job/task analyses. Where available, psychometric data for existing DoD and US Military Service proprietary personnel selection and classification tests were examined to determine the extent to which the tests measure critical RPA SAOCs and to identify measurement gaps. Seventy-eight of the 115 SAOCs received an average rating of 3 (moderately important) or higher on a 5-point scale. Of these, 57 of 78 (73%) were judged to be measured by one or more existing military proprietary tests. It is interesting to note that many of the most important SAOCs involved personality (e.g., conscientiousness, stress management, dependability, vigilance, adaptability/flexibility, integrity, responsibility, self-discipline). Table 1 provides examples of the highest-rated cognitive, personality/temperament, and other characteristics. See Williams et al. for the complete list of SAOCs.

Williams et al. (2014) made several recommendations for RPA operator test battery content. As previously noted, most of the critical SAOCs were judged to be measured by existing proprietary DoD or US Military Service tests. They recommended that a program be established to increase the reliability and reduce the fakeability of military personality tests. They also recommended that new tests be developed to fill measurement gaps (e.g., oral comprehension, vigilance) and to improve experimental measures involving task prioritization/multi-tasking and work preferences (person-environment fit).

Table 1.

Examples of SAOCS Rated Most important for RPA Pilots

Cognitive	Personality/Temperament	Other
Task Prioritization	Conscientiousness	Time Sharing
Oral Comprehension	Stress Management/Tolerance	Control Precision
Spatial Orientation	Dependability	Occupational Interests/ Work Preferences, P-E Fit
Oral Expression	Vigilance (ability & personality)	
Attention to Detail	Adaptability/Flexibility	
Critical Thinking	Responsibility	
	Self-Discipline	

Other Important Considerations in Personnel Selection

The Criterion

Many researchers spend enormous amounts of effort to develop measures of critical SAOCs based on the results of job/task analyses. They then search for available, convenient, or easy to collect job performance criteria with little thought about the theoretical meaning or psychometric properties of the criteria. The same care used to develop personnel selection methods and predictors of job performance should go into the development of job performance criteria. Failure to consider the psychometric properties of the criterion (e.g., construct validity, dimensionality, discriminability, reliability) leads to incorrect decisions about the effectiveness of selection methods and their relation to job performance. Problems also are caused by inattention to relevancy, contamination, and deficiency of the criterion.

As with measures used for personnel selection, job performance criteria vary in the constructs they measure, content, and specificity. To the extent possible, the constructs assessed by the job performance criteria should match those measured by the selection measures. As we have discussed, RPA job/task analyses have identified several critical personality traits needed for success. However, predictive validation studies have shown relatively low validities for personality compared to cognitive and other measures. One reason for this finding may be the job performance criteria used in these studies do not capture constructs for which personality is important (e.g., effort, leadership, indicators of maladaptive or counterproductive behavior). McHenry, Hough, Toquam, Hanson, and Ashworth (1990) provided an example that demonstrates the importance of criterion specificity. McHenry et al. administered a large battery of measures including ability and personality/temperament to a sample of U. S. Army trainees. Multiple criteria were used to reflect different aspects of job performance. Cognitive tests were the best predictors of criteria reflecting technical job proficiency, while measures of personality/temperament were the best predictors of criteria reflecting effort and leadership.

Special Population Norms

The assessment of human characteristics is based on comparing an individual to a representative sample of the population. Certain segments of the population vary significantly from the general population. For example, groups may differ on level of academic achievement, physical fitness, job experience, specialized knowledge/training, or other factors related to occupational performance. Moreover, differences in personality across occupational groups such as sales personnel, pilots, and engineers may occur. Military aircrew personnel are a highly selected and distinguished occupational group. Competition for pilot training assignments is great with the result that those selected differ significantly from the general adult population on cognitive, personality, and other characteristics considered during the selection process. Carretta et al. (2014) reported cognitive and personality norms for large samples of USAF pilot trainees. They observed that the mean full-scale IQ score for this group (M = 120) was about 1.33 SDs above the normative adult population mean (M = 100, SD = 15)). A pilot with a mean full-scale IQ of 105 would be slightly above the normative adult population mean, but well below the mean for USAF pilot trainees.

Significant differences also have been observed for personality scores of USAF pilot trainees compared to adult population norms. The personality portion of the USAF Neuropsychiatrically Enhanced Flight Screening (King &

Flynn, 1995) program, the forerunner of MFS, was developed to compile special population norms. The battery has been composed of the 1) Armstrong Laboratory Aviation Personality Survey (Retzlaff, Callister, & King, 1997) and 2) NEO Personality Inventory-Revised (Costa & McCrae, 1989). The ALAPS measures personality, psychopathology, and crew interaction, while the NEO-PI-R measures the Big Five domains and facets of normal personality. Figure 1 illustrates the number of standard deviations USAF pilot normative means are above or below those for the adult general population. Similar specialized norms are not presented for the ALAPS because it was normed on a USAF student pilot sample.

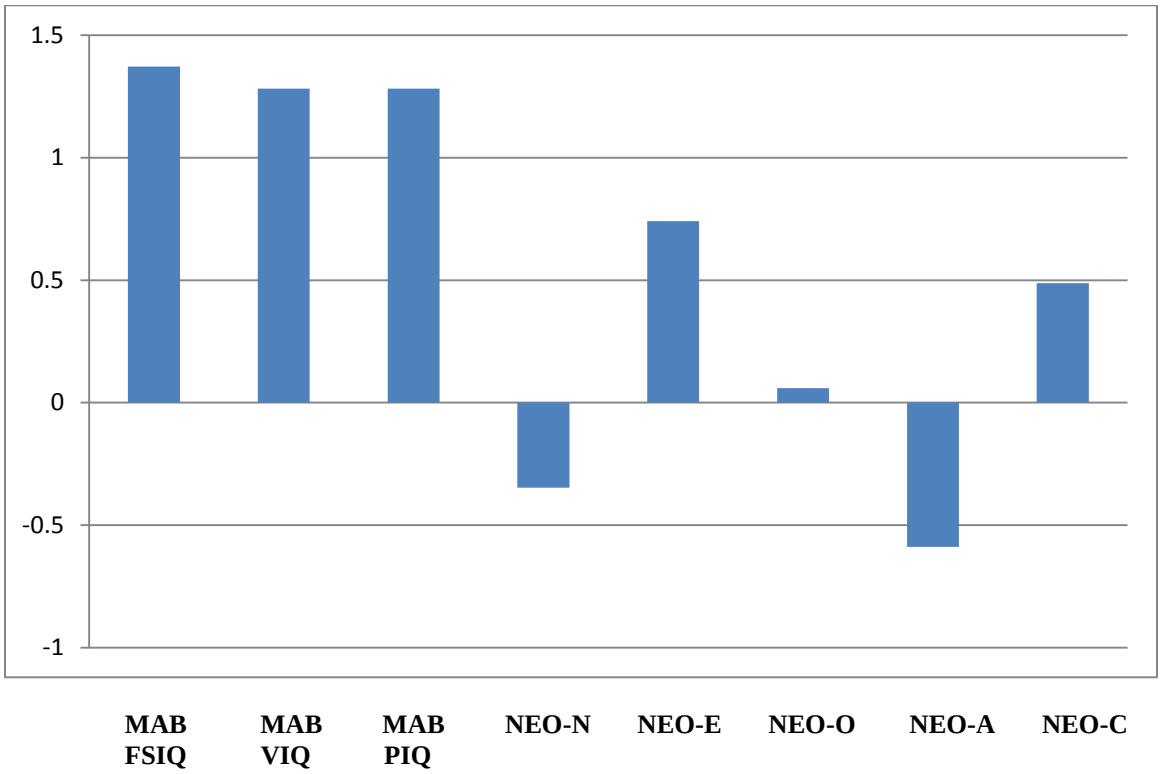


Figure 1. US Air Force pilot trainee norms versus the adult general population. The bars indicate the number of standard deviations the US Air Force pilot trainee means are above or below the adult population means. The scores are the Multidimensional Aptitude Battery (MAB) Full-Scale IQ (FSIQ), Verbal IQ (VIQ), and Performance IQ (PIQ) and the NEO-PI-R Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A), and Conscientiousness (C) scores.

To date, over 26,000 USAF student pilots have been administered some combination of these psychological tests. King, Barto, Ree, and Teachout (2011) presented a compendium of specialized USAF personality testing norms that can be used with military pilots and, cautiously, with applicants for civil airlines. This report includes profile sheets tailored specifically with these norms. A perusal of these norms demonstrates that USAF pilots differ from the general population on commercially published test norms. For example, this population has a mean *Agreeableness* T-score of 44.12 and a mean *Extraversion* T-score of 57.41, while the general population, by definition, has mean T-scores of 50 for both. This information is helpful when assessing individual pilots, as it places them in the proper context relative to their peers. The ALAPS may not be useful to those in the civilian sectors of aviation due to Federal law (the Americans with Disabilities Act) concerns, as it can be used to diagnose psychopathology in addition to measuring desirable personality traits. The problem would be administering it as part of a select-in procedure and violating Federal law by asking select-out type questions before extending a conditional employment offer. Further, it may be problematic as a selection tool due to the availability of the test manual (Retzlaff et al., 1997) in the open literature, encouraging coaching schemes, which could contribute to response inflation.

Potential Impact of New Technology on RPA Operator SAOC Requirements

RPA pilot SAOC requirements may be affected by mission objectives (e.g., manned-unmanned teaming, multi-RPA control), technology (e.g., automated take-off and landing, improved human-system interface design), and working conditions (e.g., work stressors such as shifts, number of hours, workload). It is likely that as technology advances, unmanned systems will become more autonomous, automated, and intelligent and more integrated with other manned and unmanned assets in a net-centric environment. Some tasks currently requiring manual control (take offs, landings, mission planning, sensor control) may be handled by automated systems, only requiring consent/approval by human operators. Decision aids (e.g., automatic target recognition, route planning, and timeline management) will enable the operator to assume more of a supervisory role in an integrated human-system team (van Breda, 2012). Technological developments may enable supervisory control of multiple RPAs or possibly swarms by a single operator. Under such conditions, mental and temporal workload will be high. SAOC requirements will focus on higher-order cognitive functioning. As aircraft autonomy increases, the need to manually control flight and psychomotor ability will decrease in importance. It is important that those responsible for human-system integration periodically examine the impact of changes in mission objectives and work environment and new technology on manpower, selection, and training requirements.

Discussion

Those responsible for human-system integration need to carefully consider the characteristics of human actors when developing or modifying systems. First, a job/task analysis must be done, including an analysis of cognitive, personality and other psychological characteristics needed for job success. Comparisons to the general population can be misleading. The use of specialized norms, when available and not prohibited by Section 106 of the Civil Rights Act of 1991, is highly recommended when assessing applicants as well as trained assets. Those responsible for human-system integration should also bear in mind the effects of changes in mission objectives and work environments and advances in technology on manpower, personnel, and training requirements. People, unlike machines, are prone to put their best foot forward (engage in response inflation) in an effort to influence an observer.

References

Carretta, T. R. (2005). *Development and validation of the Test of Basic Aviation Skills (TBAS),* AFRL-HE-WP-TR-2005-0172. Wright-Patterson AFB, OH, Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division.

Carretta, T. R. (2011). Pilot Candidate Selection Method: Still an effective predictor of US Air Force pilot training performance. *Aviation Psychology and Applied Human Factors, 1,* 3-8.

Carretta, T. R. (2013). Predictive validity of pilot selection instruments for remotely piloted aircraft training outcome. *Aviation, Space, and Environmental Medicine, 84,* 47-53.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Principles and practices of aviation psychology* (pp. 357-396). Mahwah, NJ: Erlbaum.

Carretta, T. R., Rose, M. R., & Bruskiewicz, K. T. (in press). Selection methods for remotely piloted aircraft systems operators. In N.J. Cooke, L. Rowe, & W. R. Bennett (Eds.). *Remotely piloted aircraft: A human systems integration perspective.* NY: Wiley.

Carretta, T. R., Teachout, M. S., Ree, M. J., Barto, E. L., King, R. E., & Michaels, C. F. (2014). Consistency of the relations of cognitive ability and personality traits to pilot training performance. *International Journal of Aviation Psychology, 24,* 247-264.

Chappelle, W., McDonald, K., Heaton, J. N., Thompson, W., & Haynes, J. (2012). Neuropsychological and personality attributes distinguishing high vs. low training performance of MQ-1B pilot trainees. *Aviation, Space, and Environmental Medicine, 83,* 261-262.

Costa, P. T., Jr., & McCrae, R. R. (1989). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). "'Normal' personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory": Reply. *Psychological Assessment, 4,* 20-22.

Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test form S: Analysis and comparison with previous forms. *Military Psychology, 22,* 68-85.

King, R. E., Barto, E., Ree, M. J., & Teachout, M. S. (2011). *Compilation of pilot personality norms* (Technical Report No. AFRL-SA-WP-TR-2011-0008). Wright-Patterson AFB, OH: U.S. Air Force School of Aerospace Medicine.

King, R. E., & Flynn, C.F. (1995). Defining and measuring the "Right Stuff": Neuropsychiatrically enhanced flight screening (N-EFS). *Aviation, Space, and Environmental Medicine, 66,* 951-956.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

Parasuraman, R., & Byrne, E. A. (2003). Automation and human performance in aviation. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Principles and practices of aviation psychology* (pp. 311-356). Mahwah, NJ: Erlbaum.

Patrick, J. (2003). Trends and contexts of pilot training. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Principles and practices of aviation psychology* (pp. 398-434). Mahwah, NJ: Erlbaum.

Powell, D., Kaplan, E., Whitla, D., Weintraub, S., Catlin, R., & Funkenstein, H. (2004). *MicroCog: Assessment of cognitive functioning, Windows ed.* San Antonio, TX: Pearson.

Retzlaff, P. D., Callister, J. D., & King, R. E. (1997). *The Armstrong Laboratory Aviation Personality Survey (ALAPS): Norming and cross-validation* (Technical Report No. AL/AO-TR-1997-0099). Brooks AFB, TX: Armstrong Laboratory.

Rose, M. R., Barron L, G., Carretta, T. R., Arnold, R. D., & Howse, W. R. (2014). Early identification of unmanned aircraft pilots using measures of personality and aptitude, *International Journal of Aviation Psychology, 24,* 35-52.

Smallwood, T., & Fraser, M. (1995). *The airline training pilot.* Brookfield, VT: Ashgate.

United States Air Force (2011). *Medical examinations and standards* (AFI-48-123). Washington, DC: Author.

United States Air Force (2014). *Air Force human system integration handbook.* Brooks City Base, TX: 711 HPW/WPO. Retrieved December 18, 2014 from http://www.wpafb.af.mil/shared/media/document/AFD-090121-054.pdf.

van Breda (Ed.) ((2012). *Supervisory control of multiple unmanned systems – methodologies and enabling human-robot interface technologies, TR*-HFM-170. North Atlantic Treaty Organization Research and Technology Organization.

Williams, H. P., Carretta, T. R., Kirkendall, C.D., Barron, L. G., Stewart, J. E, & Rose, M. R. (2014). Selection for UAS Personnel (SUPer) phase I report: Identification of critical skills, abilities, and other characteristics and recommendations for test battery development, NAMRU-D Technical Report 15-16. Wright-Patterson AFB, OH: Naval Aeromedical Research Unit – Dayton.

INVESTIGATING UAS OPERATOR CHARACTERISTICS INFLUENCING MISSION SUCCESS

Haydee M. Cuevas
Embry-Riddle Aeronautical University
Daytona Beach, FL

Kristina M. Kendrick
Embry-Riddle Aeronautical University
Daytona Beach, FL

Zane A. Zeigler
Embry-Riddle Aeronautical University
Daytona Beach, FL

David J. Hamilton
Embry-Riddle Aeronautical University
Daytona Beach, FL

The two objectives of this study were to 1) evaluate how specific operator characteristics (prior experience in manned and unmanned flight, teamwork, and gaming) influence mission success in unmanned aircraft systems (UAS) operations; and 2) evaluate the potential utility of a performance assessment tool. Mission success was assessed using a modified version of the Situation Awareness Linked Indicators Adapted to Novel Tasks (SALIANT) methodology. Eighteen participants completed a UAS scenario (port security) as part of 9 two-person crews (pilot and sensor operator). Results showed that the SALIANT measure was able to discriminate differences in performance among the UAS crews. Results also revealed significant correlations between the targeted operator characteristics and several of the SALIANT indicators. Findings from this study will be used to refine the SALIANT measure to support future research on how to optimize human performance in this domain.

The use of unmanned aircraft systems (UAS) is increasing at an unprecedented pace, with a broad range of applications including oil and gas exploration, agricultural management, wildfire mapping, weather monitoring, and emergency response (AUVSI, 2013). This trend has created significant human performance challenges such as how to: select and train UAS operators; design UAS control interfaces to minimize errors and avoid costly accidents; and safely integrate UAS into the National Airspace System (e.g., Dalamagkidis, Valavanis, & Piegl, 2008; Williams, 2006). The problems associated with these challenges are many, yet the solutions are presently few (Fern, Shively, Draper, Cooke, & Miller, 2011). Also, UAS crews differ from manned flight crews in crucial ways: crew and aircraft are not co-located; shift changeovers may occur during a mission; crew may be tasked to control multiple aircraft; control and feedback latency is common; lack standardized cockpit design and controls; lack standardized crew qualifications; and lack 'shared fate' with the aircraft (Tvaryanas, 2006). Accordingly, research is critically warranted to investigate these challenges.

Given the high consequence for errors and the high cost for attrition, the issue of UAS operator selection and training, in particular, has recently garnered considerable attention (e.g., Pavlas et al., 2009). To address this issue, this study investigated how specific operator characteristics (knowledge, skills, and abilities or KSAs) influence mission success in UAS operations. Greater experience in the targeted KSAs (prior experience in manned and unmanned flight, teamwork, and gaming) was hypothesized to be correlated with better performance during a simulated UAS scenario. Mission success was assessed using

a modified version of the Situation Awareness Linked Indicators Adapted to Novel Tasks (SALIANT) methodology, developed by Muniz, Stout, Bowers, and Salas (1998). SALIANT provides a theoretically-based assessment of the observed behaviors that are indicative of the team process behaviors that support team situation awareness (e.g., how information exchange is used as an input for building team member situation awareness; Milham, Barnett, & Oser, 2000). Thus, another important objective of this study was to evaluate the potential utility of the modified SALIANT as a performance assessment tool.

Method

Participants

Altogether, 18 participants (all males; average age = 25.29 years) participated in this study as part of two-person crews (pilot, sensor operator). Participants were recruited from the Unmanned Aircraft Systems Science (UASS) undergraduate program at a private aeronautical university in the southeastern United States. The UASS degree provides the necessary expertise for graduates to seek employment as pilots/operators, observers, sensor operators, and operations administrators of UAS. Thus, recruiting participants from this subject pool helps to increase the generalizability of the study's findings to real world UAS operations. Participants were either currently enrolled or had recently completed the *UAS Flight Simulation* course, the final capstone course in the UASS program. One crew was dropped from the analysis due to missing data, leaving a total of eight two-person crews. All participants in the study were treated in accordance with the ethical standards of the American Psychological Association.

Materials and Apparatus

Prior to participation in the study, participants were asked to review and complete an informed consent form and a biographical data form that solicited information on the targeted KSAs. Table 1 lists the items surveyed the biographical data form.

Table 1.
Biographical Data Form Items for Targeted KSAs.

KSA	Item
Manned Flight Experience	• Do you have any manned aircraft piloting experience? __ Yes __ No If yes, approximately how many hours? ___Hours • Do you have any pilot ratings or certifications? If yes, please list in the space below.
Unmanned Flight Experience	• Do you have any prior experience in operating unmanned systems? __ Yes __ No If yes, which classes have you previously taken? (Check all that apply): AS 220; AS 235; AS 403; AS 473 • How many hours have you spent in open simulation lab? (Not including class time) ___ Hours • Do you have any prior military experience operating unmanned systems? _ Yes _ No If yes, approximately how many hours? ____Hours
Teamwork Experience	• How much team experience did you have before taking part in this study? *None* (0 teams); *Very Little* (1 - 2 teams); *Some* (3 - 4 teams); *Fair* (5 - 6 teams); *Extensive* (> 6 teams) • Give an estimate of the percentage of time spent on teamwork activities as opposed to individual activities in the last week. Include both in-class and outside class activities: 0%; 0% to 20%; 20% to 40%; 40% to 60%; 60% to 80%; > 80%

Gaming Experience	• Give an estimate of the time spent (in hours) typically playing any type of video or computer game per week. If none, simply write "0" next to that game. *First-Person Shooter* (Halo, COD, Battlefield, etc.); *Racing* (Forza, Need for Speed, etc.); *Role-Playing Games* (Skyrim, Fallout, World of Warcraft, etc.); *Strategy/Puzzle* (Candy Crush, Solitaire, etc.); *Multiplayer/Online Gaming*; *Other* (please specify)

To assess the influence of these KSAs on team performance, the project team leveraged an existing UAS scenario (port security) developed for the *UAS Flight Simulation* course. In the port security scenario, the UAS crew (pilot and sensor operator) must navigate the UAS to a designated location in the harbor, conduct surveillance in the area to detect and identify the targeted vessel, gather information on the vessel, and then return the UAS to base. During each scenario, crews are presented with an emergency (e.g., oil leak, engine failure) requiring dynamic replanning and teamwork to resolve the situation.

In consultation with subject matter experts and the course instructor, the project team created a modified version of the Situation Awareness Linked Indicators Adapted to Novel Tasks (SALIANT) methodology, developed by Muniz et al. (1998) and adapted by Fiore, Fowlkes, Martin-Milham, and Oser (2000). The modified SALIANT included three new categories: Task / Equipment Knowledge, Crew Resource Management, and Mission Monitoring (see Table 2).

Table 2.
Modified SALIANT Indicators (adapted from Fiore et al., 2000).

Category	SALIANT Indicator
1. Spatial Orientation	1.1 Demonstrates awareness of location in space
	1.2 Uses available information sources
	1.3 Cross checks information
	1.4 Scans internal and external environment for abnormal conditions, changes, landmarks
2. Cue Sharing	2.1 Provides and requests backup
	2.2 Reports problems
	2.3 Informs others of actions taken
3. Problem Solving	3.1 Locates potential source of problem
	3.2 Resolves discrepancies
	3.3 Anticipates consequences of actions, decisions, and potential problem situations
4. Information Management	4.1 Provides information in advance
	4.2 Adheres to standard communication format
	4.3 Briefs status
5. Task Management	5.1 Takes action at the appropriate time
	5.2 Exhibits skilled time sharing among tasks
6. Task / Equipment Knowledge	6.1 Demonstrates knowledge of tasks
	6.2 Demonstrates knowledge of equipment/systems
	6.3 Commits minimal operational errors and mistakes
7. Crew Resource Management	7.1 Resolves conflicts with teammates
	7.2 Delegates tasks with appropriate feedback
	7.3 Asks clarification questions as necessary
	7.4 Effectively use available resources
8. Mission Monitoring	8.1 Engages in mission planning and dynamic re-planning
	8.2 Recognizes and responds to messages sent to crew

Subject matter experts carefully reviewed the UAS scenario and then mapped the naturally occurring team behaviors associated with the SALIANT indicators onto a chronological checklist based on expectations of how these behaviors would unfold during the course of the scenario. Examples of SALIANT checklist items are shown in Table 3. During performance of the UAS scenario, four subject matter experts completed the SALIANT checklist, with two trained observers per crew.

Table3.
Example SALIANT Checklist Items for Port Security UAS Scenario.

Category	SALIANT Indicator	Checklist Item
Spatial Orientation	Demonstrates awareness of location in space	Pilot raises landing gear at appropriate altitude
Crew Resource Management	Delegates tasks with appropriate feedback	Crew works together to identify emergency
Mission Monitoring	Engages in mission planning and dynamic replanning	Pilot continually updates the emergency mission entry waypoint

Results and Discussion

Given the small sample size and directional hypothesis for this initial study, alpha was set at $p < .05$, one-tailed. As illustrated in Table 4, the SALIANT indicators were able to discriminate differences in performance among the eight crews. Performance across the SALIANT categories ranged from a minimum of 0% to a maximum of 100%. Average scores ranged from 28% to 58%.

Table 4.
Descriptive Statistics for SALIANT Categories.

SALIANT Category	Minimum	Maximum	Mean	Std. Deviation
Spatial Orientation	.4188	.8182	.5490	.1351
Cue Sharing	.3281	.8438	.5800	.1607
Problem Solving	.0000	.7500	.2813	.2720
Information Management	.0833	.7167	.3177	.2229
Task Management	.0000	1.0000	.5158	.2615
Crew Resource Management	.2500	1.0000	.5313	.3010

Note. N = 16 for each category.

Bivariate correlation analysis was conducted between each of the targeted KSAs (flight experience, teamwork experience, and gaming experience) and team performance as assessed by the SALIANT. Significant correlations are reported in Table 5.

Table 5.
Significant Correlations between KSAs and SALIANT Categories.

KSA	SALIANT Category	Correlation
Manned Flight Experience		
Manned Aircraft Piloting Experience	Crew Resource Management	r (16) = .557, p = .0125
Manned Flying Hours	Crew Resource Management	r (15) = .542, p = .0185
Pilot Ratings / Certifications	Crew Resource Management	r (16) = .473, p = .032
Unmanned Flight Experience		
UAS Open-Simulation Hours	Task Management	r (16) = - .509, p = .022
Teamwork Experience		
Team Experience	Task Management	r (16) = .471, p = .0325
Team Experience	Problem Solving	r (16) = .471, p = .033

Team vs. Individual %	Problem Solving	r (16) = .465, p = .035
Gaming Experience		
First-Person Shooter	Spatial Orientation	r (16) = .503, p = .0235

Results showed a significant positive correlation between Manned Flight Experience and SALIANT indicators for Crew Resource Management (CRM). Participants with greater Manned Flight Experience performed better on the SALIANT CRM items. This result is to be expected since pilots receive CRM training during the course of their flight instruction.

Unexpectedly, results showed a significant negative correlation between Unmanned Flight Experience and SALIANT indicators for Task Management. Participants with greater Unmanned Flight Experience performed worse on the SALIANT Task Management items. It is possible that, without instructor feedback to calibrate their performance, the additional time spent practicing in the simulation during open-simulation training hours was not beneficial for enhancing their skill acquisition.

Results also showed a significant positive correlation between Teamwork Experience and SALIANT indicators for Task Management and Problem Solving. Participants with greater Teamwork Experience performed better on the SALIANT Task Management and Problem Solving items. This finding suggests that crews were able to transfer domain-general team KSAs to coordinate their activities, which, in turn, may facilitate successful task completion.

Finally, results showed a significant positive correlation between Gaming Experience with First-Person Shooter games and SALIANT indicators for Spatial Orientation. Participants with greater experience with these types of games performed better on the SALIANT Spatial Orientation items. This result likely may be due to the requirement for spatial awareness in these types of games where the player is an avatar in a virtual world. In order to succeed, the player must take in all available information to assess their situation correctly.

Conclusion

Results from this study offer initial support for the potential utility of the SALIANT methodology as a performance assessment tool. However, while promising, conclusions drawn from these results are tentative due to the study's small sample size. Thus, future research is warranted to further validate the SALIANT methodology with a larger sample size as well as with an increased number of items for the SALIANT indicators. In addition, although results revealed significant correlations between the targeted KSAs and UAS crew performance, further research in necessary to empirically evaluate the causal nature of this relationship.

In sum, the long-term goal of this research program is to promote successful UAS operations, in both the private and public sector, by optimizing human performance and minimizing human errors. Findings from this line of research may offer insights into the development of personnel selection tools and UAS operator training programs to achieve this goal.

Acknowledgements

This research was partially supported by funding from FY14 ERAU Faculty Internal Research Grant #13208 to Haydee M. Cuevas, College of Aviation Department of Doctoral Studies. The views herein are those of the authors and do not necessarily reflect those of the organizations with which the authors are affiliated. Special thanks to Alex Mirot, Dat Nghiem, and Shane Thompson for their valuable contributions to this project. Address correspondence to Haydee M. Cuevas at cuevash1@erau.edu.

References

Association for Unmanned Vehicle Systems International (AUVSI) (2013, March). The economic impact of unmanned aircraft systems integration in the United States (Economic Report). Retrieved from: http://www.auvsi.org/econreport

Dalamagkidis, K., Valavanis, K. P., & Piegl, L. A. (2008). On unmanned aircraft systems issues, challenges and operational restrictions preventing integration into the National Airspace System. *Progress in Aerospace Sciences, 44 (7-8),* 503-519.

Fern, L., Shively, R. J., Draper, M, H., Cooke, N. J., & Miller, C. A. (2011). Human-automation challenges for the control of unmanned aerial systems. *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting* (pp. 424-428). Santa Monica, CA: Human Factors and Ergonomics Society.

Fiore, S. M., Fowlkes, J., Martin-Milham, L., & Oser, R. L. (2000). Convergence or divergence of expert models: On the utility of knowledge structure assessment in training research. *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomic Society, 2,* 427-430. Santa Monica, CA: Human Factors and Ergonomics Society.

Milham, L. M., Barnett, J. S., & Oser, R. L. (2000). Application of an event-based situation awareness methodology: Measuring situation awareness in an operational context. *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society,* 2, 423-426. Santa Monica, CA: Human Factors and Ergonomics Society.

Muniz, E., Stout, R., Bowers, C., & Salas, E. (1998). A methodology for measuring team situational awareness: Situated Linked Indicators Adapted to Novel Tasks (SALIANT). The First Annual Symposium/Business Meeting of the Human Factors & Medicine Panel on Collaborative Crew Performance in Complex Systems, Edinburg, United Kingdom.

Pavlas, D., Burke, S., Fiore, S. M., Salas, E., Jensen, R., & Fu, D. (2009). Enhancing unmanned aerial system training: A taxonomy of knowledge, skills, attitudes, and methods. *Proceedings of 53rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1903-1907). Santa Monica, CA: Human Factors and Ergonomics Society.

Tvaryanas, A. P. (2006). Human systems integration in remotely piloted aircraft operations. *Aviation Space and Environmental Medicine, 77* (12), 1278-1282.

Williams, K. W. (2006). *Human factors implications of unmanned aircraft accidents: Flight-control problems*. Report Number DOT/FAA/AM-06/8. Washington, DC: Office of Aerospace Medicine. http://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/2000s/media/200608.pdf

IMPACT OF WEATHER INFORMATION LATENCY ON GENERAL AVIATION PILOT SITUATION AWARENESS

Barrett S. Caldwell
Purdue University
West Lafayette, IN

Nsikak Udo-Imeh
Purdue University
West Lafayette, IN

Mary E. Johnson
Purdue University
West Lafayette, IN

Lucero Duran
Purdue University
West Lafayette, IN

Geoffrey Whitehurst
Western Michigan University
Kalamazoo, MI

Megan M. Nyre
Purdue University
West Lafayette, IN

Vladimir Rishukin
Western Michigan University
Kalamazoo, MI

Lauren Sperlak
Purdue University
West Lafayette, IN

A critical element of situation awareness and sensemaking support for humans in complex environments is the ability to access, detect, and integrate environmental elements to recognize and project the state of the world. Some past research has suggested that new weather technology capabilities in general aviation (GA) flight settings could help improve pilot decision making and reduce accidents such as unintentional transitions from visual flight rules (VFR) to marginal VFR or even instrument meteorological conditions (IMC). This paper addresses an ongoing Federal Aviation Administration (FAA) funded research project investigating the effect of transmission delays and update latencies in presentations of weather information to pilots in the GA environment. Across a range of fixed-install, portable, and handheld (i.e. tablet, smartphone) weather information technologies, latencies of up to 15-20 minutes can be identified. These latencies may affect the use of information regarding dangerous weather conditions and timelines of pilot planning activities during VFR-to-IMC transitions.

Introduction

In General Aviation (GA) flight, pilots obtain a weather briefing before flights in what is described by the Federal Aviation Regulations (Title 14 Code of Federal Regulations) as 'preflight action' in the section Subpart B - Flight Rules. Specifically, the language in Part 91.103 includes a requirement that each pilot in command "become familiar with all available information…" including "weather reports and forecasts, fuel requirements, alternatives available if the planned flight cannot be completed" for flights "not in the vicinity of an airport" (FAA, 2014).

Traditionally, this requirement was met by the pilot in command telephoning 1-800-WXBrief and asking for a standard briefing from a Flight Service Station (FSS) weather briefer. The briefer provides in-depth weather briefing information to the pilot and records the pilot's name, aircraft N number and other pertinent information so that the specific standard briefing is retained for a period of X days. The pilot listens to the briefer, asks and answers questions, and writes down on paper the information transmitted. Recent research indicates an increase in the direct use of web-based weather products for flight preparation by GA pilots (Casner, et al., 2012; Knecht, 2011). Anecdotal evidence also indicates the proliferation of mobile aviation weather information products and tools with access to the internet while airborne has led to a reliance of web-based products and tools for enroute weather updates.

Advances in technology allow easy access to weather information elements (such as METAR, TAF, AIRMET/SIGMET, or FA), provided not only by the Federal Aviation Administration and the National Oceanic and Atmospheric Administration, but also by a number of commercial organizations, in a variety of web-based mobile devices. It is not surprising that "Pilots seem to be transitioning from a traditional means of assisted weather briefing to self-briefing" (Casner et al., 2012). This raises the question - to what extent do GA pilots actually *make use of and effectively use* the weather services that are available for them?

Results of a study of weather-related GA occurrences (Batt, 2005) identified 280 incidents out of 491 occurrences (57%). Pilots made VFR into IMC decisions while other pilots avoided bad weather only in 151 cases (30.8%), and made precautionary landings even more rarely - in 60 cases (12.2%). These results confirmed the idea that decisions made by pilots play a leading role in weather-related incident or accident outcomes.

Pilots' decisions can be affected by the timeliness of weather information presented by the technology (Bailey, 2007). For example, NEXRAD radar data and images, which can represent weather information in graphical form that could decrease workload on a pilot, cannot keep up with rapidly changing weather due to limits in data aggregation and dissemination. Pilots using software apps that display NEXRAD images can receive outdated information that decrease the accuracy or validity of pilot decision making in degrading or rapidly changing weather conditions (Bustamante, et al., 2007). These factors, combined with sometimes rapidly changing weather conditions and widely varying pilot experience, represent major concerns to the potential safety of the GA flying community.

Weather Technology in the Cockpit (WTIC)

As part of its efforts to address the future of air traffic in the United States, the Federal Aviation Administration (FAA) has requested research in the area of GA pilot decision making and behavior, including how pilot decisions are affected by new weather information tools available to and used by the pilot. This paper addresses ongoing work being conducted by researchers in the FAA Center of Excellence for General Aviation Research, known as the Partnership to Enhance General Aviation Safety, Accessibility and Sustainability (PEGASAS). The research presented in this paper is conducted as part of the PEGASAS Project 4 in Support of the Weather Technology in the Cockpit Program, or "PEGASAS WTIC". Of the four teams involved in this effort, the authors of this paper (representing PEGASAS Team 4D) were tasked to focus on technology integration factors that affect how and when weather information is presented to the GA pilot.

Original Research Questions and Gaps

Discussions during the initial stages of collaboration between PEGASAS researchers and FAA WTIC Program leadership emphasized the range of FAA-approved weather information technology systems suitable for installation and use in GA aircraft. However, there is a growing use among GA pilots of mobile devices and software applications for accessing weather information products. The initial organization of research tasks for PEGASAS Team 4D was based on the following "primary GAP":

GAP 0: There is a limited understanding of how FAA-authorized weather information sources, as presented / displayed in the range of available tools (including mobile devices and software applications), influences pilot interaction with and use of weather information in degrading weather conditions.

While conducting PEGASAS Project 4 efforts throughout 2014, it has became apparent that structuring research efforts and discussions around GAPs, as opposed to research deliverables, can substantially improve the value of the research findings for use by the FAA and the GA community. Summaries of Team 4D progress during 2014 continue to describe effort in terms of tasks and deliverables presented in this document. However, the description of GAPs provides better integration of key findings and research priorities for integration across PEGASAS activity.

Weather Information Tools and Systems Studied

The general aim of weather technology tools and products in the cockpit is to enable pilots to obtain an updated weather briefing and current conditions for a selected flight plan. The research tasks and deliverables addressed by Team 4D are intended to examine a range of available hardware device and software applications (including the proliferation of mobile devices), beyond the expected uses of certified devices only referencing authorized MET information products. Using a set of popular and varied available systems, Team 4D tasks examined pilot activity to obtain current weather information (including an updated weather briefing) for a selected flight plan, and pertinent issues that arise from efforts to obtain the updated brief.

Cataloging Weather Information Systems

The tools currently used to provide weather information to GA pilots were classified as either hardware or software, based on the following criteria. Hardware tools were taken to include handheld or dashboard-mounted devices and installed devices designed specifically for providing weather (and/or navigation) information to pilots. Software tools represent applications available on general purpose devices (such as tablets or smartphones),

including those that make use of an associated hardware project (e.g., Stratus ADS-B receiver). Team 4D evaluated nine hardware tools and 55 software tools (25 Android and 30 IOS) were identified and inventoried for consideration. Weather briefing and information update capabilities were benchmarked against the FAA 1-800-WXBRIEF Flight Service Station (FSS) service, including pilots calling FSS via radio while in-flight. It is important to note that Team 4D also considered technologically possible uses, even if they are not recommended or even subject to degraded performance (such as use of cellular network service signals above 5,000' altitude). A total of eight systems have been selected and included for additional analysis, including the web-based www.1800wxbrief.com software tool.

Goal-Directed Task Analysis and Tool Comparison

Cognitive Work Analysis and Goal-Directed Task Analysis (GDTA) tools have been previously identified by FAA as relevant HF tools for aviation task evaluations (FAA, n.d.). Consistent with the scope of this project, the primary GDTA pilot activity is to obtain an updated weather briefing for a selected flight plan. Team 4D findings were subject to how different use patterns among the range of mobile devices and software applications can affect pilot planning tasks and use of appropriate weather information sources. Differences in weather information available to pilots during pre-flight and in-flight conditions, combined with the variety of mobile device and software applications in common use, highlights additional GAPs identified by Team 4D:

GAP 1: The effectiveness of available mobile device and software application tools is affected in unknown ways due to feature availability and use of weather information sources based on device / application and relevant phase of flight.

GAP 2: Information presentation and interface design in some mobile devices and software applications may limit or prevent pilot planning activity in potentially degrading ways during adverse or degrading weather conditions.

Issues from Task Analysis Regarding Weather Information

The use of software applications and mobile devices in the GA cockpit is subject to a number of human factors and ergonomics (HFE) considerations that are more formally addressed in fixed-installation multi-function displays. However, detailed HFE of glare, vibration, or other issues was outside of the scope of Team 4D analysis. More critical elements of study included task analysis steps required to obtain information, as well as the demand on working memory or situation awareness (SA) associated with obtaining required information to support pilot decision making.

Effective weather information systems should provide information that depicts, in unambiguous ways, the important features relevant to the pilot's ability to select a proper course of action (Shattuck and Miller, 2006). Past work has clearly identified that the "picture" of the weather that the pilot develops from accessing and assessing weather information that is presented will affect pilot decision making (McAdaragh, 2002). Understanding developed about the weather situation during any phase of flight is limited by the amount of **uncertainty** associated with the weather information, the **reliability/validity** of the weather information, and the **time stress** and **task load** (Latorella et al., 2002) during the phase of flight:

- Uncertainty associated with the weather information presented to the pilot may be the level of *spatial* or *temporal* uncertainty contained in the information, which require additional mental workload to interpret.
- The reliability/validity of the weather information presented to the pilot concerns its source, which must deliver *accurate* and *complete* information as well as the *availability* of the information, which may be impacted by the type of data link used.
- Time stress and task load are interconnected and impact pilot cognitive workload and the time available to process information, make a decision and take action.

Weather Event Triggers, Latency, and SA

The interplay between information presentation and dynamic weather factors is critical during flights in or around adverse weather, where action must be taken to avoid the potential for transitioning from VMC to IMC, or dangerous exposure to severe weather events. Naturalistic decision-making dominates under situations with uncertainty and high time stress (Wiggins and O'Hare, 1995, Elgin and Thomas, 2004). In these conditions, decision-making tasks and resulting actions are more likely to be automatic, executed intuitively rather than analytically (Caldwell, 2008). The general model of SA describes the pilot's awareness being comprised of *perceiving* the components of the weather-relevant enviroment (Level 1 SA), *integrating / comprehending* those

components (Level 2 SA), and *projecting* those components into the future (Level 3 SA) (Endsley, 2000). Specifically, relevant weather event components were identified in Team 4D based on interviews with seven experienced aviation weather and flight instructors. Results of the interviews highlighted the following weather transitions, which are shown in Table 1. For the purposes of the PEGASAS WTIC research, these components influencing pilot SA can be described as "weather event triggers". Note that these findings replicate a number of past research studies regarding critical weather event triggers that should (but do not always) cause pilots to consider alternate flight path / diversion / return activities (Johnson and Wiegmann, 2011).

Table 1.
Results of seven interviews regarding possible Weather Event Triggers (# reporting trigger)

On the Ground (# reporting)	In the Air (# reporting)
Thunderstorms (3)	Clouds below form a CIG (1)
MVFR or close (2)	Descending clouds (3)
Clouds (night) (1)	Thunderstorms (3)
High winds (2)	Lowering visibility (2)
Snow (1)	Shapes of clouds (1)
General bad, don't go (1)	Precipitation (2)
Convective outlook (1)	Tall buildups in clouds (2)
Visbility dropping (1)	Convection (1)
Advisories from FSS (1)	Wind shear (2)
	Moderate / greater turbulence (2)
	Icing (1)
	High winds (1)
	Advisories from FSS (1)
	ATIS / ASOS reports (1)

Planning, near-term, and immediate decision-making activities are influenced by the level of time stress of the situation the pilot faces (Elgin and Thomas, 2004); guidelines indicate separations into planning (> 20 minutes), near-term (3-20 minutes) and immediate (< 3 minutes) decision-making regimes (FAA, 2014; RTCA SC-206, 2014). Pilot SA is maintained by two forms of environmental data: "out-the-window" (OTW) input and the instrument input. OTW input provides a clear and contemporary, yet limited, view of what the conditions are directly outside the aircraft. Instrument input provides other types of information in the general vicinity, certainly beyond the immediately visible range. However, due to the technology capabilities of collecting, integrating, and broadcasting NEXRAD data, "real time" NEXRAD information regarding potentially dangerous and fast moving event triggers (e.g., convection, thunderstorm fronts) are subject to considerable delays. Members of Team 4D collected NEXRAD latency data using a fixed-installation Garmin 1000 hardware system in an actual SR-20 GA aircraft on the flight line at the Purdue University Airport. After initial loading, "time now" and the screen "time stamp" were recorded every minute using the aircraft clock to indicate "time now" and G1000 XM weather screen time as "time stamp". The "time stamp" doesn't include the time it took for the image to be generated and sent to Sirius XM, but is an estimate of the latency from satellite upload to image appearance on the G1000. The latencies presented in Table 2 below are confirmed by similar intervals reported in AC 00-63A (FAA, 2014). Additional investigation suggests that NEXRAD latency may actually be an additional 2 or 3 minutes from when the image was taken to satellite upload.

These findings, plus additional Team 4D investigations of flight simulator display capabilities, identified two additional GAPs focused on weather information presentation latencies.

*GAP 3: **Reported** aviation weather update capabilities and use of FAA-approved weather information sources differ in latency or availability from **achieved** updates presented to the GA pilot during actual flight.*

GAP 4: Identification of adverse weather event triggers (and impact on pilot planning efforts) differs between out the window and mobile device / software application presentations of weather conditions; differences in awareness of trigger severity and potential impact affects pilot planning task and time sequences.

Table 2.

Tools by Latencies and Refresh Rates

Tool	Latencies (min)	Refresh Rates (min)
NEXRAD	2 – 7 minutes	4 – 5 minutes
Echo Top	1 – 16 minutes	7 – 9 minutes
Cloud Top	4 – 21 minutes	11 – 13 minutes
XM Lightning	0 – 7 minutes	3 – 7 minutes

Note: Cell Movement presentation experienced no additional latencies beyond 2 minutes

Popular commercial aviation training device simulators (ATDs) available by December 2014 are not capable of presenting weather information to weather displays that are not synchronized to the computer-generated "out the window" (OTW) display. Team 4D team efforts during Phase I determined that modifications of existing ATDs are infeasible to demonstrate NEXRAD latencies of 10-20 minutes in a General Aviation (GA) operational scenario. Previous research, as well as Team 4D empirical data, has demonstrated that such latencies are characteristic of actual GA operations (NTSB, 2012). Discrepancies in OTW and radar-based presentations of environmental components represent major challenges to developing and maintaining SA, in either individual or team-based performance scenarios; even minor shifts in information presentation modality could severly degrade task coordination performance (Caldwell and Everhart, 1998).

Ongoing Research Questions and Gap Resolutions

Based on the GAPs identified, and tasks completed, by the PEGASAS WTIC Team 4D in 2014, a number of additional research and technology development activities are planned for Phase II work in 2015. These activities include:

Select and use of existing PEGASAS Phase I weather event scenarios, and generate additional scenarios, to examine influence of information latencies on planning capabilities and diversion activities

The outcomes of this activity will be used to empirically test available GA pilot planning capabilities in low- and high-fidelity aviation simulation environments, as affected by weather information presentation latencies.

Determine feasibility of PC-based GA aviation training simulator prototypes integrating realistic (up to 20 minute) weather information presentation latencies

If a potential PC-based aviation training device (PC-ATD) is feasible and can be developed with realistic presentations of weather information presentation latencies, such a device could provide substantial education and training benefits to the GA community.

Comparative testing of weather information latency effects on pilot planning capabilities and tasks in both low-fidelity (PC prototype) and high-fidelity aviation simulator environments

The use of experimenter-controlled weather information presentation latencies can help identify and quantify the effects of those latencies on pilot immediate (0-3 minute) and near-term (3-20 minute) planning capabilities. If similar experimental conditions can be run in both the low-fidelity and high fidelity contexts, and directly compared, additional benefits can be obtained.

As part of the PEGASAS Project 4 in support of the WTIC Program, we believe that the creation of suitable PC-ATD capabilities (with suitable latencies) and sentinel weather event scenarios can be an important contribution to pilot education and knowledge development to increase awareness and reduce risks of dangerous decision making currently affecting the GA pilot community.

Acknowledgements

Funding for this project was provided to the PEGASAS Team 4D (Barrett Caldwell, PI) as part of FAA Air Transportation Center of Excellence for General Aviation Research, Cooperative Agreement 12-C-GA-PU, Amendments 007 and 017. The findings and perspectives presented here are those of the authors, and do not represent official positions of the FAA or any other agency.

References

Bailey III, W.R., Bustamante, E.A., Bliss, J.P., Newlin, E.T. (2007). Analysis of Aircrews' Weather Decision Confidence as a Function of Distance, Display Agreement, Communication, Leadership, and Experience. The International Journal of Applied Aviation Studies, 7(2), 272-294.

Bustamante, E.R., Fallon, C.K., Bliss, J.P., Bailey III, W.R., Anderson, B.L. (2007). Pilots' Workload, Situation Awareness, and Trust During Weather Events as a Function of Time Pressure, Role Assignment, Pilots' Rank, Weather Display, and Weather System. *The International Journal of Applied Aviation Studies*, 5(2), 347-367.

Caldwell, B. S. (2008). Knowledge sharing and expertise coordination of event response in organizations. *Applied Ergonomics*, **39**, 427-438.

Caldwell, B. S., & Everhart, N. C. (1998). Information Flow and Development of Coordination in Distributed Supervisory Control Teams. *International Journal of Human-Computer Interaction, 10*(1), 51-70.

Casner, S.M., Murphy, M.,P., Neville, E.C., Neville, M.R. (2012). Pilots as weather briefers: The direct use of aviation weather products by general aviation pilots. *The International Journal of Aviation Psychology, 22(*4), 367-381. doi: 10.1080/10508414.2012.718241

Elgin, P. D., & Thomas, R. P. (2004). An integrated decision-making model for categorizing weather products and decision aids. Technical Report NASA/TM-2004-212990, National Aeronautics and Space Administration, Langley Research Center, Hampton, Virginia.

Endsley, M. R. (2000). Theoretical Underpinnings of Situation Awareness: A Critical Review. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

FAA (2014). Use of Cockpit Displays of Digital Weather and Aeronautical Information. Advisory Circular 00-63A. Published April 7, 2014. Washington, DC: Federal Aviation Administration.

FAA (n.d.). "FAA Human Factors HF Tools." Electronic site, retrieved Feb 20, 2015. Available at http://www.hf.faa.gov/workbenchtools/default.aspx?rPage=ToolList&subCatID=28

Johnson, C. M., & Wiegmann, D. A. (2011, September). Pilot Error During Visual Flight Into Instrument Weather An Experiment Using Advanced Simulation and Analysis Methods. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, No. 1, pp. 138-142). Sage Publications.

Knecht, W. R. (2011). Testing Web-Based Preflight Weather Self-Briefing for General Aviation Pilots (No. DOT-FAA-AM-11-05). Oklahoma City, OK: FAA Civil Aerospace Medical Institute.

McAdaragh, R. M. (2002). Toward a Concept of Operations for Aviation Weather Information Implementation in the Evolving National Airspace System. (NASA/TM-2002-212141.) NASA Langley Research Center.

NTSB (2012). "In-Cockpit NEXRAD Mosaic Imagery: Actual Age of NEXRAD Data Can Differ Significantly from Age Indicated on Display." NTSB Safety Alert, available at http://www.ntsb.gov/safety/safety-alerts/Documents/SA_017.pdf. Retrieved February 20, 2015.

Shattuck, L. G., and Miller, N. L. (2006). Extending Naturalistic Decision-making to Complex Organizations: A Dynamic Model of Situated Cognition. *Organization Studies* **27** (2): DOI: 10.1177/0170840606065706. Retrieved December 29, 2014 from http://www.dtic.mil/dtic/tr/fulltext/u2/a487772.pdf.

Wiggins, M. W., & O'Hare, D. (1995). Expertise in aeronautical weather-related decision-making: A cross-sectional analysis of general aviation pilots. *Journal of Experimental Psychology: Applied, 1*, 305–320.

AVIONICS TOUCH SCREEN IN TURBULENCE: SIMULATION FOR DESIGN

Sylvain Hourlier & Xavier Servantie
Thales Avionics
Bordeaux, France

As touch screens are everywhere in the consumer market Thales has launched in depth evaluations on their introduction in the cockpit. One of the challenges is to verify its compatibility with in flight use under turbulence conditions, including light, moderate and severe. In flight accelerometer collections were performed to provide us with a baseline for choosing between possible simulation solutions. Thales recognized early on the need for such a tool as it would enable us to define recommendations for our HMI designs. The objectives were first to validate specific complex touch/gestures using all the potential of touch interactions for novel cockpit Human Machine Interfaces and second to look into the various physical anchoring solutions capable of facilitating touch screens interactions in aeronautical turbulent environments. Given the 6 axis accelerometer profiles that were collected, a number of potential candidate simulation platforms were selected. They were reviewed in terms of performance and cost. Our final candidate is an Hexapod structure capable of reproducing those profiles with acceptable validity. This paper presents the works that enabled us to validate such an hexapod as a viable simulator for our tests and the development of an avionics platform for touch interactions under light to severe turbulences. Pilots were asked to evaluate 6 simulated profiles designed to mimic the “inflight” references. Tests were performed to validate the best profiles for each level of turbulence. The selected profiles were then used to evaluate our touch screen propositions in light, moderate and severe turbulent conditions. Preliminary results are presented.

The ubiquity of touch technology and its upcoming in cockpits

The trend of touch technology for interaction is undisputed. DisplaySearch, a market analysis firm, forecasts it to grow to over $16 billion by 2016 and $31.9 billion by 2018. The market growth is being driven by increased demand from applications such as iPads and other tablet PCs, smart phones, and emerging notebook PC designs. (Sieh, 2010). More recently another analyst confirms the trend and the touch screen market grew from $1.5 billion in annual revenues in 2008 to over $6 billion in 2011 (Blanco, 2012). Since the uprising of the inevitable Iphone, touch interactions overtook the cellphone industry. Nowadays, kids try to interact spontaneously on any screen they come by as if it “obviously” had to be a touch screen.

Facing such an inevitable trend, the AV2020 full touch screen cockpit concept has been developed (2020). It comprises multiple seamless touch screens in an integrated approach to pilots’ HMI demands. Yet implementing touch technology in a liner cockpit means complying to part 25 aircraft certification. The process is thorough and specifies that the design of systems should take into account aeronautical effects (such as turbulence) and the way they affect the efficiency of pilots’ interactions. Hence, an human factors evaluation was decided to alleviate the

risk on usability of touch displays in turbulence, refine design recommendations for interactions with touch technology (HMI design and physical installation) and prepare certification.

Characterizing aeronautical turbulence

Origin of turbulence

Even with limited flight experience one can relate to the term “turbulence” in flight. Usually the captain orders passengers to their seat with their seat belt tightened due to upcoming turbulence. Atmospheric turbulence is defined as “small-scale, irregular air motions characterized by winds that vary in speed and direction” (Encyclopedia Britannica, 2013). One must note that turbulence does not compare to a vibration, as it is chaotic by nature and not cyclic.

Intensity of turbulence

Turbulence is separated into four levels of intensity. Each different level of intensity can be described from both ‘reaction of the aircraft’, as well as the ‘reaction inside the aircraft’. These four levels are described below.

Table 1.

Turbulence Reporting Criteria Table, (Aeronautical Information Manual, FAA).

Intensity	Aircraft reaction	Reaction inside aircraft
Light	Turbulence that momentarily causes slight, erratic changes in altitude and/or attitude (pitch, roll, yaw). Report as **Light Turbulence** or Turbulence that causes slight, rapid and somewhat rhythmic bumpiness without appreciable changes in altitude or attitude. Report as **Light Chop.**	Occupants may feel a slight strain against seat belts or shoulder straps. Unsecured objects may be displaced slightly. Food service may be conducted and little or no difficulty is encountered in walking. • Occasional−Less than 1/3 of the time. • Intermittent−1/3 to 2/3. • Continuous−More than 2/3.
Moderate	Turbulence that is similar to Light Turbulence but of greater intensity. Changes in altitude and/or attitude occur but the aircraft remains in positive control at all times. It usually causes variations in indicated airspeed. Report as **Moderate Turbulence** or Turbulence that is similar to Light Chop but of greater intensity. It causes rapid bumps or jolts without appreciable changes in aircraft altitude or attitude. Report as **Moderate Chop.**1	Occupants feel definite strains against seat belts or shoulder straps. Unsecured objects are dislodged. Food service and walking are difficult.
Severe	Turbulence that causes large, abrupt changes in altitude and/or attitude. It usually causes large variations in indicated airspeed. Aircraft may be momentarily out of control. Report as **Severe Turbulence**	Occupants are forced violently against seat belts or shoulder straps. Unsecured objects are tossed about. Food Service and walking are impossible.
Extreme	Turbulence in which the aircraft is violently tossed about and is practically impossible to control. It may cause structural damage. Report as **Extreme Turbulence**	

The objective of such a description is to recognize turbulence by its effects to enable reporting. As our objective is to analyze the effect of various controlled turbulence levels (in a simulator) on touch screen usability, we had to analyze beyond that description to come up with metrics on what such levels of turbulence mean in terms of displacement and acceleration.

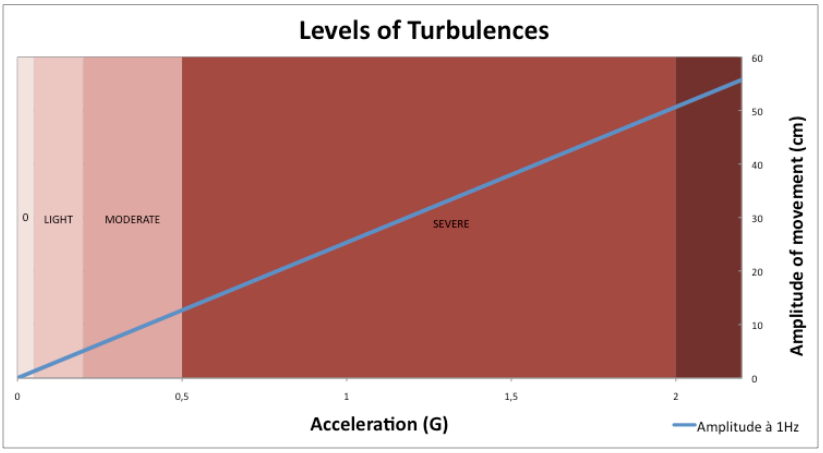


Figure 1. Level of turbulence as a function of acceleration and displacement

Figure 1 represents the relationship between displacement and accelerations. The blue line characterizes the effects at 1Hz. In a sinusoid, displacement of 25 cm per second implies a maximum acceleration of 1m/s.s (1G). One can undergo a maximum of 2 Gs when submitted to a displacement of 50 cm per second. Using such relationship, the various levels of turbulence were approximated with regards to maximum acceleration and maximum displacement withstood. We focused on the effects of vibration being between 0,2 and 7Hz as they are predominant on the control of hand/arm movement (Berthoz, 1981). At one end, for a frequency of 0,2Hz one would need 12 meters of displacement to reach an acceleration of 2Gs. On the other hand, the higher the frequency, the flatter the line, at 7 Hz, one would reach 2gs for a displacement of only 1cm. This preliminary analysis enabled us to focus our search for an adequate simulation platform. what we are looking for should be able to reproduce large displacements at low frequencies (i.e. vibration pods are no solution, as they produce small displacements at high frequencies).

The best solution was the Hexapod. There are many types of hexapods and only the high end ones are able to reproduce the levels of movement characteristic of aeronautical turbulence. We need: 3 axis of acceleration, X, Y & Z, 3 angular accelerations and ultimately a certain capacity of displacement coherent with those encountered in a real aircraft.

Environment simulation design

To complete our initial analysis we started collecting in flight data on a Socata TBM700 aircraft. We used a SGB IG-500N GPS enhanced miniature Attitude and Heading Reference System (AHRS) that delivers attitude and position measurements. It was installed near the center of gravity of the aircraft to collect movement and accelerations (3 angular + 3 linear) at 100Hz when submitted to various levels of turbulence.

Hexapod limits integration (tech evaluation)

The inflight recordings provided flight path (georeferenced) and 100Hz sampling of accelerations (3 angular + 3 linear) on any given path. The data had to be transformed, as an Hexapod cannot process them directly (being fixed to the ground the machine cannot understand georeferenced movements…). The mathematical transformation produced XYZ & 3 Angular accelerations around a stabilized georeference that would be the center of the hexapod, hence producing the turbulence profiles. The Hexapod we chose being the property of the Ecole Nationale Supérieure des Arts et Métiers (ENSAM), we verified that the profiles were within the maximum displacement (+/-50cm) and the maximum accelerations (+/-2G). Minor adjustments were made mostly by limiting replay frequency between 0,2 and 7Hz.

Profile adaptation (expert evaluation)

Working with a Flight test pilot, we adjusted the profiles. First a 22,5 seconds sample was chosen based on diversity (maximum displacements and accelerations within the sample) and lack of symetry (the cahotic nature of turbulence had to be preserved). That sample was the reversed and joined to the original one making a 45s profile. The profile was run sevel times at ¼ displacement (¼D) then at ½ displacement (½D) then full (1D) on an empty seat, for security reasons. Next our test pilot was submitted to the same progressive runs to perform an initial assessment of the profiles. We optimised then the initial Sample (1D) playing on maximum range of displacement, dilating or compressing parts of the sample, adding or reducing accelerations, mostly Z and Y (the front back acceleration being rare in an AC). Every alteration implied a progressive ¼, ½ and Full test with our test pilot. The objective was to provide 3 profiles of turbulence, for the light, moderate & severe levels of turbulence. In the end we selected 6 profiles that should cover the desired turbulence levels to be reproduced on an Hexapod.

Subjective Acceptability Evaluation

Means & Method

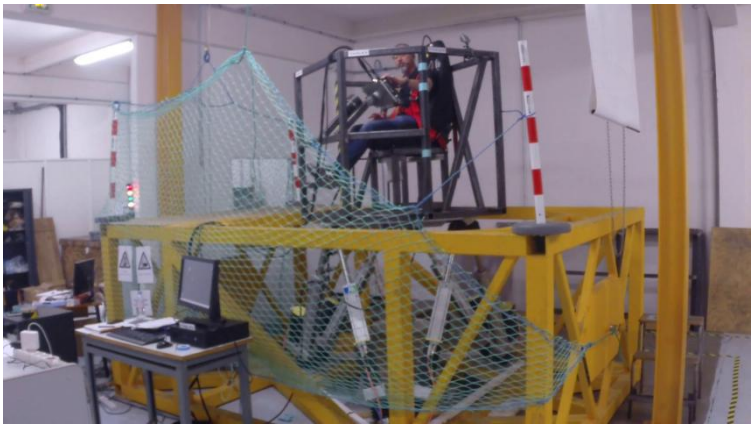


Figure 2. The Hexapod at ENSAM with the test bench on top

The evaluation took place at ENSAM in Bordeaux and had a double objective, first a pilot assessment of the levels of turbulence played by the Hexapod (figure 2) and second a in depht evaluation of touch interactions performance when subjected to various levels of

turbulence. Only the first evaluation is presented here. The Hexapod (+/-2g, +/- 50cm Y,X,Z displacements and 3 axis angular acceleration), property of ENSAM Bordeaux was fitted with a specific "cage" replicating the conformation of the AV2020 cockpit design. The design of the cage was contracteed to ENSAM on detailed specification to ensure the realism of multiple screen positions. Six 45s profiles (table 2) were pre programed on the hexapod and could be played on demand.

Subjects

30 subjects performed this evaluation: 5 left handed, 25 Right handed; 4 women, 26 men; 6 aged 20—29, 11 aged 30-39, 8 aged 40-49, 5 aged 50-59; 7 men had more than 100h of piloting experience (5 with significant flight experience); 9 reported being sometimes sea sick or simulator sick.

Turbulence acceleration profiles

Table 2.

Turbulence profiles to be tested (acceleration in m/s^2)

Profile	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**
Maximum	1,38	2,29	5,51	4,12	5,52	8,11
Mean	0,35	0,65	1,33	1,28	1,53	2,60
Median	0,31	0,57	1,15	1,11	1,32	2,29
Turbulence level	Less than light	Light moderate	Moderate high	Moderate low	Moderate High	Severe

A typical run would comprise the 6 turbulence profiles comparative evaluation then the touch screen evaluation under turbulence and would last 1h 30mn on average. A pause in the middle was added to accommodate the test subject, the experience being somewhat tiring.For the subjective evaluation of the turbulence level, the protocol was quite simple. 7 pilots (more than 100h of piloting experience) ran each profile and were asked to evaluate the realism of the profile as a turbulence one could encounter in an aircraft, second to rate the level of turbulence the profile would compare to. An example of the questionaire is shown table 3.

Table 3.

Subjective evaluation of simulated turbulence profiles

Turbulence profile played	**Does it feel like real in-flight turbulences?**	**Please estimate the level of this turbulence profile**
- - -	not at all ◆——— Perfectly	0 ◆——— light ◆——— moderate ◆——— Severe

Results

Though our sample of pilots was small, our results show a great coherence and little variability. Since that experiment, more pilots have assessed the levels of turbulence that the

hexapod can simulate but with no significant change in the results. Results are shown on figure 3, all profiles have a rating superior to 5/10, 5 out of 6 profiles are juged higher than 8/10 and for all profiles there is very little dispersion in the ratings. The higher the level of turbulence the smaller the dispersion of the pilots evaluation. Levels P1 &P2 were juged light, the levels P3,P4 & P5 were juged as moderate and the last profile P6 was rated severe.

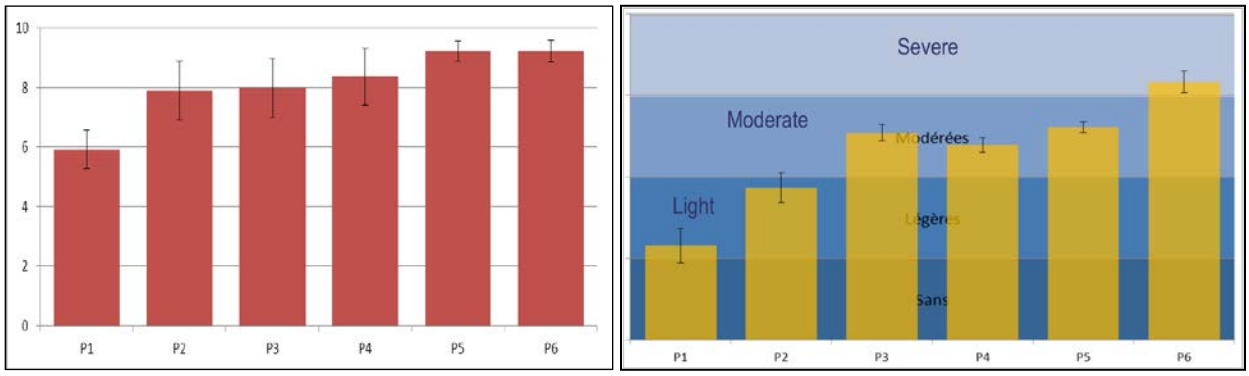


Figure 3. Results: on the left, estimated "realism" of the profile on a scale from 0 (not realistic at all) to 10 (extremely realistic). On the right: estimated level of simulated turbulence profiles.

Conclusion

The pilots interviewed are all agreeing on the quality and representativeness of the hexapod as a means to reproduce turbulence (small distribution of answers).The Hexapod movements are juged similar to real turbulence with a high level of confidence, except for the lowest level. It appears to be less realistic than the others (though still over the average). Pilots reported on debriefing that the low displacements as witnessed on the lowest profile were harder to feel thus to compare to a memorized experience. Though P1 could still be accepted as representative of levels of turbulence, it was not selected in the end for future trial. The Hexapod was juged adapted to the silmulation of light to severe turbulence profiles and while there is a pilot consensus on the quality and representativeness of all the profiles, only 3 levels were chosen on the 6 prepared as being more representative for future evaluations: P2 to simulate light turbulence; P5 to simulate moderate turbulence; P6 to simulate severe turbulence.

References

Berthoz, A. (1981). Effets des vibrations sur l'homme In J. Scherrer et al. ed. lit.-Précis de Physiologie du Travail.

Blanco, R. (2012), Nothing Touches This Market… for The Penny Sleuth. Retrieved from http://pennysleuth.com/nothing-touches-this-market/

Encyclopedia Britannica, aeronautical turbulence definition. Retrieved from http://global.britannica.com/EBchecked/topic/41528/atmospheric-turbulence

Hsieh, C. (2010), Touch Panel Market Research, Retrieved from http://www.displaysearch.com/cps/rde/xchg/displaysearch/hs.xsl/touch_panel_market_analysis.asp

Regulations, F. A. (2008). Aeronautical Information Manual. Retrieved from **http://www.faa.gov/atpubs**

Wagtendonk, W. (2003). *Meteorology for Professional Pilots.* Bay of Plenty, New Zealand: Aviation Theory Centre (NZ) Ltd.

IDENTIFYING REPRESENTATIVE SYMBOLOGY FOR LOW VISIBILITY OPERATIONS/SURFACE MOVEMENT GUIDANCE AND CONTROL SYSTEM (LVO/SMGCS) PAPER CHARTS

Andrea L. Sparko
Stephanie G. Chase, PhD
U.S. Department of Transportation
John A. Volpe National Transportation Systems Center
Cambridge, MA

The Volpe Center developed a questionnaire to examine the representativeness of symbol shapes and the usefulness of information depicted on Low Visibility Operations/Surface Movement Guidance and Control System (LVO/SMGCS) paper charts. One-hundred forty-four pilots were shown a series of symbol shapes and responded "Yes" or "No" to whether they considered each symbol shape representative of a given information type. Symbol shapes were presented at increasing levels of context. Pilots then rated the usefulness of information depicted on LVO/SMGCS charts. Pilots identified representative symbol shapes for a geographic position marking, instrument landing system (ILS) hold line, runway guard lights (RGL), stop bar, and the combination of RGL *and* a stop bar. The general shape was usually perceived as representative regardless of variations in features such as border or fill. Pilot opinions of usefulness generally reflected findings for symbol shape representativeness.

Low Visibility Operations/Surface Movement Guidance and Control System (LVO/SMGCS) is a set of special procedures and visual aids designed to enable safe airport operations below 1,200 ft runway visual range (RVR). Each airport that operates under LVO/SMGCS conditions must have LVO/SMGCS charts that illustrate these procedures and visual aids (FAA, 2012). It is important that symbols shown on LVO/SMGCS charts are easy to recognize and understand, since pilots rarely operate under LVO/SMGCS conditions. The symbols currently used on LVO/SMGCS charts vary across chart providers and airports. Current FAA guidance on LVO/SMGCS (FAA, 1996; 2012) does not contain recommendations specific to LVO/SMGCS charts and, to date, no research has examined human factors considerations for LVO/SMGCS chart symbology. Thus, the FAA requested that the Volpe Center gather data to help identify best practices for LVO/SMGCS symbology.

The current study had two goals. The first goal was to identify what symbol shapes pilots consider representative of information shown on LVO/SMGCS charts. The study also examined whether pilots needed context to identify representative symbol shapes or whether they could identify the symbol shapes alone. The second goal of this study was to gather pilot opinions on the usefulness of depicting different types of information on LVO/SMGCS charts.

Methodology

This study was conducted using an online questionnaire. The following sections describe the participants and questionnaire tasks.

Participants

A total of 144 air transport pilots participated in the study. Participants were required to have category (CAT) III qualified experience (preferably 5+ years) or have military LVO/SMGCS training. To

thank pilots for participating, the names of all participants were entered into a random drawing to receive one of fifty $50 gift cards to Amazon.com.

Symbol Shape Representativeness Task

The task focused on seven information types, defined in Table 1. Definitions were not provided during the task.

Table 1.
Information Types and Definitions.

Information Type	Definition
Geographic position marking (GPM)	Pavement marking used to verify aircraft position
Clearance bar	Lights at the holding position of a taxiway/taxiway intersection
Instrument landing system (ILS) hold line	Pavement marking indicating a holding position at the boundary of an ILS critical area
Runway guard lights (RGL)	Lights at the runway hold short point position of a taxiway/runway intersection, indicating the presence of an active runway
Stop bar	Lights at the holding position of a taxiway or runway intersection, used to indicate clearance to enter a runway when turned off
Combination of RGL *and* stop bar	Collocated RGL and stop bar
Non-movement area	Pavement marking outlining the boundary of an area not under air traffic control

For each information type, pilots were shown a symbol shape and asked to respond "Yes" or "No" to whether they considered the symbol shape to be representative of a particular information type. A total of 60 symbol shapes were shown: 27 of the symbol shapes were real symbols currently used on LVO/SMGCS charts to depict the information type in question, and 33 of the symbol shapes were "foils" that are not currently used on LVO/SMGCS charts to depict the information type in question. Note that a foil symbol shape could be a fake symbol, designed by the researchers, or a symbol that is used on LVO/SMGCS charts to depict a different information type. Foil symbol shapes were used to determine whether pilots accepted variations in symbol shape features (e.g., line thickness or shading) to represent the same information as long as the shape was consistent (e.g., all squares).

Symbol shapes were presented to pilots alone (i.e., on a white background) as well as at increasing levels of context. Most information types were shown at four levels of context, presented one at a time in increasing order:

1. Symbol shape shown on a white background (no context)
2. Symbol shape shown with a single taxiway
3. Symbol shape shown with adjacent taxiways and runways
4. Multiples of the symbol shape, shown with adjacent taxiways and runways

The chart background used to provide context was based on FAA prototype LVO/SMGCS charts. The chart background changed for each information type, but it was the same for all symbol shapes shown within each information type. An example question is provided in Figure 1 for the GPM information type.

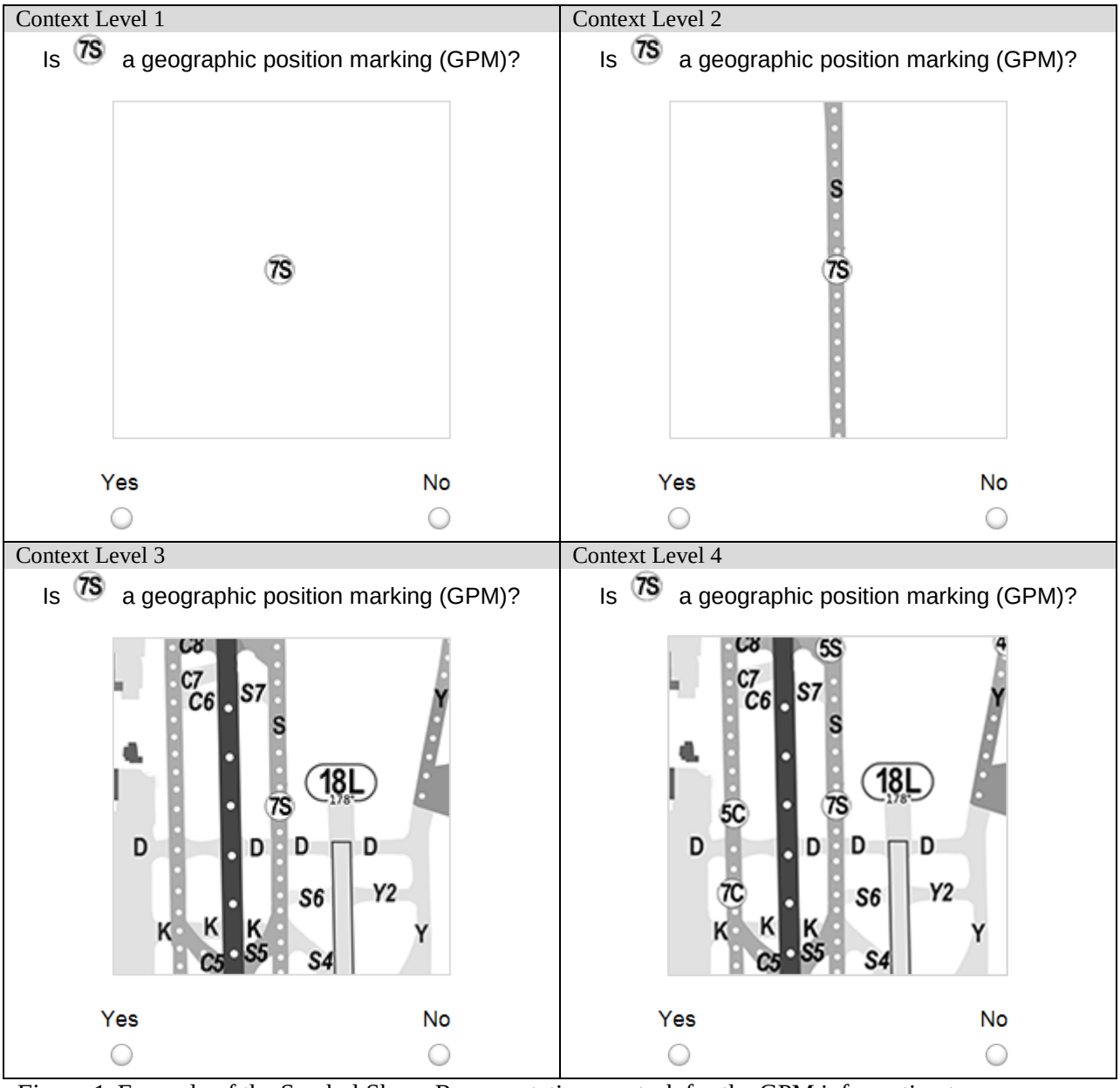


Figure 1. Example of the Symbol Shape Representativeness task for the GPM information type.

Information Type Usefulness Task

In the *Information Type Usefulness* task, pilots were asked to rate the usefulness of nine information types on LVO/SMGCS charts. The nine information types included the seven examined in the *Symbol Shape Representativeness* task (see Table 1) plus two more:

- Approach hold: Pavement marking indicating a holding position at the boundary of a protected approach hold containment area for a runway
- Apron holding point: Pavement marking indicating a holding position at the boundary of an apron

All nine information types were provided in a table with definitions. An excerpt from the table is provided in Figure 2.

Please rate the usefulness of the following information on LVO/SMGCS charts:

	Very Useful	Somewhat Useful	Not Very Useful
Geographic position marking (GPM): Pavement marking used to verify aircraft position	❍	❍	❍

Figure 2. Excerpt from the *Information Type Usefulness* task.

Data Analysis and Results

This section presents the preliminary data and results. More details on this effort and a detailed analysis are documented in Sparko & Chase, in preparation.

Symbol Shape Representativeness

Symbol shape representativeness data were analyzed using chi-square tests comparing the number of "Yes" (representative) responses to the number of "No" (not representative) responses for each symbol shape at each context level. Results were deemed statistically "significant" if there was less than a 5% probability ($p < .05$) that the results occurred by chance. Symbol shapes were considered "representative" if they received significantly more "Yes" responses than "No" responses. Effects of context were observed when the perceived representativeness of a symbol shape changed as context increased. In some cases, pilots needed context to identify a symbol shape as representative.

The representative symbol shapes are shown in Table 2 by information type and need for context. Unless marked as a foil, all of the symbol shapes are used on LVO/SMGCS charts to depict the information type in question. The results show that pilots identified circle shapes as representative of a GPM, regardless of shape outline, fill, text (regular or italic), or context. Ladder shapes were considered to be representative of an ILS hold line, regardless of color or the number rungs. Context was used to identify the foil symbol shape as an ILS hold line; this shape was designed by the researchers to be a variation of the ladder shape with thicker rungs. The fact that pilots needed context to identify this foil symbol shape but not the other ladder shapes suggests that participants associated the the thickness of the rungs with the representativeness of the symbol shape.

Pilots only identified one symbol shape as representative of an RGL, the foil symbol shape , which is used on LVO/SMGCS charts to depict a clearance bar, not an RGL. Context was needed to identify this symbol shape as representative. Pilots identified two *different* symbol shapes as representative of a stop bar. The foil symbol shape , which was designed by the researchers, needed context to be identified as a stop bar. The foil symbol shape , which is used on LVO/SMGCS charts to represent the combination of RGL *and* a stop bar, was identified as representative of a stop bar regardless of context. The same symbol shape was also identified as representative of the combination of RGL *and* a stop bar, suggesting that pilots may not distinguish between these information types on LVO/SMGCS charts. No representative symbol shapes were identified for a clearance bar or a non-movement area. Given that non-movement areas delineate a boundary, this result is not surprising as it may be more relevant to identify whether there is a representative line style used to designate that information type.

Table 2.
Representative Symbol Shapes by Information Type and Need for Context.

Information Type	Context Not Needed	Context Needed
GPM	 (foil)	
ILS hold line		 (foil)
RGL		 (foil)
Stop bar	(foil)	 (foil)
Combination of RGL *and* stop bar		

Information Type Usefulness

Pilot ratings of information type usefulness were analyzed using chi-square tests that compared the number of "very useful," "somewhat useful," and "not very useful" ratings. The results showed that the majority of pilots rated the following information types as "very useful" (all results are statistically significant at $p < .05$):

- ILS hold line (69% of pilots)
- Approach hold (65%)
- RGL (65%)
- Clearance bar (61%)
- Stop bar (60%)
- Combination of RGL *and* stop bar (60%)
- GPM (56%)

Apron holding points received approximately equal numbers of "very useful" (46%) and "somewhat useful" (37%) ratings. Pilots most often rated non-movement areas as "somewhat useful" (49%).

Summary and Conclusion

The results of this study provide input as to what symbol shapes are considered to be representative on LVO/SMGCS charts. For the information types considered here, the findings suggest that pilots may base their perception of representativeness on the overall symbol *shape*. Pilots accepted variations in symbol shape features, such as border, fill, or color, as long as the overall shape was consistent. However, care should be taken when designing symbol shape features to ensure that variations in those features do not alter the symbol shape.

Context helped pilots to identify some "fake" symbol shapes as actual symbols, suggesting that the location of the symbol shape in relation to other chart elements may sometimes be more important than the symbol shape itself. Context may have been helpful in identifying symbol shapes that were unfamiliar or unintuitive. Other symbol shapes, mostly symbols used on LVO/SMGCS charts to depict the information type, were identified regardless of context. Note that the current study did not address information types depicted using variations in linear patterns, which may inherit their meaning based on context.

Pilots identified one symbol shape as representative of both a stop bar and the combination of RGL *and* a stop bar, suggesting that pilots may not distinguish between these information types on LVO/SMGCS charts. Future research might examine the need and operational acceptability of using one symbol shape to represent both a stop bar and the combination of RGL *and* a stop bar on LVO/SMGCS charts.

When asked to give their opinions of information type usefulness, pilots' ratings generally complemented the symbol shape representativeness findings. Pilots identified representative symbol shapes for most of the information types that they considered "very useful" on LVO/SMGCS charts. The one exception was for clearance bars, which were rated "very useful" even though pilots did not identify any representative symbol shapes. Clearance bars, which are usually collocated with GPMs on the airport surface and on charts, may not stand out to pilots on LVO/SMGCS charts, even though pilots believe they are useful. It is also possible that pilots may not know what a clearance bar is by name due to its association with GPMs.

This study is intended to provide data to help the FAA develop best practices for LVO/SMGCS charts. The results provide a general understanding of what symbol shapes may be perceived as representative of certain information types depicted on LVO/SMGCS charts. For some information types, however, additional research may be needed.

Acknowledgements

This paper summarizes the results of a study conducted by the Aviation Human Factors Division at the John A. Volpe National Transportation Systems Center. This research was completed with funding from the Federal Aviation Administration (FAA) Human Factors Division (ANG-C1) in support of the Flight Operations Branch (AFS-410). A comprehensive technical report on this work is currently in progress (Sparko & Chase, in preparation).

We would like to thank our FAA program managers Michelle Yeh, Stephen Plishka, Gina Bolinger, and Tom McCloy, as well as our technical sponsors, Bruce McGray and Terry King. We would also like to thank Andrew Burns, Randy DeAngelis, Sean Flack, and Philip Saenger for their help and advice with this project. Thank you to Katarina Morowsky and Andrew Kendra of the Volpe Center Aviation Human Factors Division for their assistance with study design and administration. Also thank you to the Air Line Pilots Association (ALPA) and FedEx who helped to publicize this study and recruit participants. We also appreciate those participants who told their colleagues about the study and helped us recruit additional pilots.

The views expressed herein are those of the authors and do not necessarily reflect the views of the Volpe National Transportation Systems Center or the United States Department of Transportation.

References

Federal Aviation Administration. (1996). *Surface Movement Guidance and Control System* (AC 120-57A). Washington, D.C.

Federal Aviation Administration. (2012). *Procedures for Establishing Airport Low-Visibility Operations and Approval of Low-Visibility Operations/Surface Movement Guidance and Control System Operations* (Order 8000.94). Washington, D.C.

Sparko, A. L., & Chase, S. G. (in preparation). *Low Visibility Operations/Surface Movement Guidance and Control System (LVO/SMGCS) Chart Symbology*. Washington, D.C.: U.S. Department of Transportation.

HUMAN SPAN-OF-CONTROL IN CYBER OPERATIONS: AN EXPERIMENTAL EVALUATION OF FAN-OUT

Vincent F. Mancuso[1], Gregory J. Funke[2], Monica B. Eckold[2], Adam J. Strang[2]

[1]Oak Ridge Institute for Science and Education, Wright-Patterson AFB, OH
[2]Air Force Research Laboratory, Wright-Patterson AFB, OH

Modern cyber operations require operators to maintain supervisory control of remote computer agents. A current operational concern is the number of agents an operator can control at once. This issue resonates with similar "span-of-control" research conducted in UAV operations (e.g., Cummings & Mitchel, 2008). One way to identify operator span is via "fan-out," a numeric calculation that provides a span-of-control estimate based on system and environmental variables. However, fan-out is a mechanical representation that only accounts for task-characteristics and environmental variables, thus providing an upper bound of human performance that does account for cognition, workload, or work interruptions. The present study compares fan-out estimates against actual human performance in a supervisory control cyber task. Results are discussed and future research trajectories proposed.

Over the last decade, cyber security has become a primary concern for homeland security. This concern is only likely to grow as we continue to develop technologies that employ computer networks to manage our major private (e.g., banking) and government (e.g., nuclear power plant) assets.

Traditionally, cyber security research focuses mostly on computer science and engineering problems, such as the development of algorithms and systems to detect, identify, and mitigate specific threats that exist on computer networks. While this research is critical to our national defense, it does not acknowledge the critical role that human operators play in cyber security. Recently, the Human Factors community has recognized this critical research gap and in response has started an initiative to explore human-centered aspects of cyber operations. Current research has focused on identifying important dimensions of cognition within cyber operations, such as situation awareness (Giacobe, 2012), team knowledge structures (Mancuso & McNeese, 2012), and team collaboration (Rajivan et al, 2013). However, little research has focused explicitly on specific issues related to task load and individual operations management.

Current cyber operations exist within a complex system of human-machine interaction, where operators are tasked with monitoring the activities, efficacy, and progress of intelligent and autonomous computer systems (Tyworth et al., 2013). In these environments, cyber operators supervise intelligent systems as they execute tasks across the network. This task places significant cognitive demand on operators due to the need to maintain situation awareness while dividing attention across a set of dynamic tasks and managing a deluge of information. While novel within the context of human-centered cyber research, these environments share many commonalities with human supervisory control tasks.

In Human Supervisory Control (HSC) tasks, a human-in-the-loop provides an autonomous asset with high-level plans, instructions, and goal directives (Miller, 2004). In operations such as Unmanned Aerial Vehicles (UAVs), a task that is currently performed using manual control but will likely shift to HSC (at least partially) in next-gen operations, researchers have examined human performance in HSC simulations to improve system design, mitigate operator workload, account for individual differences, and identify/optimize human span-of-control. However, similar research has not been conducted on cyber operations despite that fact that current-gen work environments often involve HSC tasks.

Based on this, the purpose of this research was to conduct a preliminary exploration of the translation of previous HSC research to cyber operations. Specifically, in this paper we perform an experimental evaluation of the predictive span-of-control metric known as "fan-out." To assess fan-out's

potential application for cyber-operations, we will compare fan-out estimates against observed human performance in a cyber supervisory control task that requires human operators to control multiple automated agents across a simulated area network.

Fan-Out

Nehme et al. (2008) and Cummings & Mitchel (2008) developed a work-flow model of HSC tasks that parse performance into three components: service time, productive time, and wait time. Service time, also known as interaction time (IT) is the amount of time that it takes a human to service an autonomous asset. This value includes the amount of time it takes an operator to allocate the asset, determine the necessary inputs, and expresses those inputs to the asset via the interface (Olsen & Goodrich, 2003). Next, productive time, also known as neglect time (NT), represents the average amount of time an asset can operate without an operator's intervention (Olsen & Goodrich, 2003). Finally, wait time (WT) is the amount of time that an asset spends in the queue waiting to be serviced by the human. Maximum efficiency can be achieved by minimizing service and wait times, while maximizing productive time.

To help improve HSC systems for maximal performance, some of the variables mentioned above are combined to obtain computational estimates of fan-out (FO) according to the equation, $FO = (NT / IT) + 1$ (Dixon, Wickens & Chang, 2005; Miller, 2004; Olsen & Goodrich, 2003; Sheridan, 1992). From a linear throughput perspective, FO indicates the number of homogenous assets a single operator can control at once without interference or drag (Olsen & Wood, 2004). Thus, FO can (and has been) used as a theoretical upper bound for estimates of operator span-of-control. For example, Cummings & Mitchell (2008) found that operators performed at approximately 36% below fan-out estimates in a UAV simulation and speculated three reasons for sub-optimal performance: a) WT in the human decision-making queue (WTQ), b) interaction wait time (WTI), and c) wait time due to loss of situation awareness (WTSA). WTQ occurs when an asset goes unattended while an operator completes their decision making task. WTI, which is very similar to IT, includes time that the operator spends determining the appropriate action and communicating it to the asset. Finally, WTSA includes time that the operator spends away from the asset as they perceive elements in the environment, comprehends their meanings and makes future predictions of their status. While conceptually quite simple, capturing these metrics for the purpose of predicting human span of control in a computational model can be quite complex, especially WTQ and WTSA.

Based on such evidence, we expected that, assuming accurate inputs, that estimates of FO in a cyber HSC simulation (BotNET Operator Ratio Determination; BOARD v1.5) would correlate with human span-of-control. However, due to the inherent complexity of cyber-operations, we also expected that actual span-of-control will be significantly lower than the estimates generated by the fan-out equation. Together, these finds would allow us to pinpoint the theoretical and realistic span-of-control for HSC cyber operations.

Methods

Simulated Task Environment

BOARD is a human-in-the-loop scaled world simulation that is set in the context of cyber supervisory control operations. In this simulation, participants are tasked with remotely controlling computer agents using a command line interface. During the tasks, participants interact with three main components: a) the agent control window, b) the agent beacon monitor, and c) the mission commander messaging system (Figure 1).

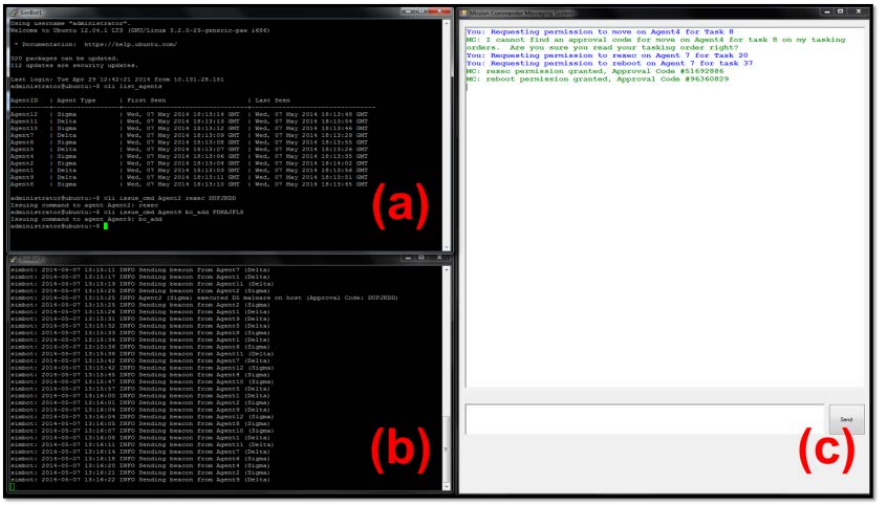


Figure 1. The BOARD simulation user interface depicting the agent control window (a), the agent beacon monitor (b), and the mission commander messaging system(c)

These interface elements allow participants to remotely control computer agents using a set list of commands, monitor the progress and state of varying missions, and communicate with a mission commander to obtain permission to execute restricted commands.

At the start of each experimental trial of BOARD performed in this experiment, participants received an Operating Tasking Order (OTO) print out and were given access to a set of computer agents. Each OTO comprised a mission set, where missions within the set were characterized by an assigned "agent type" and three steps that must be accomplished (in serial order) to successfully complete the mission (see Table 1 for an example). In each mission step, participants were provided with a command they must execute and an authorization code to run the command. In order to complete the mission, the participant needed complete all three steps in order, using an agent of the correct type. If they enter the incorrect authorization code, the command may not execute and they cannot complete the mission.

Table 1.
Example BOARD Operating Tasking Orders.

	Type	Step 1	Step 2	Step 3
1	Delta	getlog "yxdembfl"	grab "xtizjluw"	phish "sblukllp"
2	Sigma	getlog "hlhsggwg"	rename "lgxustfx"	bl_url "rahtdkcl"
3	Delta	bc_add "hihhohwv"	bc_add "upfyvedo"	rexec

In some cases, authorization codes were not provided to participants (e.g. Mission 3, Step 3 in Table 1). In these instances, the participant was forced to utilize the Mission Commander Messaging system to request an authorization code. This manipulation was included to represent cyber rules of engagement, which require command authorization before executing sensitive missions. After requesting the code, the mission commander responded with an authorization code in a variable amount of time (between 10 and 60 seconds). While participants waited for authorization, they were permitted to move on and complete other missions (but not subsequent steps in the same mission).

Measures

In this study, fan-out measures were customized for each participant from typing ability and average response time for agents during a pre-experiment simulation. Prior to the experimental task, each participant was asked to complete a typing test lasting two minutes. Words-per-minute results were adjusted based on errors, and then multiplied by 5 to calculate estimated characters per minute (CPM;

Arif & Stuezlinger, 2009). Using the equation proposed by Olsen & Goodrich (2003), IT was calculated by comparing the CPM to the average character length of agent execution commands (37 characters) and NT was calculated from the average time it took an agent to execute a command (45 seconds).

Observed operator span-of-control was calculated as the maximum number of agents operating in a one-minute window averaged across each 20 minute trial.

Participants

Twenty-one participants (17 Male, 4 Female), drawn from local universities and United States Air-Force agencies, were financially compensated for their participation (on-duty Airmen received no compensation outside of their regular duty pay). All participants were between the ages of 19 and 45 (M = 23.95, SD = 5.15) and had some level of experience with Command Line Interfaces (CLI) and/or computer programming languages. Four participants had previous cyber security experience (classes, professional experience, etc.).

Procedures

Experimental sessions were conducted in the AFRL Cyber Integrated Performance and Human Effectiveness Research (CIPHER) Laboratory. Prior to the task, participants completed three phases of training. First participants completed a short self-paced computer based training (CBT) which took approximately 15 minutes. Following CBT, the researcher guided the participant through a short practice scenario using the BOARD environment to familiarize themselves with the computer interface, task requirements, and rule. Following the guided training scenario, participants were instructed to complete a second training scenario independently, while the researcher observed their progress and corrected any mistakes that were made. After the participant had reached satisfactory performance on the independent training scenario, participants completed four experimental trials, each lasting 20 minute. During each trial, participants were provided a different number of autonomous agents (4, 8, 12 or 16) with which to complete their assigned missions. These manipulations were part of a larger investigation and are peripheral to the central question of the current study (i.e., evaluation of the accuracy of fan-out measures in cyber). As such, results pertaining to their effects on performance are being prepared for presentation elsewhere. Estimates of operator span-of-control presented herein are means calculated across the four scenarios. The total duration of the experiment was approximately 3 hours.

Results

Fan out estimations were calculated for each individual based on their typing speed (cpm; M = 264.29, SD = 70.33) and the average NT (45 seconds). Results showed no significant correlations between fan-out estimates and actual span-of-control, r = 0.19, p > .05 (Figure 2).

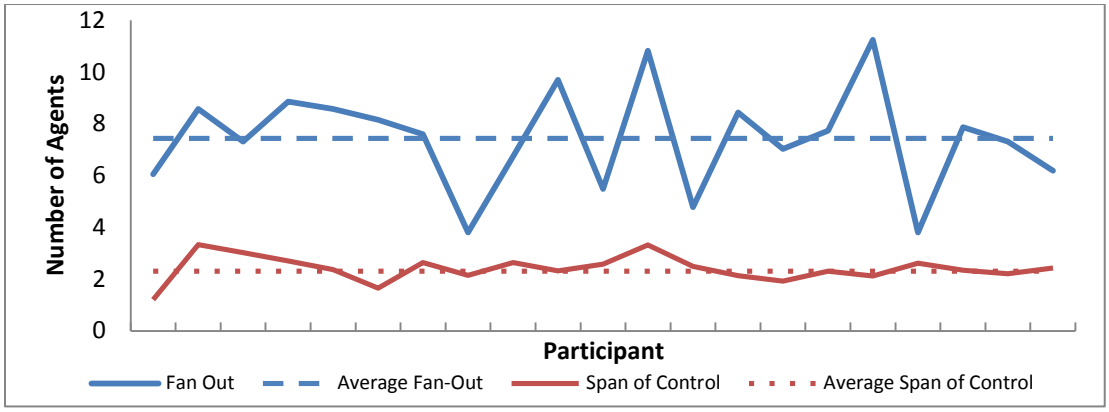


Figure 2. Participants Span-of-Control compared to Fan-Out Estimates

However, consistent with prior research, fan out estimates were found to be significantly higher than the actual span of control. A paired sample *t*-test revealed a significant difference between the predicted fan-out values ($M = 7.43$, $SD = 1.99$) and average span-of-control ($M = 2.27$, $SD = 0.49$), $t = 12.70$, $p < 0.05$.

Discussion

Previous research has presented fan-out as a model upper bound of human span-of-control in several HSC tasks (Crandall & Cummings, 2003). While it is to be expected that actual span-of-control will fall below this value, generally these metrics have been found to be somewhat correlated (Cummings & Mitchell, 2008). Surprisingly, we found that fan-out estimates were not significantly correlated with observed span-of-control in the BOARD cyber HSC simulation. However, we did find a consistent difference between the predicted values (via fan-out) and observed span-of-control. Given that observed span-of-control fell well below the estimates (close to a 100% difference, as compared to the 30-50% difference presented in other HSC tasks, e.g., Cummings & Mitchell, 2008), the current results suggest that the inherent complexity of cyber operations entails a higher cognitive load, reducing operator span well below fan-out estimates.

One caveat of these findings, however, is that fan-out was calculated from a relatively simplistic formulation proposed by Cummings & Guerlain (2004); other researchers have proposed more complex approximations. In their research, Mitchell, Cummings and Sheridan (2003) proposed the addition of wait times to the denominator of the equation. In their expanded formula, wait time is a combination of the time the asset spends in the queue before receiving instructions from the operator and time attributed to operator reorientation and activities supporting situation awareness. Another interpretation by Crandall and colleagues (2005) utilizes other metrics, such as neglect and interaction impact, to calculate an overall performance metric for each asset based on all possible values for interaction and neglect time.

Conclusion

Building from previous HSC research, the purpose of this study was to investigate how the metric of fan-out translated to cyber operations. Using a simulated autonomous agent control task, we evaluated participants' actual span-of-control, and compared them to the estimates calculated by a popular interpretation of the fan-out equation. Our findings indicated that fan-out did not provide an accurate representation of human performance in our task. Future research should focus on identifying the cognitive requirements of cyber work, so that more mature equations for predicting span of control can be developed.

References

Arif, A. S., & Stuerzlinger, W. (2009, September). Analysis of text entry performance metrics. In 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH). Toronto, CA 26-27 September (pp. 100-105). IEEE.

Crandall, J. W., & Cummings, M. L. (2007). Identifying Predictive Metrics for Supervisory Control of Multiple Robots. IEEE Transactions on Robotics, 23(5), 942 - 951. doi:10.1109/TRO.2007.907480

Crandall, J. W., Goodrich, M. A., & Nielsen, C. W. (2005). Validating human-robot interaction schemes in multitasking environments. IEEE Transactions on Systems, Man, and Cybernetics, 35(4), 438-439. doi:10.1109/TSMCA.2005.850587

Cummings, M. L., & Guerlain, S. (2004, September). An interactive decision support tool for real-time in-flight replanning of autonomous vehicles. AIAA 3rd Unmanned Unlimited Technical Conference, Workshop and Exhibit, Chicago, Il. doi: DOI: 10.2514/6.2004-6526

Cummings, M. L., & Mitchell, P. J. (2008). Predicting Controller Capacity in Supervisory Control of Multiple UAVs. IEEE Transactions on Systems, Man, and Cybernetics, 38(2), 451-460. doi:10.1109/TSMCA.2007.914757

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission Control of Multiple Unmanned Aerial Vehicles: A Workload Analysis. Human Factors, 47(3), 479-487. doi:10.1518/001872005774860005

Giacobe, N. A. (2013, September). A Picture is Worth a Thousand Alerts. In Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society, 30 September - 4 October (pp. 172-176). San Diego, CA, HFES.

Mancuso, V. F., & McNeese, M. D. (2012, September). Effects of Integrated and Differentiated Team Knowledge Structures on Distributed Team Cognition. In Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society, 22-26 October (pp. 388-392). Boston, MA. HFES

Miller, C. (2004). Modeling human workload limitations on multiple UAV control, Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, CA: HFES, 526

Mitchell, P. M., Cummings, M. L., & Sheridan, T. B. (2005, May). Management of multiple dynamic human supervisory control tasks. In 10th International Command and Control Research and Technology Symposium. MacLean, VA (pp. 1 - 11).

Nehme, C. E., Kilgore, R. M., & Cummings, M. L. (2008, September). Predicting the impact of heterogeneity on unmanned-vehicle team performance. In Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society, New York, NY 22-26 September (pp. 917-921). HFES.

Olsen Jr, D. R., & Wood, S. B. (2004, April). Fan-out: measuring human control of multiple robots. In Proceedings of the SIGCHI conference on Human factors in computing systems. Vienna, Au 24-29 April (pp. 231-238). ACM.

Olsen, D. R., & Goodrich, M. A. (2003, September). Metrics for evaluating human-robot interactions. In Proceedings of PERMIS, Gaithersburg, MD (p. 4-11).

Rajivan, P., Janssen, M. A., & Cooke, N. J. (2013, September). Agent-Based Model of a Cyber Security Defense Analyst Team. In Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society, 30 September - 4 October (pp. 314-318). San Diego, CA, HFES.

Sheridan, T. B. (1992). Telerobotics, automation, and human supervisory control. Cambridge, MA: MIT Press.

Tyworth, M., Giacobe, N., Mancuso, V., McNeese, M., & Hall, D (2013). A Human-In-The-Loop Approach to Understanding Situation Awareness in Cyber Defense Analysis. EAI Endorsed Transactions on Security and Safety, 13(2), 1-10.

PERSPECTIVES OF UNSUCCESSFUL AIR TRAFFIC CONTROL SPECIALISTS

Linda G. Pierce
FAA Civil Aerospace Medical Institute
Oklahoma City, OK
Cristina L. Byrne
FAA Civil Aerospace Medical Institute
Oklahoma City, OK

Approximately one-quarter of air traffic controller trainees fail field training at their first facility assignment. In some cases, those who fail the training qualifications at their first air traffic control facility assignment are allowed to transfer to a less complex facility. We surveyed a sample of these controllers to identify their perceptions of work-related and external factors that contributed to their failure and subsequent request for reassignment. For example, although these controllers were selected to work at their first facility, in part, based on their aptitude for the job, some said they simply could not do the work at the level that was required. Others indicated that factors such as facility culture or training methods used by on-the-job training instructors might have contributed to their failure. This research is a first step in understanding why controllers fail training.

The Federal Aviation Administration (FAA) employs approximately 14,500 air traffic control specialists (ATCSs). These ATCSs, referred to as controllers, work at en route and terminal air traffic control (ATC) facilities across the United States. Applicants for ATCS vacancies must meet test criteria that have a demonstrated relationship with achieving certification as an ATCS. The vast majority of applicants do not meet these criteria. Those relatively few applicants meeting the criteria and subsequently hired by the FAA as trainees, also called developmental controllers, will be in training for the first one to three years of their employment. Most will attend training at the FAA Academy before undergoing site-specific training at an ATC facility. However, some developmental controllers with prior experience as civilian or military controllers may bypass the Academy and begin field qualification training at their first assigned facility immediately after hiring.

Field qualification training for developmental controllers includes a combination of classroom instruction, computer-based instruction, simulated exercises, and on-the-job training (FAA, 2013a). If successful in all stages of field qualification training, the developmental controllers become certified professional controllers (CPCs). Most often, developmental controllers who fail field qualification training are terminated from employment. However, in some cases, field-training failures may request reassignment to a lower volume and less complex ATC facility. If reassignment is requested, the training history of the developmental controller is reviewed by the National Employee Services Team (NEST), a team within the FAA's Air Traffic Organization, and a determination is made to retain or terminate the employee. Training failures from either an en route or terminal facility may request transfer but they are usually allowed only to transfer into lower-level terminal facilities (FAA, 2013b).

Training failures are costly, whether the developmental controller leaves the FAA or is reassigned to a lower level ATC facility. On average, the cost to train one developmental controller for one year is about $180,000 (FAA, 2014). Our research objective is to determine why developmental controllers fail field qualification training at their first facility. Are unsuccessful developmental controllers simply unable to control air traffic, despite demonstrating the aptitude for ATC during the selection process and, in most cases, succeeding at the FAA Academy? Are there factors other than ability that contribute to failure in ATC field qualification training? If other factors are involved, what are they? Are these factors internal to the FAA such as the culture of the facility or perhaps the training policies or practices of the on-the-job instructors or are these factors external in nature, related to family issues or facility location?

This research project is a first attempt to understand the factors that contribute to training failures from the perspective of unsuccessful developmental controllers, specifically those who request and are allowed to transfer to lower-level ATC facilities. While these developmental controllers may be biased or have a limited understanding of some of the factors involved, understanding their perspective is a useful first step in considering strategies to reduce training failures.

Method

Participants

The 100 developmental controllers who volunteered to participate in this study were solicited from among all developmental controllers sent to the FAA Academy from February 2014 through January 2015 for training after failing field qualification training at their first facility and prior to beginning training at their second facility. Initially, we did not collect demographic data (e.g., age, gender, or race or national origin (RNO)) to encourage participation by ensuring anonymity. We added demographic questions to our survey approximately half way through data collection after determining that participants were willing to provide the information. The average age of the 42 participants reporting was 31.32 (*SD* = 3.17). There were 35 males and 7 females in our sample, and their RNO was reported as follows: 31 White; 4 Hispanic; 4 Black; 2 Asian or Pacific Islander; and 1 American Indian or Alaskan Native.

Materials and Procedure

Researchers at the FAA's Civil Aerospace Medical Institute (CAMI) created the Controller Transfer Questionnaire (CTQ) for use in this research. While the FAA encourages completion of an on-line exit survey for those who leave the agency or federal service (including retirement), there is nothing similar for use with ATCSs who fail training at their first facility but are allowed to transfer to a lower level ATC facility. In addition, the FAA's exit survey does not cover the factors thought to contribute to training failure, the subject of this research.

The CTQ has 46 questions divided into the following sections: entry, reassignment, training, performance, culture, and feedback. In each section, participants were asked questions regarding their perceptions of factors that had contributed to their failure in training and subsequent request for reassignment. They were also asked how satisfied they were with each type of training, their opinion about the best and worst parts of training, and how they would improve training. We used a variety of question response formats in developing the CTQ. The most predominant type of question uses a Likert response format, with responses ranging from one to seven on a defined scale (e.g., agreement, satisfaction, and difficulty). A secondary type of question required respondents to mark all applicable items. For example, one question asked respondents to identify what they liked best about being a controller and provided several response options like salary, benefits, prestige, or the challenge of the work. We used these response formats to minimize the number of open-ended questions on the CTQ, thereby facilitating response consistency and supporting data analysis. Each question formatted in this way included an *Others* item to allow for responses not listed as options. The CTQ was slightly modified after analyzing data collected from the first 58 participants. The modifications, based on frequent responses to the *Others* item, were used to increase the number of response options provided for selected items. We analyzed a subset of CTQ items specifically related to participants' perceptions of job factors that contributed to their failure.

Results

Work-Related Factors

Work-related factors seen as contributing to requests for reassignment are shown in Table 1. Of the 100 participants, 83 selected at least one response on this question, and 22 of the 83 selected more than one response. Work-related factors selected most often as contributing to requests for reassignment were *Could not do the work* and *Did not like the facility*. Most of the reasons listed in the *Others* category explained item responses selected or mentioned external factors that were not work-related, and are addressed in another section of the survey.

Table 1.
Work-Related Factors.

Were there work-related reasons for requesting reassignment? (Mark all that apply)

Response Options	Frequency
Could not do the work	24
Did not like the work	8
Did not like work hours/schedules/shiftwork	1
Did not like my co-workers	13
Did not like my trainer(s)/instructors(s)	8
Did not like my managers	7
Did not like the facility	18
Others (Please list below)	48

Note. Data are based on the number of respondents selecting a particular response.

Could Not Do the Work. In the comments under *Others*, one participant selecting *Could not do the work* said, "Level 12 radar work was above my capabilities." However, another participant, also selecting *Could not do the work* said "I still think I can do EnRoute, but I didn't pass the skill checks for some reason. Just as they let me go, I really started to get the hang of it." The latter comment is consistent with responses made by many survey participants on other survey questions related to perceptions of training performance (see Table 2). Whether selecting or not selecting *Could not do the work*, most participants thought they could have certified as a controller if they had stayed in training at their first facility or moved to a different facility of the same type. They also thought they were progressing well in training. Although, as shown in Table 2, those selecting *Could not do the work* for all three items scored significantly lower than those who had not selected *Could not do the work* as a work-related factor in requesting reassignment.

Table 2.
Individual Perceptions of Training Performance

Question 1 Definitely Not to 7 Definitely Yes	*Could not do the work*		
	Selected	Not Selected	Significance of the Difference
Do you think you could have certified as a controller if you had stayed at your facility?	M = 5.33, SD =1.17 (N = 24)	M = 5.96, SD = 1.48 (N = 74)	$t'(48.88) = -2.13$, $p = .01$
Do you think you could have certified as a controller at a different	M = 5.79, SD =1.06 (N = 24)	M = 6.55, SD = .83 (N = 74)	$t'(32.60) = -3.21$, $p = .003$

facility of the same type?			
Did you feel you were progressing well in training at your facility?	M = 4.26, SD = 1.54 (N = 23)	M = 5.19, SD = 1.44 (N = 75)	t'(34.57) = 2.57, p = .01

Did Not Like the Facility. A participant selecting *Did not like the facility* said "I felt poorly treated. The work was antiquated and done in a way that I felt was beneath me or any smart person." Another participant who had also selected *Did not like the facility* said, "Most people in the center are miserable and angry. The ones that try and help get ignored." Eight of the 18 participants selecting *Did not like the facility* also selected *Did not like my co-workers*. The perceived culture of the facility may have contributed to participants not liking the facility. Of the 84 participants responding to a question on their perception of the facility culture, slightly more participants rated the culture as being unsupportive/apathetic or hostile than friendly or competitive (27 selected more than one response). Very few participants said their facility culture was supportive (see Table 3). Of the 18 participants who selected *Did not like the facility* as a reason for requesting reassignment, only one choose *Supportive* as the predominant organizational culture at their facility. Most (14 of 18) said the facility culture was either *Unsupportive/apathetic* or *Hostile*.

Table 3.
Facility Culture

What was the predominant organizational culture at your facility? (Mark all that apply)

Response Options	Frequency
Friendly	20
Competitive	22
Supportive	9
Unsupportive/apathetic	25
Hostile	30
Others (Please list below)	36

Note. Data are based on the number of respondents selecting a particular response.

Did Not Like My Trainer(s)/Instructor(s). Although only eight participants selected the item response *Did not like my trainer(s)/instructor(s)* most participants thought the training process needed to be improved. On the item *Do you believe that the training process needs to be improved?* the average response for all participants was 6.20 (SD = 1.31) on a scale from 1 (Definitely Not) to 7 (Definitely Yes). Thirty of the 71 recommendations made by the participants for improving training methods specifically mentioned facility trainers or on-the-job training instructors. Sample comments were:

> "Trainers must be better trained in how to teach. In what way specifically I cannot say for sure, but the ability to do a job is absolutely not the sole requirement for being an effective teacher of that job."
> "Trainers have to love to train people, it is voluntary work. But it seems that they just love the extra income, and feel the power to ridicule the non-CPCs."
> "Find CPC's who want to train. Don't have trainees change trainers. Stick with one or two trainers who want to train. Most CPC's don't want to train and adopt bad attitudes with the trainee until he/she fires them for someone else."
> "I believe that controllers who really want to train are those who should train and go to an OJTI class (extensive class). Training and being a new hire is already stressful enough; training process should be a team effort (whole crew)."
> "Have people that want to train, train. A lot of OJTIs didn't want to train."

"Do a better job selecting trainers, not everyone is capable of being a good trainer."

External Factors

External factors seen as contributing to the participants' request for reassignment are shown in Table 4. Of the 100 participants, 53 selected a response on this question, and 16 of the 53 selected more than one response. Family, location, and cost of living were selected most often. *Other* reasons listed were primarily to explain item responses or were work-related, not external factors. An explanation of an external circumstance reported by several participants was their own health or health of a family member.

Table 4.
External Factors.

Were there external circumstance that drove your request for reassignment? (Mark all that apply)

Response Options	Frequency
Family	24
Childcare	5
Spouse	4
Cost of Living	11
Housing	3
Schools	3
Location	13
Commute	5
Others (Please list below)	23

Note. Data are based on the number of respondents selecting a particular response.

Discussion

The FAA categorizes developmental controllers transferring from higher-level to lower-level facilities as training failures (FAA, 2011). As mentioned previously, training failures are costly to the FAA. In a recent study by Pierce, Broach, Byrne, and Beckley (2014), the failure rate for developmental controllers who started field qualification training from 2007 to 2011 and completed training by June 2014 was 26.3%. The percentage of training failures varied greatly by type of ATC facility, ranging from 15% at tower only facilities to 45% at terminal radar approach control (TRACON) facilities. Training failures are also costly to the developmental controller. While difficult to quantify the emotional costs, we know that developmental controllers often leave employment in other occupations when selected for ATC training that they may or may not be able to return to, if they fail training as a controller. They often also must move to duty stations far from their current residence. Thus, to reduce costs to both the FAA and the developmental controller, strategies to decrease training failures are needed.

Our goal was to understand what developmental controllers who had failed training, but been allowed to transfer to a lower-volume and less-complex ATC facility, thought contributed to their failure to succeed at their first facility. Identifying contributing factors could potentially lead to the development of strategies or interventions to decrease the likelihood that developmental controllers would fail training at their first facility.

Based on the data collected thus far, it would seem important to examine issues related to organizational culture. Is the ATC environment hostile toward developmental controllers? What kind of

support do developmental controllers need to manage family matters during this time of transition? It also seems that some strategy may be needed to support and improve the performance of on-the-job training instructors. Should all CPCs be allowed to be on-the-job training instructors? Should there be some way of assessing the effectiveness of on-the-job training instructors? How should they be trained and what is the best strategy for matching developmental controllers with on-the-job training instructors?

These and other questions will be addressed in follow-on research. We plan to continue administering the CTQ to developmental controllers allowed to transfer to lower level ATC facilities after failing at a first facility and to counter a limitation inherent in this project by extending our data collection to include others involved in the training process. For example, the perspective of (a) successful developmental controllers, (b) developmental controllers who failed and were terminated, (c) OJTIs, and (d) other facility training personnel should be gathered to broaden our understanding. In addition, we plan to supplement the survey-based data with additional analyses, based on more quantitative, performance-based data. For example, assessing the extent to which developmental controllers who are allowed to transfer are successful in training at their second facility might indicate a need for a succession plan in which developmental controllers enter at less complex facilities and move to more complex ones after reaching CPC. In addition, identifying facilities with relatively high training failures may allow for a targeted approach to data collection and implementation of interventions. The research reported in this paper represents the first step in a multi-year, multi-method approach to decrease field training failure rates of air traffic controllers.

Acknowledgement

Research reported in this paper was conducted under the Air Traffic Program Directive/Level of Effort Agreement between the Human Factors Division (ANG-C1), FAA Headquarters, and the Aerospace Human Factors Research Division (AAM-500) of the Civil Aerospace Medical Institute.

References

Federal Aviation Administration (2011, May). *National Training Database Guidelines*. Washington, DC: FAA.

Federal Aviation Administration (2013a). *JO 3120.4N Air Traffic Technical Training*. Washington, DC; FAA.

Federal Aviation Administration (2013b). Human Resources Policy Manual, EMP-1.14a (ATCS) Employment Policy for Air Traffic Control Specialist in Training. Washington, DC; FAA.

Federal Aviation Administration. (2014). *Post-hearing questions for the record submitted to Ms. Patricia McNall from Senator Claire McCaskill.* Hearings before the Subcommittee on Financial and Contracting Oversight of the Senate Committee on Homeland Security and Government Affairs, 113th Congress , 2nd Session, January 14, 2014. (http://www.hsgac.senate.gov/download/?id=2B709C2C-DB78-4B00-9B7F-325B940B8EF7).

Pierce, L.G., Broach, D., Byrne, C.L., & Bleckley, M.K. (2014). *Using biodata to select air traffic controllers.* (Report No. DOT/FAA/AM-14/8). Washington, DC; FAA Office of Aerospace Medicine.

AN EVALUATION OF THE UTILITY OF AT-SAT FOR THE PLACEMENT OF NEW CONTROLLERS BY OPTION

Cristina L. Byrne and Dana Broach
FAA Civil Aerospace Medical Institute
Oklahoma City, OK

In this study, we investigated the utility and fairness of using the Air Traffic Selection and Training (AT-SAT) test battery to place Air Traffic Control Specialist (ATCS) applicants into terminal or en route facilities. While results of statistical analyses indicated that AT-SAT could be considered a valid tool for use in placement, based on technical considerations only, it was concluded that it should not be used in that way due to lack of utility and potential for adverse impact. If the FAA were to use AT-SAT for placement, the risk of additional adverse impact and pay disparities should be evaluated against the marginal utility of placement in terms of changes in field training success rates.

The air traffic control specialist (ATCS, or controller) occupation is considered to be an intellectually challenging, important, and prestigious career field by the majority of recently hired developmental controllers (Cannon & Broach, 2011). The Federal Aviation Administration (FAA) projects hiring approximately 1,300 new controllers per year over the next five years to replace retiring controllers (FAA, 2014). Excluding rehires or others with previous ATCS experience, it is required that applicants receive a passing score on an aptitude test to be hired into the occupation (U.S. Office of Personnel Management [OPM], 2013). Currently, the computer-administered Air Traffic Selection and Training (AT-SAT) test battery is the aptitude test used by the FAA to assess applicants under the OPM occupational qualification standards.

The validity of AT-SAT as a predictor of ATCS job performance was demonstrated in two concurrent, criterion-related validation studies. The first study was reported in 2001 by Ramos, Heil, and Manning (2001a, b). Approximately 1,000 incumbent en route controllers took the proposed test battery and job performance data were collected. The correlation between scores on the test battery and a composite of the job performance measures collected was .51 without any corrections for range restriction or criterion unreliability. With correction for incidental range restriction, the correlation was .68 (Waugh, 2001). The American Institutes for Research (AIR®, 2012) conducted the second study, named the Concurrent Validation of AT-SAT for Tower Controller Hiring (CoVATCH). Incumbent air traffic control tower (ATCT) controllers (N = 302) took the current operational version of the AT-SAT test battery and job performance measures were collected (see Horgen et al., 2012). The correlation between a regression weighted composite of AT-SAT subtest scores and the composite of the two criterion measures was .42 without any statistical corrections (AIR®, 2012). These two studies independently demonstrated that AT-SAT is a valid predictor of ATCS job performance.

Before placement can be discussed, it is useful to understand the nature and structure of the organization within the FAA responsible for air traffic control operations and facilities. This organization, called the Air Traffic Organization, or ATO, can be divided into two major partitions, also referred to as options: Terminal Services and En Route/Oceanic Services. New hires can be placed into either the Terminal option or the En Route option. At en route centers, controllers handle high altitude air traffic between airports, work that is generally considered very complex and demanding. At TRACONs, controllers direct traffic within about 50 miles of an airport, usually during initial climb and final descent of the aircraft. This work can also be considered very demanding. The work at ATCTs involves directing air traffic on the runways and in the immediate vicinity of the airport, as well as issuing takeoff and landing clearances. This type of air traffic control (ATC) is generally considered somewhat less complex and less demanding than radar ATC, but that can vary greatly by location. Historically, the failure rate in on-the-job training for new controllers has been higher in en route facilities (Manning, 1998). Controller positions at en route centers generally have the highest pay grades in the occupation. Controller positions in towers generally have lower pay grades than en route positions. Thus, there are both organizational (success and failure rates in facility on-the-job training) and individual economic consequences attached to placement decisions. Moreover, because placement affects the terms and conditions of employment (especially starting pay), it is an employment decision as defined by the Uniform Guidelines on Employee Selection Procedures (29 C.F.R. § 1607.2B) (EEOC, 1978). Therefore, using AT-SAT scores for placement, as recommended by the Department of Transportation Inspector General (2010), must be validated.

To use a test score for placement purposes, the relevant professional standards and principles require “evidence that scores are linked to different levels or likelihoods of success among jobs” (American Educational Research

Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 160). Relevant evidence might include a pattern of differential relationships between predictors and criteria by job type (Society for Industrial and Organizational Psychology [SIOP], 2003). In their analysis of the CoVATCH data, AIR® (2012) was not able to provide any evidence suggesting that AT-SAT could be used for placement by facility level which is based on complexity and traffic load. However, they did report evidence for validity by option in that the regression equation (i.e., the weight given to each subtest) for tower was not identical with the equation for en route as reported by Ramos et al. (2001a, b). Unfortunately, there were few differences in the recommended option placement when the two equations were used to hypothetically place individuals in their sample of 300 tower controllers. AIR® concluded that AT-SAT might be used for placement by option but further analyses were needed. Thus, the purpose of the current study was to extend the AIR® analysis by using AT-SAT validation data collected from both en route and terminal controllers.

Placement Rules

AT-SAT scores might be used for placement in many different ways. For example, persons with scores above some cut-off might be assigned to the en route option, while persons with scores below that cut-off might be assigned to terminal. Or scores might be categorized into ranges, with persons in the lowest range assigned to one option, persons in the highest range assigned to the other option, and persons with scores in the middle range assigned to either option, depending on agency needs. Thus, the first step in this analysis was to decide how AT-SAT scores would likely be used for placement. To use AT-SAT for placement decisions, AIR® suggested computing a score for each option, based on the option-specific regression equations. As no AT-SAT validation study has been conducted specifically for TRACON positions, the equation derived from the tower sample was used to represent the entire terminal option. The applicants would then be assigned to a score band within each option. For example, an applicant could be classified as Well-Qualified Terminal and Qualified En Route (or vice versa), Well-Qualified in both, or Qualified in both. Current use of AT-SAT defines Well-Qualified as a score of 85-100, Qualified as a score of 70-84.9, and Not Qualified as 69.9 and below.

The placement procedure suggested by AIR® is feasible but has three drawbacks. First, the overall ranking of an individual (which impacts hiring decisions) is confounded with their ranking within an option (which impacts placement decisions). This might make the initial selection of a candidate more complicated and less systematic with more judgment and consideration being required of decision makers for each individual case. Second, given the width of the categorical bands and the correlation found between the current en route score (used by AIR® as the basis for en route placement) and the tower score (r = .65, see Table 1), it would be expected that if the placement rules suggested by AIR® were used, a good portion of candidates would receive the same categorical ranking for both options (i.e., Well-Qualified or Qualified in both options). Third, the en route equation was reweighted to find an optimal balance between validity and the reduction of adverse impact, but the tower equation reported by AIR® was not weighted in a similar way. This suggests that the en route equation used by AIR® in their analysis would produce a different option score for reasons other than "true" subtest relationships to performance.

Table 1.
Correlations between Current, En Route, and Terminal AT-SAT Scores and 1st Facility Success.

	Current	En Route	Terminal
Current			
En Route	.880		
Terminal	.651	.793	
1st Facility Success	.120	.210	.176

Note. All correlations significant at $p < .01$, $n = 2{,}332$. Current, En Route, and Terminal AT-SAT scores are based on similar, but slightly different equations developed through two AT-SAT validation studies.

Taking these drawbacks into account, we investigated an alternative approach to placement. The first step would be to categorize individuals using the current operational AT-SAT equation, which was weighted to mitigate adverse impact (Wise, Tsacoumis, Waugh, Putka, & Horn, 2001), into Well-Qualified, Qualified, and Not Qualified categories using the current cut scores as a basis for initial selection. Second, two additional composite scores would be computed based on (a) the original, unadjusted weights for en route (Ramos et al., 2001a, b), and (b) the tower equation developed by AIR® (2012) for terminal. For convenience, these will be referred to as the Current, En Route, and Terminal scores, respectively, throughout the rest of this paper. The applicant's hiring status would first be determined by using the Current score to determine the initial categorical rankings. Persons categorized as "Not Qualified" on the basis of their Current score would be removed from further consideration. Next, the En Route and Terminal scores would be computed for each person using the respective option-specific weights. Whichever score

was highest would serve as the placement recommendation. In the rare event of a tie, the applicant would be given a recommendation of "Either." The initial categorization (i.e., Qualified or Well-Qualified) based on the Current score would then be attached to this option recommendation.

Evaluation of Proposed Placement Approach

The following analyses were conducted to evaluate the proposed placement approach. First, logistic regression analyses were completed to verify the relationship of AT-SAT scores (computed using the three equations) to first facility training success, a criterion measure not used in the two previous concurrent, criterion-related validation studies. First facility training success refers to whether or not developmental controllers achieved certified professional controller (CPC) status at their first facility. Second, cross-tabulations were computed to examine the potential outcomes and utility of using AT-SAT for placement. Third, given that placement would constitute an employment decision encompassed by the Uniform Guidelines on Employee Selection Procedures, the potential for adverse impact was assessed against the 4/5ths rule (29 C.F.R. § 1607.4D) (EEOC, 1978).

The data used for these analyses were extracted from FAA AVIATOR, the Air Traffic National Training Database (NTD), the AT-SAT database, and the FAA Personnel and Payroll System (FPPS). Extracted information included AT-SAT test scores, race, gender, pay, and developmental training status. The sample used for the adverse impact analyses included anyone who had taken AT-SAT (N = 18,663) and who had race/gender information available (Race: N = 15,052; Gender: N = 14,115). The sample used for all other analyses included individuals who had AT-SAT data and a finalized first facility outcome (i.e., achieved CPC, failed, or transferred from first facility due to performance) by July 2012 (N = 2,332). In both samples, individuals had submitted an application for an ATCS position between 2007 and 2009.

Results

Logistic Regression

The results of the logistic regression analyses showed that the Current – $R^2 = .022$, χ^2 (1, 2332) = 32.71, $p \leq .001$, En Route – $R^2 = .064$, χ^2 (1, 2332) = 99.32, $p \leq .001$, and Terminal – $R^2 = .047$, χ^2 (1, 2332) = 71.88, $p \leq .001$ scores (based on the previously derived equations) were statistically significant predictors of first facility training success. The raw correlations between these scores and first facility training success, uncorrected for range restriction, were similar but not identical to each other and can be found in Table 1. These findings parallel the results obtained during both concurrent validation studies to assess the predictive of AT-SAT using ordinary least squares regression analyses and other types of job performance measures, as well as the results of a longitudinal validation of AT-SAT using first facility training success as the criterion (see Broach et al., 2013). Additionally, when logistic regression analyses were run separately for the En Route and Terminal samples (not restricting subtest weights based on the previously derived equations), the subtest scores were differentially correlated with first facility training success similar to the findings of AIR® that the subtest weights using a sample of tower controllers were not identical to those found in the original en route validation study. Taken together, this evidence demonstrates some degree of differential validity and prediction, both technical requirements as described by the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), as well as the Principles for the Validation and Use of Employee Selection Procedures (SIOP, 2003).

Cross-Tabulation Analysis

The next step in the analysis was to cross-tabulate actual vs. hypothetical placement. "Actual Placement" was the official assignment of newly hired controllers to en route or terminal facilities without regard to their Current score on AT-SAT. "Hypothetical Placement" was the decision that would have been made using the En Route and Terminal scores derived from AT-SAT. Placement was "correct" when the actual placement matched the hypothetical placement based on En Route and Terminal scores; placement was "incorrect" when the actual placement did not match the hypothetical placement.

The cross-tabulation was computed for those who were placed "correctly" or 'incorrectly" by their success in field training (Table 2), and the results were adjusted based on the typical proportion of hires assigned to each option (Table 3). The results of this analysis suggest that the utility of using AT-SAT for placement is marginal to slightly negative, depending on how the data are examined. The FAA could potentially see a 3% increase to 80% in the success rate of controllers "correctly" placed into the en route option as compared to the baseline success rate (without AT-SAT guided placement) of 77%. However, this gain could be offset by a 5% reduction to 74% in the

success rate of those "correctly" placed into the terminal option as compared to a baseline success rate (without AT-SAT guided placement) of 79% (see Table 3) for a net reduction in success rates across both options of 2%.

Table 2.
Cross-tabulation of Training Completion Rates at 1st Field Facility.

Actual Placement	Hypothetical Placement	Unsuccessful		Successful		
		N	%	N	%	Total
En Route**	En Route	111	20%	436	80%	**547**
En Route	Terminal	79	27%	218	73%	**297**
Terminal	En Route	153	17%	728	83%	**881**
Terminal**	Terminal	159	26%	448	74%	**607**
Total		**502**	**22%**	**1,830**	**78%**	**2,332**

Note. **Indicates "correct" placement – meaning that applicants were actually placed in the option that AT-SAT would have predicted had it been used for this purpose at the time of hire.

Table 3.
Training Success Rates at the 1st Field Facility with and without Placement.

	Success Rate without Placement	Success Rate with Placement
En Route (36% of positions)	77%	80%
Terminal (64% of positions)	79%	77%*
Across Options (weighted by number of positions)	78.28%	78.08%

Note. *Indicates rate adjusted for likelihood of filling 40% of terminal positions with applicants initially recommended for *En Route* placement

However, this loss must be reexamined and weighted within the context of the number of positions available in each option and the number of controllers being hypothetically placed in the en route option. The overall baseline success rate in terminal without placement is driven upwards by the higher success rate of individuals that would hypothetically have been placed in the en route option. Given the number of applicants that scored higher on the En Route equation, as compared with the number of positions typically available for en route controllers in recent years, it is estimated that approximately 40% of available terminal positions could be filled by individuals with en route recommendations. This would likely be the preferred policy given their apparent ability to succeed in either option.

Thus, to accurately estimate the overall success rate with AT-SAT guided placement for terminal, given the likely situation that 40% of the positions could be filled by applicants scoring higher on the En Route equation (who would likely have higher success rate – 83% vs. 74%), a weighted average was computed. The overall success rate, assuming placement of some applicants with En Route placement recommendations into the terminal option then becomes 77% [(83% success rate x 40% of the positions) + (74% success rate x 60% of the positions)] instead of 74% for terminal positions. This computation results in a success rate 2% lower than the current terminal success rate seen without using AT-SAT for placement.

In sum, if AT-SAT is used to guide placement by option, there is a potential increase in success rates for those placed in en route of 3% but a potential decrease for those placed in terminal of 2%, for an overall 1% increase in success rates. However, this estimate must also be considered within the context of the ratio of people hired into each option. Generally speaking, because more people are hired into the terminal option (accounting for approximately 64% of open positions yearly), the decrease in the terminal success rate must be weighed more heavily in the calculation of overall success rates computed with and without placement (Table 3). Taking the higher hiring rate in the terminal option, the net effect of using AT-SAT for placement would likely be a very slight reduction in the overall success rate across both options (Table 3).

Adverse Impact Analysis

As with other employment decisions, a placement decision carries with it the potential to impact an individual's ability to earn. Given the nature of this decision, the potential for adverse impact against members of protected groups must be considered. Using data from FPPS, it was determined that en route controllers earn on average approximately $20K more per year than terminal controllers. The difference in annual salaries was calculated using a snapshot of the FPPS data captured in July 2012. On average, receiving a recommendation for placement into the en route option would likely provide an individual with a greater opportunity to earn more over the course of employment and is, thus, considered the preferred option for calculating adverse impact.

Table 4.
Adverse Impact from Placement Decision.

	Hypothetical Placement			En Route Placement Rate	Adverse Impact Ratio[a]
	En Route	Terminal	Total		
By Ethnicity					
Asian	228	228	456	.50	.95
Black	713	2,324	3,037	.23	**.45**
Hawaiian-Pacific Island[b]	26	49	75	.35	**.66**
Hispanic-Latino	269	556	825	.33	**.62**
Native American-Alaskan Native	30	35	65	.46	.88
White	4,632	4,209	8,841	.52	
Multi-racial	462	569	1,031	.45	.86
No groups marked	357	358	715	.50	.95
Total	**6,717**	**8,328**	**15,045**		
By Sex					
Female	1,103	2,320	3,423	.32	**.66**
Male	5,350	5,686	11,036	.48	
Total	**6,453**	**8,006**	**14,459**		

Note. [a]Adverse impact ratio calculated with respect to whites for ethnicity and male for sex.
[b]Groups comprising less than 2% of the applicant pool are italicized. Bold ratios are less than what is acceptable under the 4/5^{ths} Rule (0.80).

Using the placement rules previously described, assigning controllers to an option using their AT-SAT scores could result in differential placement rates by race and sex into the terminal and en route options (Table 4). For example, just 23% of black candidates would be recommended for placement in en route, compared to 52% of white candidates (adverse impact ratio = .23/.52, or .45, where the threshold for adverse impact is defined as a ratio of .80 or less by the Uniform Guidelines on Employee Selection Procedures [EEOC, 1978]). The adverse impact ratio for Hispanic/Latino applicants was .62 and for females was .66. This adverse impact would result in addition to the adverse impact these protected groups already face in assignment to the Well-Qualified category ranking for initial selection considerations.

Summary

Looking at both of the concurrent validation studies and this current set of analyses together, there is sufficient evidence to suggest that the abilities required to perform air traffic control tasks do vary, to some limited degree, by option. The regression analyses (calculated repeatedly using different samples and at different times) have, in fact, derived different equations for the two options, which overlap but are not completely identical. This evidence can help provide the technical justification required, if the FAA were to pursue the use of AT-SAT for placement.

However, it is not clear that the variation by option is of a sufficient degree to justify differential placement given the minimal utility observed. Moreover, the utility of using AT-SAT to guide placement is minimal – and might be slightly counterproductive for the FAA. The cross-tabulations indicated that the success rate in en route would increase if AT-SAT is used for placement but would decrease in terminal. Taken across both options, field training success rates would not likely change in a meaningful way provided that the number of candidates typically hired for each option in recent years remains consistent. Finally, placement using AT-SAT could potentially have adverse impact on individuals in protected classes. That is, members of protected classes would be placed into higher paying en route facilities at less than 80% of the rate of majority members of each class (race and gender). Differential placement rates on the basis of AT-SAT scores could create troubling pay disparities by race and sex. If the FAA were to use AT-SAT for placement, the risk of additional adverse impact and pay disparities should be evaluated against the marginal utility of placement in terms of changes in field training success rates..

Acknowledgements

Research reported in this paper was conducted under the Air Traffic Program Directive /Level of Effort Agreement between the Federal Aviation Administration Headquarters and the Aerospace Human Factors Division of the Civil Aerospace Medical Institute sponsored by the Office of Aerospace Medicine and supported through the FAA NextGen Human Factors Division.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Psychological Association.

American Institutes for Research. (2012). *Validate AT-SAT as a placement tool.* (Draft report prepared under FAA contract DTFAWA-09-A-80027 Appendix C). Oklahoma City, OK: Federal Aviation Administration Aerospace Human Factors Research Division (AAM-500).

Borman, W. C., Hedge, J. W., Hanson, M. A., Bruskiewicz, K. T., Mogilka, H. J., Manning, C., Bunch, L. B., & Horgen, K. E. (2001). Development of criterion measures of air traffic controller performance. In Ramos, R. A., Heil, M. C., & Manning, C. A. (2001). *Documentation of validity for AT-SAT computerized test battery, Volume II*. (Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Broach, D., Byrne, C. L., Manning, C. A., Pierce, L. G., McCauley D., & Bleckley, M. K. (2013). *The validity of the Air Traffic Selection and Training (AT-SAT) test battery in operational use.* (Report No. DOT/FAA/AM-13/3). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Cannon, M. M. & Broach, D. (2011). *Studies of next generation air traffic control specialists: Why be an air traffic controller?* (Report No. DOT/FAA/AM-11/12). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290-39315.

Federal Aviation Administration. (2014). *A plan for the future: 10-Year strategy for the air traffic control workforce 2014-2023.* Retrieved from http://www.faa.gov/air_traffic/publications/controller_staffing/media/CWP_2014.pdf

Horgen, K., Lentz, E. M., Borman, W. C., Lowe, S. E., Starkey, P. A., & Crutchfield, J. M. (2012, April). *Applications of simulation technology for a highly skilled job.* Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Manning, C. A. (1998). Air traffic controller field training programs, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992.* (Report No. DOT/FAA/AM-98/23). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Nickels, B. J., Bobko, P., Blair, M. D., Sands, W. A., & Tartak, E. L. (1995). *Separation and control hiring assessment (SACHA) final job analysis report* (Deliverable Item 007A under FAA contract DFTA01-91-C-00032). Washington, DC: Federal Aviation Administration, Office of Personnel.

Ramos, R. A., Heil, M. C., & Manning, C. A. (Eds.). (2001a). *Documentation of validity for the AT-SAT computerized test battery, Volume I.* (Report No. DOT/FAA/AM-01/5). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Ramos, R.A., Heil, M.C., & Manning, C.A. (Eds.). (2001b). *Documentation of validity for the AT-SAT computerized test battery, Volume II.* (Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of employee selection procedures* (4th ed.). Bowling Green, OH.

U. S. Department of Transportation Office of the Inspector General. (2010). *Review of screening, placement, and initial training of newly hired air traffic controllers.* (Report. No. AV-2010-049). Retrieved from http://www.oig.dot.gov/audits?tid=71

U.S. Office of Personnel Management. (2013). General schedule qualification standards: Air traffic control series - 2152. *OPM Classification and Qualifications.* Retrieved from http://www.opm.gov/policy-data-oversight/classification-qualifications/general-schedule-qualification-standards/2100/air-traffic-control-series-2152/

Waugh, G. (2001). Predictor-criterion analyses. In Ramos, R. A., Heil, M. C., & Manning, C. A. (2001). *Documentation of validity for AT-SAT computerized test battery, Volume II*. (Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Wise, L. L., Tsacoumis, S. T., Waugh, G. W., Putka, D. J., & Hom, I. (2001, December). *Revision of the AT-SAT.* (Report No. DTR-01-58). Alexandria, VA: Human Resources Research Organization (HumRRO).

LEARNING OF LOCATION-IDENTITY BINDINGS: DEVELOPMENT OF LEVEL 1 SITUATION AWARENESS IN AN AIR TRAFFIC CONTROL-LIKE TASK

Alexander Nalbandian and Esa Rantanen
Rochester Institute of Technology, Rochester, NY

Knowing "what is where" is essential to human perception and performance. This knowledge corresponds to the concept of Situation Awareness (SA), specifically Level 1 SA. The underlying research paradigm concerns tracking of identical objects moving on a screen (Multiple Object Tracking, MOT). This method has been useful to investigate the fundamentals of visual tracking, but it lacks a connection to real-world scenarios. In another paradigm, objects tracked have unique identities (Multiple Identity Tracking, MIT) requiring a combination of peripheral and focal perception in tracking. This model has been used to examine air traffic controllers' SA. This paper will report results from an experiment where objects with identities similar to typical air traffic control (ATC) call signs were displayed on a plan-view display. The 4- and 8-object conditions replicated previous research and a 12-object condition simulated the density of typical ATC displays. The displays were static to examine the time required to create identity-location bindings (Level 1 SA). The displays were periodically blanked and participants queried about locations of target objects. Location errors and response times were recorded. The object array size of 4 had the highest accuracy and shortest time to acquire good SA. As the object set size increased from 4 objects to 12 objects, location errors increased dramatically. In sum, participants could reliably retain locations and identities of only about three objects. This finding has implications to the capacity of controllers to maintain adequate SA in future ATC systems.

This research is premised on two main ideas, one applied, and the other theoretical. The applied premise is the need to accurately model air traffic controllers' performance. The primary task of air traffic controllers is to ensure that aircraft within their areas of responsibility maintain at least the minimum standard vertical and horizontal separation from each other at all times. To perform this task, controllers must at all times know the current positions of aircraft they are responsible for as well as their trajectories to predict their future locations. Controllers refer to their ability to track multiple aircraft as a mental "picture". In the human factors research literature an analogous—if not identical—construct to the "picture" is situation awareness (SA). SA is typically defined to have three distinct levels (Endsley, 1995a). Level 1 refers to the perception of elements in an environment, such as aircraft call signs and their spatial locations on the controller's display, Level 2 pertains to the meaning of those elements, and Level 3 to controllers' ability to predict their future states.

With the introduction on the NextGen technology in ATC in the near future, controllers' tasks will shift from active control of air traffic to monitoring of automation (Durso & Manning, 2008; Metzger & Parasurman, 2001). At the same time, the number of aircraft under individual controllers' responsibility at any given time will likely increase. Under such conditions, acquisition and maintenance of Level 1 SA in not a trivial task. Niessen and Eyferth (2001) demonstrated a cognitive model (MoFi) of controllers' SA that posits a monitoring cycle of the radar screen to acquire and refresh all aircraft information, suggesting a continual cycle of learning and forgetting. Forgetting about aircraft is a common error in ATC (Shorrock, 2005), as are perceptual errors of misidentification of aircraft based on both auditory and visual input (Shorrock, 2007). Such errors can be attributed to an inadequate Level 1 SA.

Research Paradigms

The second, theoretical, premise of our work concerns the choice of an appropriate research paradigm for examining human performance in tasks that are relevant to and closely resemble those of air traffic controllers. The multiple (moving) object- or identity tracking (MOT/MIT) paradigm is plainly analogous with ATC tasks. In previous research on MOT it has been constantly argued that the mechanism behind tracking is parallel in nature. The size of the parallel capacity is most explicitly defined by Pylyshyn and Storm (1988), who posited 4-5 visual indexes that move automatically along with the moving objects, as if these pointers were glued to the tracked objects. Very recently, however, this view has been questioned. Oksama and Hyönä (2004) have provided evidence that the efficient tracking performance is based on continuous attention switching between the tracked targets with the help of visuospatial working memory. To study dynamic identity tracking or binding, i.e., the observer's

awareness of the location and identity of visual elements at any given time, Oksama and Hyönä (2004) developed a new multiple identity tracking paradigm, which is plainly analogous with ATC tasks. They found that multiple identity tracking performance deteriorated linearly as a function of set-size, tracking time, object speed, and target familiarity. Furthermore, they found that tasks measuring visuospatial memory and attention switching proved to be significant predictors of multiple object tracking performance. These findings are not consistent with the notion of parallel tracking, which is assumed to carry out the task efficiently and automatically without recourse to featural or semantic properties of the objects. On the other hand, they are more consistent with a higher-order post-attentive serial switching account, which operates with the help of temporary memory buffer(s).

Hope, Rantanen, and Oksama (2010) incorporated the concept of entropy, or the magnitude of direction changes in object trajectory, and traditional aircraft call signs to produce an experimental paradigm that could test how well a participant could track moving objects on a PVD. The experimental design was modified from a moving identity tracking (MIT) task, but instead of tracking a set of predetermined objects, all object were potential targets and ATC call signs were used in place of the line drawings or faces. The call signs were so small that the participants had to foveate on each identity. After a viewing period of 20 seconds the objects were stopped and masked and participants prompted to click on a particular call sign. The participant then chose the object in question from the stationary masked call signs. The participant would answer by the clicking on the call sign he thought was the correct one. Independent variables were the number of objects on the screen (4, 9, 14, 19 identities) and the amount of entropy (0.00 meaning a straight line, and 0.69 and 1.00 resulting in constantly changing direction).

A main effect for the number of objects was found. More importantly, a small effect for entropy was found but in opposite direction than hypothesized. Better performance was achieved in higher entropy conditions than when the objects moved along straight lines. Hope, Rantanen, and Oksama (2010) suggested that the level of SA was not at the level 3 that has been suggested by previous literature (Endsley & Garland, 2000) but at level 1 SA. In other words, it appears that the participants did not encode the trajectory information of the moving objects but looked for them at their last known positions. These results suggest that maintenance of SA in ATC is a static visual search task rather than a tracking task.

ATC Display Considerations

The above conclusion is further supported by examination of the speeds symbols of aircraft typically move on ATC displays. Although aircraft move through the airspace at high speeds and control decisions have to be made and actions taken within temporal “windows of opportunity” (Rantanen, 2009), at a perceptual level it may be argued that the movements of aircraft symbols on plan view displays (PVDs) represent a static rather than dynamic situation. To determine the speed of a moving object (e.g., aircraft symbol) on a PVD in terms of degrees of visual angle per second (Deg. VA/s), four parameters need to be known: (1) the ground speed of the object displayed, and in the case of airborne displays, the ground speed of the ownship, (2) the scale of the display, or the area depicted on the display, (3) the actual size of the display, and (4) the viewing distance.

In ATC applications all of the aforementioned parameters vary widely. Controllers can sit back or lean forward as they view their displays, their viewing distance ranging from as much as 1 m (or 1,000 mm) to as little as 250 mm. For our calculations here, the typical viewing distance of computer displays in experimental conditions of about 500 mm seems to be a reasonable average of also the viewing distances of ATC PVDs, and hence we will use this value in subsequent calculations.

The second source of variability is the scale of the PVD, which can be freely selected by controllers from a wide range of values. There are no hard limits for display range, but in practice these typically vary from about 200 nm across the display for large enroute sectors with little traffic to about 20 nm across the display in approach control operations. Finally, aircraft ground speeds vary from supersonic for military jets (since the retirement of the Concorde) to about automobile highway speeds for small aircraft. Hence, we can estimate the range of aircraft true air speeds (which equals their ground speed in the absence of wind) to be from 1,150 kts (Mach 2 near tropopause) to 80 kts.

Given the display range options and the range of aircraft speeds as described above, and assuming a modern 20 x 20 in PVD, we can calculate the range of target (or object, i.e., aircraft position symbol) velocities on ATC PVDs in terms of VA to be from 0.25 deg/s for a military jet maneuvering at about Mach 1.2 at low level and viewed on a PVD with a range of 50 nm to 0.02 deg/s for a general aviation on a cross-country flight at 100 kts, viewed on a display showing 100 nm across. In Table 1 some typical target velocity values on PVDs have been calculated. Given these speeds and the perceptual cycle of controllers monitoring traffic on their displays and continually updating each aircraft’s position information, we argue that the displayed information is more static than dynamic in nature. Futhermore, contrasting the velocities in Table 1 to those used in experimental MIT research (2.6

to 10.7 DegVA/s in Oksama & Hyönä, 2008, and 4.32 DegVA/s in Hope et al., 2010) casts doubt on the usefulness of the MIT paradigm to study air traffic controllers' performance in operational settings.

Table 1.
Some Typical Values of Object Speeds on PVDs in ATC. The Calculations Are Based on a 20-in Diameter PVD Viewed at a Distance of 500 mm.

Type of flight	Target GS (kts)	PVD range (nm)	Tgt speed (mm/s)	Tgt speed (DegVA/s)
Fighter jet, maneuver	790	50	2.23	0.26
Airliner, approach	230	50	0.65	0.07
Fighter jet, cruise	1150	200	0.81	0.09
Airliner, cruise	550	200	0.39	0.04
GA, cruise	100	80	0.18	0.02

Performance Measurement Considerations

Hope, Rantanen, and Oksama (2010) masked the object identities on the display when their participants were asked to click on a given identity. The participants could move the cursor on any masked identity which was unmasked, allowing for verification of the identity. If the first guess was wrong, the participants checked another masked identity. In this paradigm response accuracy could be expected to be 100% (i.e., the correct identity could eventually be found) and performance was measured by response time. According to this paradigm, if the participant has good Level 1 SA and knows where the queried object is, he or se will click on the right (masked) object the first time. If the participant's SA is poor, he or she must check multiple objects before finding the right one, which is reflected in an increasing response time (RT), which should be a multiple of the number of objects checked before finding the queried one.

The above paradigm is primarily concerned with identity error, or confusion between identities of displayed objects. It does not measure location error, as the objects remain visible (albeit with masked identities) throughout the experiment. Another paradigm may be considered to also measure location error, that is, uncertainty of the actual location of the queried object. To accomplish this, the objects on the screen could be masked completely, that is, the display is blanked upon query of a particular object. Now the participant must not only recall the approximate location of the queried object, but click on the exact location from memory. Identity errors would be manifested by clicks closer to the location of a wrong object than the queried one. Finally, instead of examining performance in a snapshot fashion, it may be of interest to do a timeline analysis of multiple trials. Such analyses would reveal learning of location-identity bindings over time.

Method

Participants

A total 45 participants were recruited from the student population at Rochester Institute of Technology. Most participants had normal or corrected to normal vision. Six participants were removed from the data because of difficulties completing the task.

Apparatus

A MacbookPro laptop computer was used to run the experiment. The computer had a 15" LCD screen to display the experimental objects. Screen resolution was 1024 × 768 pixels. A standard Microsoft desktop mouse was used to move the cursor. The PEBL programming language was used to create the experimental visualization and to collect the data.

Stimuli

The stimuli were objects that mimicked ATC call signs and consisted of 3 letters and 4 numbers. The call sign list was predetermined before set of trials. The objects subtended 2-3 degrees of visual angle and required the

participant to foveate on each object to resolve their unique identities. The locations of the objects on the display were random and configured before the experiment. To achieve an appropriate level of difficulty of the objects' identities, same letters or numbers were used at the first, last, and center positions in each of the call signs in various sets (Tydgat & Grainger, 2009). For the remainder of positions the order of the letters and numbers were shuffled to create unique identities for each object. Therefore, in a given set of objects, the same three letters were used and the same four numbers were used to make the identities sufficiently confusable and to make the participants to memorize more than just a single letter from each identity.

Independent Variables

The number of the objects varied between trials (4, 8, 12 identities). These set of objects were used to replicate previous research (Pylyshyn & Storm, 1988; Oksama & Hyönä, 2004, 2008; Alvarez and Franconeri, 2007) and to provide a semi-realistic simulation of the density of ATC displays. Each of the object sets was displayed stationary to examine the study time required to effectively create identity-location bindings.

Dependent Variables

The central dependent variable for this experiment was the error between the location of the object in question and the participant's response (a mouse-click on the display). The error was measured as the distance between where the participant clicked and the location of the queried object. Additionally response time, measured from the end of the study period to the mouse click on the queried object, was recorded.

Experimental Task

The task of the participant was to study the entire set of objects throughout the display for a period of time. Based on a pilot study, the study time was 0.7 seconds per object (or 0.7 s × number of objects). After that the screen went blank and the participant heard an audio clip of a particular object identity. Simultaneously, the mouse cursor reappeared at the center of the display and the participant then clicked on the location the queried object. This process was be repeated for 50 trials for each set of objects.

Procedure

The experimenter explained the task to the participant and answered any questions about the experiment to them. The participant then read through and filled out an informed consent form. Upon completion of the form, the participant was situated in front of the display. The participant was reminded to study all the objects on the display and to click on the location of a particular object upon hearing an object identity on an audio clip. Before the experiment, 5 practice trials were used to help the participant get used to the system. The participant was offered breaks after each set of trials. The order of the set of objects was randomized to minimize any order or fatigue effect. Once the trials were completed, the participant was asked about any strategies used and if they were satisfied by the amount of study time during the debriefing. Participants received extra credit in a psychology course for completing the study.

Results

There was a total of 150 trials for each participant, which adds to a total of 5,850 data records. In the initial analysis of the data, it was determined that some of the response times occurred before any time to distinguish the object identities were possible (below a threshold of 2,777 ms). A Pearson correlation was used to examine response time and accuracy (error measure) tradeoffs for each object set size. Only very weak correlations were found (r = 0.13, 0.08, and 0.11 for the 4, 8, and 12 objects conditios, repectively; all $p < 0.01$). If the participants took longer to respond, their accuracy improved only marginally.

The object with the shortest Euclidean distance away from the click location was considered the intended target. By matching the intended target of the user and the queried object, the identity error could be determined (i.e. match or no match). If the intended target matched the queried object, the distance was the location error.

Plots of accuracy versus trial number suggested learning curves . The power law equation ($y = ax^b + c$), was fitted to the data (Newell & Rosenbloom, 1981). The results are depicted in Figure 1. It is apparent that little learning took place after about 10 trials, or that after studying the objects for about half a minute (4 objects), or one

minute (8 objects), or minute and a half (12 objects), memory of their locations and identities improved only incrementally. Most importantly, the location errors for 8 and 12 objects remained quite large even after all 50 trials. This finding appears to support our hypothesis that performance in tracking multiple moving identities suffers primarily from poor Level 1 SA.

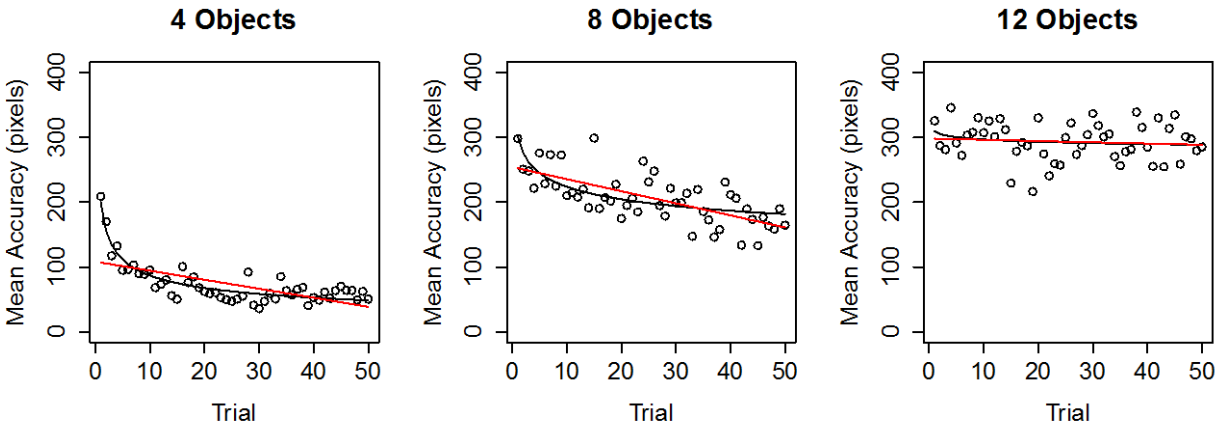


Figure 1. Learning curves of identity-location bindings in an experimental task involving 4, 8, or 12 objects with unique identities on a display. Performance even after 4.6 or 7 minutes of study time for 8 and 12 objects, respectively, was remarkably poor. N = 39.

Learning curves that emerge from average performance of multiple individuals are famously misleading and often mask individual differences and true effects of learning. Therefore, we examined performance of individual participants in each object set size condition. This examination revealed that the participants' performance in the 4-objects condition was generally good and reasonably high accuracy was achieved too quickly for learning effects to emerge. In the 8- and 12-object conditions, however, the performance was much worse and also much too variable to suggest any learning. Because the power law equation fitted very poorly to individual participants' data, no meaningful further analyses (functional data analysis) could be performed.

A holistic view of the results is provided in Table 2. The generally very poor performance is noteworthy. Even in the easiest condition with only 4 objects to memorize over 10% of the objects were misidentified. The means and ranges of the location errors should be judged relative to the display size of 1024 × 768 pixels. Using the percentage of correctly identified objects and the number of objects, it was calculated that only about 3 objects' could be retained in memory during the experiment.

Table 2.
Mean Location Error, Location Error Range, and Identity Error Percentage for each Condition.

Objects	M Location Error (pixels)	Error Range (pixels)	Ident. Error (%)	Approx No. Recalled Objects
4	38.17	1-188	88.24	3.53
8	40.68	2-194	41.11	3.29
12	50.27	3-151	23.37	2.80

Note. The mean location error is only for the correctly identified targets, not for all data points.

Discussion

Air traffic controllers tasks are quite complex and modeling controller performance therefore at least equally complex. Even reduction of controllers' tasks to experimental paradigms such as multiple identity tracking (Hope, Rantanen, and Oksama, 2010) or learning of initial location-identity bindings of several objects on a display (this study) have proved to involve myriad poorly understood variables. Our results also seem to suggest that many experiments on multiple identity tracking may have provided insufficient time for the participants to acquire

adequate Level 1 SA for good performance on tracking moving objects, confounding Level 1 SA and tracking performance.

The debriefing questions provided insight into strategies and the study time for the experiment. Participants described using spatial patterns of the objects, chunking the objects identities (Cowan, 2001), and splitting the screen in half to determine which object was where (Alvarez & Cavanagh, 2004). As for the study time, most participants felt they had sufficient study time for the 4 object condition, but inadequate time for the larger set sizes.

The average inspection time of 0.7 seconds per object was not sufficient for higher levels of object set sizes, apparent in the very poor accuracy with 8 and 12 objects. Either learning requires much longer time than this experiment provided, or, more likely, controllers are not committing all the information (object identity and location) to memory, but continually search for the aircraft they need to attend (Hope et al., 2010).

As is usually the case in research of ATC and controller performance using naïve college students as participants, this study, too, suffers from some serious limitations. The "callsigns" or identities of the experimental objects used in our experiment had no meaning whatsoever to the participants. In ATC, specific callsigns have meaning to the controllers, signifying type of the flight (corporate or private aircraft vs. airliners) and possibly much flight plan-related information, such as route and altitude. This information certainly is helpful in maintenance of Level 1 SA. Confusability of realistic callsigns is also not uniform and estimation of appropriate—or realistic—levels of confusability is therefor difficult. Finally, because of the number of variables influencing performance even in simple laboratory experiments, a single study does not allow drawing of many conclusions. For that, series of experiments are necessary. Nevertheless, methodological issues are also important, and we hope that this study makes a contribution in that regard.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106–11.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7*, 1-10.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

Durso, F. T., & Manning, C. A. (2008). Air traffic control. In C. M. Carswell (ed.) *Reviews of Human Factors and Ergonomics, 4* (pp. 195-244). HFES.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Endsley, M. R., & Garland, D. J. (2000). *Situation awareness: Analysis and measurement*. CRC Press.

Hope, R. M., Rantanen, E. M., & Oksama, L. (2010). Multiple identity tracking and entropy in an ATC-like task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(13), 1012-1016.

Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. *Human Factors, 43*(4), 519-528.

Mogford, R. H. (1997). Mental models and situation awareness in air traffic control. *The International Journal of Aviation Psychology, 7*(4), 331-341.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Lawrence Erlbaum.

Niessen, C., & Eyferth, K. (2001). A model of the air traffic controller's picture. *Safety Science, 37*(2), 187-202.

Oksama, L, & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher order cognition? An individual difference approach. *Visual Cognition, 11*(5), 631-671.

Oksama, L., & Hyönä, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology, 56*(4), 237-283.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179-197.

Rantanen, E. M. (2009). Measures of temporal awareness in air traffic control. *Proceedings of the 53rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 6–10). Santa Monica, CA: HFES

Tydgat, I., & Grainger, J. (2009). Serial position effects in the identification of letters, digits, and symbols. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 480-98.

AN OPERATIONAL ANALYSIS OF HUMAN FACTORS IN AN UNMANNED AIR SYSTEM

Saskia D. Revell
Royal Air Force Centre of Aviation Medicine
RAF Henlow, United Kingdom
Victoria J. Cutler
Royal Air Force Centre of Aviation Medicine
RAF Henlow, United Kingdom

Previous research has highlighted Human Factors (HF) issues associated with operating Unmanned Air Systems (UAS). This research has examined the human-machine interface, error types found in UAS mishaps, and examined specific factors such as workload or situation awareness. Fewer studies have examined the HF issues experienced during live military UAS operations in a conflict zone. Accordingly, a HF analysis was undertaken of a UK UAS unit operating in Afghanistan. The analysis was conducted using the Operational Events Analysis (OEA) approach, which is a structured, qualitative method of identifying flight safety HF issues. The OEA included UAS operators and maintenance personnel. HF issues were identified that included the aviation culture, characteristics of the task, fatigue and shift management, and the work environment.

Accident and incident rates for UAS are generally higher than for manned aircraft. Analysis of accident and incident data has shown that Human Factors (HF) are involved in up to 68% of UAS incidents (Williams, 2004; Tvaryanas, Thompson and Constable, 2006). Accordingly, researchers have sought to understand and address the HF issues associated with UAS operations.

HF studies of UAS operations have examined the human-machine interface, error types found in UAS mishaps, and explored specific factors such as workload or situation awareness. For instance, Thompson, Lopez, Hickey, DaLuz, Caldwell and Tvaryanas (2006) examined the effects of shift work and sustained operations on UAS operators and found subjective boredom amongst operators and decrements in vigilance over the course of a shift. UAS aircrew have also been found to report decreased mood and increased fatigue levels relative to traditional aircrew (Tvaryanas and Thompson, 2006). A number of studies have examined the nature of personnel who are recruited to fly and maintain UAS. McCarley and Wickens (2004) highlighted that there was a lack of consistency in the standards for UAS pilot selection across the US military for the extent to which the pilots had previous aviation experience. The impact of personnel background on culture was identified by Hobbs and Herwitz (2005), identifying a potential negative impact for maintenance personnel who had not come from a mainstream aviation background.

The wide range of factors considered reflects the novel field of UAS HF research, and the varied ways in which UAS are operated. For instance, Thompson et al.'s (2006) study analysed personnel involved in UASF MQ-1 Predator missions in support to operations where personnel were based at Nellis Air Force Base rather than the conflict zone itself. In the UK, the Hermes 450 (H450) UAS has provided operational Intelligence, Surveillance, Target Acquisition and Reconnaissance (ISTAR) capability while being controlled by operators situated within the conflict zone.

In 2011, the Royal Air Force Centre of Aviation Medicine (RAF CAM) identified a number of HF issues associated with operating the H450 whilst deployed at Camp Bastion, Afghanistan. These issues included the aviation culture, training, supervision, procedures and working environment in which H450 was operated. Five months after this work was conducted, a H450 accident occurred and the subsequent inquiry found that many of the issues identified in the RAF CAM analysis had contributed to the accident (MAA, 2012). Since the accident inquiry, changes were implemented by the unit and a follow-up HF review was requested in 2013 to provide a 'health check' for H450 operations. The aim of the follow-up HF review was to identify what HF issues were influencing the work of the team so that awareness of those issues could be raised and action taken to address.

Method

A HF review was undertaken of the British Army H450 unit operating at Camp Bastion, Afghanistan in summer 2013 using the Operational Events Analysis (OEA) approach. The OEA is a structured and proactive method of identifying HF issues that have the potential to influence flight safety (Harris, 2011). The OEA is based on the Accident Route Matrix (ARM) (Harris, 2011) which is a framework used as part of UK military HF air accident investigations.

The OEA is a mixed methods qualitative design which uses a combination of semi-structured interviews and observations. The OEA conducted for H450 in 2013 followed the first six stages outlined by Revell, Harris and Cutler (2014). The first three of these stages were associated with setting up the OEA and stage four was the OEA visit; stages five (analysis) and six (output) are described in the results section.

OEA Set Up

The requirement for the OEA was specified as being to conduct a HF 'health check' of H450 operations. The scope of the OEA was agreed to include H450 operations in Afghanistan and use of the Hermes 450 simulator in the UK. The HF specialist established a point of contact in both the UK and Afghanistan and familiarised themselves with the previous OEA that had been conducted and the subsequent accident investigation (MAA, 2012). A number of logistical considerations also were made to enable the HF specialist to integrate with the H450 team during the visit. H450 personnel were provided with key information in advance of the visit to allow personnel enough time to understand the OEA and decide whether to participate.

OEA Visit

The OEA visit took place over a six day period during summer 2013. At the time of the OEA visit the majority of personnel had been deployed for approximately four months and the Rest and Recuperation (R&R) cycle was in progress, as such, some personnel had recently returned from R&R while others, were due to go. H450 crew, however, had only recently deployed and the majority had not been in Afghanistan longer than a month.

Following a familiarisation tour of the unit, the HF specialist commenced an iterative process of interviews and observations to gather information regarding any HF issues that were present on the unit.

Interviews. 24 one-to-one interviews and multiple informal discussions were held with personnel. The interviews included nine H450 crew, seven maintainers, three management personnel and five people who were in operational support roles. Interviews were conducted across the spectrum of job roles, and across the rank structure. Participants were recruited voluntarily and on an anonymous basis. A semi-structured interview was used for the one-to-one interviews. These interviews took up to one hour and used the same interview form as used in HF accident investigations (Harris, 2011). Accordingly, the interview included questions on a wide range of HF issues including organisational factors, supervision, tasks, equipment, environment, behaviours and actions, and operator conditions.

Observations. Observations were undertaken of: flight planning, briefing, and debriefing; personnel operating H450 from the Ground Control Station (GCS) and from next to the runway; engineering tasks, including aircraft launch and recoveries; team meetings; and the H450 simulator used in the UK.

Results

Thematic analysis was used to review the data collected based on the six phases detailed by Braun and Clark (2006) applied as described by Revell et al. (2014). The output of the OEA was provided in a formal written report to the unit, which described the HF issues identified and the role of these issues in potential hazard sequences. Positive factors were also highlighted and a set of recommendations made to address the HF issues identified. In this section, some of the key findings from OEA are described. These findings are highlighted due to their applicability to a range of UAS operations, rather than being specific to characteristics of the H450 operation in Afghanistan.

Aviation Culture

Personnel felt that aviation culture and practices had improved since the introduction of H450, however, a particular challenge to maintaining that culture related to the fact that live flying of H450 in UK airspace was not permitted. Therefore, some personnel felt there was a lack of aviation focus, particularly when not deployed. H450 crew were drawn from a non-aviation, Army background and reported that when not deployed they returned to more typical 'Army' tasks and performed few aviation related duties. Operators then had to re-build their aviation experience and be re-immersed in aviation culture in preparation for their next deployment.

Task Characteristics

H450 sorties lasted up to fourteen hours and each flight was operated by two separate crews, with one crew responsible for the aircraft launch and first half of the mission and the second continuing the mission before recovering the aircraft. Therefore, crews were operating H450 for an extended period of time. The tasks to operate the aircraft during this time were perceived as being simple and non-demanding, linked to the high level of automation offered by the system. The tasks personnel performed were largely supervisory in nature, monitoring systems and responding to issues and changes that occurred. As a result, personnel perceived that operating H450 was mundane with little mental stimulation, and difficulties were reported in maintaining vigilance and alertness. Indeed, when operating the H450 simulator, crews were observed to appear bored and have difficulty maintaining attention to the displays.

Maintenance personnel also stated that they found working on H450 simple and monotonous, and that H450 was less technically challenging than other aircraft they had worked on (typically helicopters, as maintenance personnel were drawn from the Army aircraft engineering cadre).

Working Environment

The majority of H450 crew stated that the GCS was a cold environment to work in. The GCS had a large number of computer systems that required constant air conditioning to maintain performance. As a result, crews were instructed not to touch the temperature controls and mitigated the cold temperature by wearing extra layers including hats, scarves and gloves during winter periods. Operating the H450 involved sitting for long periods of time with little movement, which decreased the perceived temperature even further.

Maintenance tasks were often performed outside, and so could be undertaken in a wide temperature range. A particular issue was heat in the summer where personnel could be working without shade for an extended period of time, although personnel also reported that temperatures could get very low during winter periods. Extended periods of time working outside was particularly common for aircraft launch and recovery tasks where personnel would be on the runway either waiting for ATC and the GCS to release the aircraft or waiting for the H450 to land.

Accommodation

H450 operators and maintainers were accommodated in tents at Camp Bastion which, at the time of the OEA, was a major military base located within a conflict zone. Therefore, there could be a high level of noise from military equipment and personnel, including other aircraft. Personnel reported having disturbed sleep and rest due to noise. Sleep disruption was also linked with unreliable air conditioning inside the tents. Although a few reported that they became accustomed to the noise and heat, reduced sleep quality increases the risk of both acute and cumulative fatigue. Indeed, a number of personnel reported observing others being tired including those operating and working on aircraft.

H450 Crew Shift Pattern

In order to launch the UAS within the system design constraints and support operations 24 hours a day, five H450 were launched per day with staggered take-off times and two crews to cover each flight. As a result, there were ten different start times for H450 crew. Crews operated on each start time for two days before moving to the next start time through the cycle. Many of the start times changed in small (one hour) increments but that at some points crews had a greater change in start time, up to a maximum of seven hours. H450 crew generally perceived their shift pattern to be okay, however, some found it difficult to adjust their body clock and acclimatise due to the constantly changing shifts. During the twenty-day shift cycle, crews' body clocks were likely to be continuously lagging behind, as there was not sufficient time given to enable their body clock to catch up. The constant changing of body clock could impact on alertness and fatigue. As a result, some crew reported feeling tired, particularly at the end of their shift.

Fatigue Management of Maintenance Personnel

Maintenance personnel were not subject to the same duty time regulations as aircrew and worked a "24 hours on, 24 hours off" shift pattern. Maintenance personnel were generally happy with this shift pattern as personnel were given a six hour rest period during the shift, had facilities to sleep at the work location, and after the shift had 24 hours off to recuperate. However, a number of potential issues were identified with this shift pattern that would increase the risk of a fatigue related error. These included extended time on task, as well as limited and disrupted rest periods.

Morale

When personnel had recently arrived in Afghanistan, morale was reported to be high. However, morale was found to decrease across the deployment. This was particularly noticeable when comparing the recently deployed H450 crew, to personnel who had been deployed for four months. There were a number of reasons given for the diminishing morale such as the task characteristics, being away from home for a number of months, and the pressures of working in the same team for an extended period of time.

Discussion

The OEA method was applied to identify the HF issues associated with the operation of H450 by the British Army in Afghanistan in summer 2013. The OEA identified a range of issues including aviation culture, task characteristics, fatigue and shift management, and aspects of the operational environment. Many of these issues reflect those identified in previous HF research undertaken into different types of UAS operations, but the H450 work has also identified additional factors which reflected the operational requirements and environment in which the H450 was flown.

Similar to Hobbs and Herwitz's (2005) findings, the aviation culture amongst H450 operators was influenced by the background of the personnel. However, the culture was found to have changed

over time and was also influenced by other factors such as the nature of the task and the ability to operate H450 when not on operations.

Inline with Thompson et al.'s (2006) findings, the H450 operator's task in the GCS was perceived to be boring, and there were reports and observations that attention was difficult to maintain. The present study also found that similar issues occurred amongst UAS maintenance personnel. The perceptions of the task were also contributing to low morale, along with a number of other factors that may be specific to personnel deployed away from home, who are working and living together within close proximity.

With regard to fatigue, although a comparison was not made between UAS aircrew and traditional aircrew, the findings of the present study indicate that the nature of H450 operations meant some H450 crew were struggling to adjust their body clock to the shift pattern resulting in fatigue, particularly at the end of their shift. Further, maintenance personnel were not subject to the same working hour regulations as aircrew, which increased the scope for shift patterns to increase the risk of fatigue related errors.

HF issues that were raised in the present study that had not been identified in previous UAS research were the temperature in the working environment, be it too hot or too cold, and issues with gaining adequate rest when flying from a major base in a conflict zone. Few studies had analysed UAS in a conflict zone and the application of the OEA approach in this environment has enabled different insights to be provided into the HF challenges that military UAS operators face.

Looking towards the future, UAS will continue to be utilised in UK military operations to provide ISTAR capability, however, the nature of military operations is set to change since the draw down of personnel from Afghanistan and a new UAS has been introduced. UAS operators will need to be adaptable to the new UAS and changing operational environment, which is likely to bring in new HF challenges. Therefore, the HF issues associated with operating military UAS will need to be continually assessed to ensure optimal performance is maintained. Further, it is recommended that research and design of UAS continues to consider factors such as fatigue and shift management of UAS personnel, aviation culture, environmental temperature, mental underload and perceived boredom of both UAS aircrew and maintainers, and develop methods to address these issues.

Conclusion

This paper presented the results from an OEA which was conducted on a UAS unit operating at Camp Bastion, Afghanistan. The HF challenges faced by military personnel operating UAS included the aviation culture, tasks characteristics, fatigue and shift management, and the characteristics of the working environment. The OEA has enabled the British Army to mitigate the HF identified, for example, reducing tour lengths from six to four months and making improvements to the shift cycle. In order to gain optimal performance from military UAS operators, the HF challenges of operating UAS in conflict zones will need to be continually assessed and actions taken to address in the changing military environment.

References

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101.

Harris, S. (2011). *Human Factors Investigation Methodology*. Paper presented at the 16th International Symposium of Aviation Psychology, Dayton, OH.

Hobbs, A. & Herwitz, S.R. (2005). *Human Factors in the Maintenance of Unmanned Aircraft*. Moffett Field, CA: NASA Ames Research Center, San Jose State University Foundation.

McCarley, J.S. & Wickens, C.D.(2004). *Human factors concerns in UAV flight*. Urbana-Champaign, IL: Institute of Aviation, University of Illinois.

Military Aviation Authority (2012). *Service Inquiry Investigating the Accident Involving Unmanned Air System (UAS) Hermes 450, ZK515 on 2 October 201*1. Retrieved 23 January 2015 from https://www.gov.uk/government/publications/service-inquiry-investigating-the-accident-involving-unmanned-air-system-uas-hermes-450-zk515-on-02-oct-11.

Thompson, A.T., Lopez, N., Hickey., DaLuz, C., Caldwell. J.L. & Tvaryanas, A.P. (2006). *Effects of Shift Work and Sustained Operations: Operator Performance in remotely Piloted Aircraft (OP-REPAIR)*. Brooks City-Base, TX: 311th Performance Enhancement Directorate, Performance Enhancement Research Division. Retrieved January 22, 2015 from http://www.dtic.mil/dtic/tr/fulltext/u2 /a443145.pdf

Tvaryanas A.P. & Thompson WT. (2006). Fatigue in military aviation shift workers: survey results for selected occupational groups. *Aviation, Space and Environmental Medicine, 7*7, 1166-1170.

Tvaryanas, A.P., Thompson, W.T. & Constable, S.H. (2006). Human Factors in Remotely Piloted Aircraft Operations: HFACS Analysis of 221 Mishaps Over 10 Years. *Aviation, Space and Environmental Medicine, 77,* 724-732.

Revell, S.D., Harris, S.L. & Cutler, V.J. (2014). *A Preventative Approach to Identifying and Addressing Flight Safety Human Factors Issues*. Paper presented at the 31st European Association of Aviation Psychology Conference.

Williams, K.W. (2004). *A Summary of Unmanned Aircraft Accident/Incident Data: Human Factors Implications.* Oklahoma City, OK: Civil Aerospace Medical Institute, Federal Aviation Administration. Retrieved January 23, 2015, from http: www.hf.faa.gov/508/docs/uavFY04Mishaprpt.pdf

Acknowledgements

This paper presents a summary of key findings from the OEA that was conducted on Hermes 450 in Camp Bastion, 2013. As such, it does not present the full findings of the HF analysis. The formal OEA report has been released internally to the organisation.

EXPERIMENTAL EVALUATION OF VARYING FEEDBACK OF A COGNITIVE AGENT SYSTEM FOR UAV MISSION MANAGEMENT

Elisabeth Denk, Sebastian Clauss, Annike Borchers, Josef Werner & Axel Schulte
Universität der Bundeswehr München (UBM)
Institute of Flight Systems (LRT- 13), 85577 Neubiberg, Germany
{elisabeth.denk, sebastian.clauss, axel.schulte}@unibw.de

In this study we investigate on a cognitive delegation agent for UAV task-based mission management. Particularly, we advocate a specific high-level feedback provided by the agent to the human operator to enhance mission effectiveness. As extension to human supervisory control we suggest to introduce the concept of agent supervisory control where the agent is delegated by high-level operator commands and controls several sub-systems aboard the UAV fulfilling the mission in highly automated fashion. Results of our experimental human-in-the-loop study focusing on the effects of high-level feedback are presented. Therefore, two configurations are compared, one with basic feedback and one with full feedback. The results show that particularly during in-flight re-planning situations the full feedback is beneficial. Interaction times and task related activities are significantly lower. In the pre-flight mission preparation phase no significant effect were found. Results can be used to further develop highly automated (multi)-UAV mission management systems.

Agent Supervisory Control

Nowadays, military *Remotely Piloted Aircraft System* mission management is in focus of research (Clauss & Schulte, 2014; Theissing & Schulte, 2015). In modern UAV-systems, conventional automation (i.e., auto-flight systems) relieves the operator of high-frequent sensor-motor tasks and improves precision and performance for mission execution. Instead of manual control, the operator controls the aircraft intermittently through automation. The automation has to be monitored more or less continually by a *Human Supervisor* (HS). This type of control relationship was described as *Human Supervisory Control* (HSC) by Sheridan (1992).

The HS generally performs five *Supervisory Functions*. Determining the current objective and exploring a strategy to achieve it, using the given means (*plan).* The HS conveys its commands to the automation (*teach*) and monitors the automation to ensure proper execution (*monitor*). If necessary the HS intervenes (*intervene*) and finally may learn from experience to perform better next time (*learn*) (Sheridan, 1992). The cognitive capabilities of the HS, allow the overall system to react to individual challenges in the environment and the status of the UAV-system and enable it to compensate for unforeseen events.

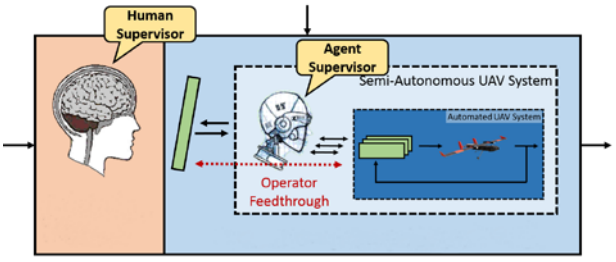


Figure 1. Work system of a semi-autonomous UAV with cognitive agent.

In this context, feedback information must be perceived, interpreted and processed by cognitive functions. To extend the operators support, a rather conventional approach of introducing more complex automation could constrain his work; increase the complexity of the overall system as well as the number of automation functions.

In this context, Bainbridge (1983) describes two *Ironies of Automation*, the first is a shift of human errors from manual control to designing and implementing of automation functions and the second is *Clumsy Automation* (Wiener, 1988). It supports the operator in low-stress situations, but cannot provide support in highly intense situations. For the supervision of automation functions in manned flight, Billings (1997) describes four *Costs of Automation*: *complexity*, *brittleness*, *opacity* and *literalism*.

To tackle some of these issues, cognitive capabilities are displaced such as decision-making, problem solving and planning aboard the aircraft. A cognitive agent, implementing cognitive capabilities, is introduced onboard the UAV to manage and control its existing conventional automation systems. In terms of *Cognitive Automation* (Onken & Schulte, 2010), the agent works within the supervision of the human operator and thus serves as a link between the mission management layer in the responsibility of the human pilot and the mostly automated UAV navigation, guidance and control.

Figure 1 shows the resulting work system from integrating a cognitive agent into the automated UAV-system. The human operator interacts with the single cognitive agent, rather than the multitude of automation functions. The agent supervisor formulates discrete commands for the conventional automation and monitoring their execution. The cognitive capabilities of the agent allow deriving action plans from human delegated objectives. For this purpose, the agent plans and coordinates the application of the underlying automation. Still, the semi-autonomous agent does not have authority to modify or specify its own objectives (Onken & Schulte, 2010). In analogy to the definition of HSC, the relationship between the cognitive agent, the conventional automation and the UAV-system may be best described by the term *Agent Supervisory Control* (Clauss, Kriegel, & Schulte, 2013). But the cognitive agent is always acting like an *intelligent* subordinate to the human within the concept of HSC, using its cognitive capabilities to execute human tasks in a flexible manner.

Figure 2 shows the resulting management hierarchy of the UAV-system including, the additional echelon as the guidance layer of the cognitive agent. The agent is introduced between the operator and the conventional automation. In this role the cognitive agent combines commanding and monitoring as well subordinating to the human operator. The agent supervisor behavior and the interaction with the human will have a combined effect on the operator's perception.

Agent Feedback within a Task-Based Guidance Approach

The human pilot acting as a supervisor of the semi-autonomous UAV-system requires information, which allows monitoring its performance and to intervene (re-plan) when necessary. Additionally, the human behavior and criteria for delegating the cognitive agent are depending on the operator's information about the agent's capabilities and performance. So the operator has to decide, which tasks must be delegated and which could be done manually (Leana, 1986; Parasuraman & Riley, 1997). For supervisory control of conventional automated systems, sophisticated concepts of interactions already exist. In the following we examine an approach to a bidirectional information flow. This allows a calibrated delegation of tasks to the cognitive agent as well as it provides adequate feedback to the operator.

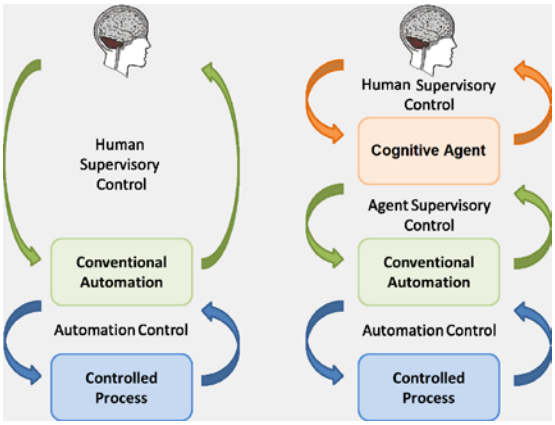


Figure 2. Agent Supervisory Control (ASC) as additional guidance loop between operator and conventional automation.

As a concept of delegation we chose a Task-Based UAV Guidance (TBG) approach (Uhrmann & Schulte, 2011). Herein the operator solely defines the objectives for the agent as commanded intents, instead of formulating step-by-step instructions for multiple automation components. Therefore, the operator defines what the semi-autonomous system shall accomplish, instead of providing *how* (i.e. through what actions) this shall be achieved. Tasks might be military reconnaissance missions. TBG stems from an inter-human delegation relationship and relieves the supervisor from the tedious task to derive automation action instructions from intentions (Clauss & Schulte, 2014).

With respect to the operator's tasks, the performance of the system is mainly affected by its ability to decide which tasks to delegate to the agent and how to formulate these tasks. Parasuraman and Riley (1997) presented criteria for the task delegation to subordinate automation (cognitive agent) and indicates those criteria that can be directly influenced by automation design. The central criterion is *reliance*, resembling the affinity of the operator to delegate a task. The human reliance on automation is directly influenced by the *confidence* in a manual executing of the task, the current level of *fatigue*, the *perceived risk* associated with task failure and *trust in automation* for a satisfying task executing.

Three specific criteria in this context (*machine accuracy, trust in automation and workload*) can be identified as affected directly by automation behavior and accordingly at least to some extent controllable by design. Machine accuracy describes the level of sufficiency with which a delegated task is executed by the automation. Trust in automation comes from the operator's perception of the of automation behavior and is used to predict behavior in future situations. The operator's sensor-motor and cognitive workload is directly influenced by the

interaction with automation during task execution. The likelihood of trust in task delegation depends on the complexity of the underlying system (Lee & See, 2004). The basis for trust development, regarding the agent's capabilities, is the information feedback received by the operator. This information can be categorized (Lee & Moray, 1992). The operator's monitoring task is accomplished with the help of the assessment of agent feedback. The agent itself monitors the heterogeneous conventional automation systems and creates a symbolic representation, which it communicates to the human. With regard to the operator's workload, the human is supported by the agent, if the cognitive task (interpreting the symbolic information) is less complex than the processes needed to monitor the conventional automation itself. The desired form of feedback, with respect to its application domain, can be described by the term *etiquette* (Miller, 2002). Etiquette means an established form of interaction that expresses the role of the transmitter and which rules aim a better understanding and enhancing the effectiveness and the safety of the communicating system.

An extended feedback of a cognitive agent was developed (see figure 3 right side, *Enhanced Feedback Configuration*) to minimize mission errors and to raise human awareness about UAV resources and capabilities. The cognitive agent provides event-independent information about the current system status of the UAV, the current tactical situation information and perception results (threats, tactical elements) and the status and the current objective of the agent, its (reviewed) task-agenda and its execution progress. Further, the agent provides information only on currently available automated capabilities. As a feedback to the task-agenda delegated by the operator, the agent presents its execution plan containing a list of activities to be performed in order to transition from its current state into the specified goal state (Agent Plan). In case of plan execution errors, the agent uses its knowledge-based reasoning to independently derive an alternate action plan. If no solution can be derived within the boundary conditions specified by the operator, the agent reports an error reason to the operator. Figure 3 shows the map displaying a mission for each configuration. Figure 3, right (i.e., *Enhanced Feedback Configuration*) shows the task agenda (reviewed by the cognitive agent). The blue line indicates the flight plan created by the agent.

Figure 3, left (i.e. *Baseline Configuration*) shows the setup where the agent's feedback is limited to only graphical information about the position of the UAV and its current flight plan. In this configuration the agent is fully functional, but no information about its intent or task execution plan is conveyed to the operator. Tasks may be delegated by the operator with no visibility of the UAV's actually available capabilities.

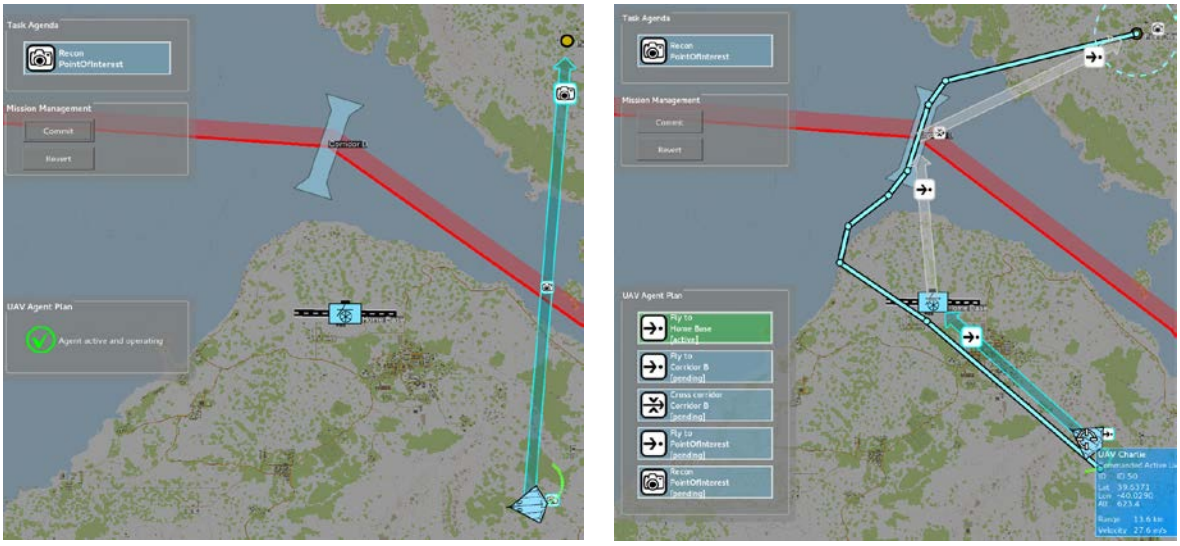


Figure 3. Moving map and planning display in mission and payload control station (left: *Baseline Configuration*, right: *Enhanced Feedback Configuration*).

Experimental evaluation

Communication between the operator and the automation on a symbolic level is an essential part of task based guidance. We hypothesize that the feedback provided by the agent will affect the work result of the system, even if its decision-making and control functions stay unmodified. In an experimental campaign the impact of the agent's feedback behavior on planning, commanding and re-planning of the operator is examined. The developed cognitive agent (Clauss & Schulte, 2014) with advanced feedback abilities is evaluated with respect to human

performance, re-plan capability and human trust in the agent. In our experiment we compare the full feedback system to a configuration with reduced feedback information.

Research setup and configurations

The experimental design is a within-subject design with a secondary task (Borchers, 2014; Werner, 2014). Its factor is the feedback behavior of the cognitive agent during mission execution (*configuration A* and *B*). For its evaluation, a comparative experiment was conducted, in which two missions (m*ission I* and *II*) are performed. Both missions are very similar, so they are comparable (see experimental procedure). The participants perform *mission I* and *II* while agent configuration A (*Baseline Configuration*) and separately with agent configuration B (*Enhanced Feedback Configuration*) (see figure 3). All participants were exposed to the two configurations of the system, while performing missions and completing the questionnaires. To eliminate sequence effects and spillover effects, the missions were randomized.

The experimental hypothesis says that configuration B reduces workload, planning effort and error rate, compared to configuration A. It can also be assumed, that configuration B leads to higher situation awareness, distribution of attention, trust in automation, visual perception and acceptance, as compared to configuration A. In order to prove the hypotheses the constructs were operationalized using the following dependent measures.

Objective performance measurements examine the objective performed tries for planning or re-planning a mission. An additional performance variable is the amount of the error rate while planning and re-planning. Human-system interactions were measured to determine behavior changes. Therefore, the interaction time and the interaction activity were counted. The interaction time is defined as time where the operator is actively interacting with the cognitive agent in a particular situation. The interaction time can be quantified by the observation of manual actions (clicks) on the touch displays or by use of eye tracking data during monitoring tasks. The interaction activity is measured with the number of touch-screen clicks over time. Subjective dimensions were used to measure the subjective workload and performance by the standardized NASA-TLX (Hart, 1986; Hart & Staveland, 1988). The six-scale questionnaire was answered after planning and commanding, monitoring and re-planning in both missions. Additionally to the performance measuring, to observe the situation awareness of the operator the SAGAT questionnaire (Endsley, 1988) was performed after re-planning. At the end of both missions the operator had to complete a questionnaire for the subjective evaluation of acceptance and trust in automation (Lee & Moray, 1992; Lee & See, 2004). The questionnaire consists of four dimensions (system interaction, system behavior, system information and overall system).

Participants and experimental conditions

The sample consists of 13 officers of the German Armed Forces. The participants aged between 21 to 27 years (M_{age}=24.2) are recruited from the University of the Bundeswehr Munich. Participants include 12 male and one female student.

The operators control the UAV-system from a *Ground Control Station* (GCS), using a Human-Machine Interface (HMI) consisting of two multi-touch displays. Inputs can be made using touch screen or mouse controls. Room dividers shield the GCS to avoid visual distractions for the operators, while a headset damps external sounds and facilitates intercom. The lower screen features the MPCS (Theissing & Schulte, 2015) for task-based UAV guidance and automation monitoring using a map display of the mission area and a graphical representation of the tactical situation. The interaction with mission elements allows the formulation of tasks provided to the UAVs. The upper screen of the GCS features a modular sensor interface showing the live sensor-feed for the UAV. The sensor display is used to observe the surveillance area and to detect vehicle movements. The GCS experimental setup includes an eye-tracking system, measuring the operators' focal point on the screens during the experiments. The experimenter uses an external workstation for the manipulation and control of the experiment's tactical situation and events (Borchers, 2014; Werner, 2014).

Experimental procedure

During the experiment, each participant performed *mission I* and *mission II*. The experiment has a total duration of approximately three hours, including the mission preparation, the actual missions and standardized briefings and debriefings.

The missions have an identical general layout with similarly complex mission scenarios but differing mission events and dynamic threats. The sample is split into two, whereas the first half performs *mission I* in system

configuration A followed by *mission II* in system configuration B. The second half performs both missions with switched system configurations. The assignment to the conditions was randomized to minimize systematic effects.

Each subject performs two consecutive missions (*mission I* and *mission II*), each with similar mission layout and tasks. An island, conquered by hostile forces, is to be retaken from an adjacent island and therefore periodic reconnaissance missions have to be performed by the operator and his UAV. In this scope, own troops should be supported, hostile targets detected and identified as well as areas scanned (reconnoitered). At the beginning of each mission the UAV takes off from its home base (indicated by the blue square below the UAV symbol) and crosses the FLOT (red line) through transit corridors (blue). Enemy *Surface-to-Air Missile Sites* (*SAM-Sites*), indicated in red, generally have to be avoided by the UAV. Reconnaissance targets are objects or areas (yellow). The main objective for *mission I* is to perform a detection task in two areas (A, B) and a reconnaissance in two additional areas (C, D). The operator should plan and complete the mission. After completing the main objective the mission is interrupted and the operator has to re-plan. In *mission II*, area A and B should be detected and area C cleared. During the mission positions are also changing and the UAV must land on an alternative airfield (re-planning). The procedure and all instructions are standardized.

Experimental Results

The explorative hypotheses are calculated with the Wilcoxon signed-rank test. For the exploratory data analyses, the significance level was set to 5%.

While planning and commanding the mission, there are no differences between configuration A and B in the following variables. The interaction time is in both configurations not significant different (Z= -0.629; p=.277), as well as the interaction activity (clicks) (Z= -0.594; p=.294). The data of the subjective workload shows no significant differences between configuration A and B (Z= -0.874; p=.207), as well as configuration A to baseline data (Z= -0.314; p=.395) and configuration B to baseline data (Z= -0.735; p=.242). Figure 5 presents the results of the number of clicks and interaction time, while planning and commanding.

While monitoring the mission, in configuration A (MW = 4.24, SD = 3.79) the interaction time is significant shorter than in configuration B (MW = 6.04; SD = 3.32; Z= -1.784; p=.042). The longer interaction time during configuration B during the monitoring phase might be an indication for a higher task load. But the higher task load does not cause a higher subjective workload (Z= -0.559; p=.300). The operator must perceive more information, which needs more time. An explanation for no differences between configuration A and B might be that a different interaction behavior does not exist because of the short latency, till planning and commanding the mission.

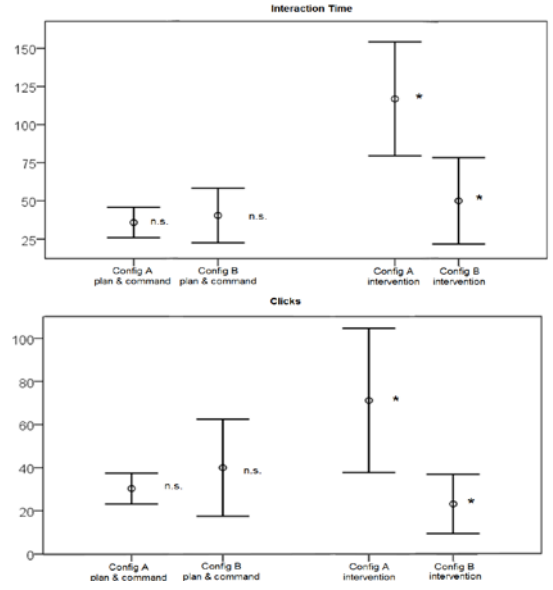


Figure 5. Result of the interaction time and the number of interactions during planning and commanding as compared to re-planning in configurations A and B (n.s. = Not significant, * = p<0.01).

While re-planning the mission, in configuration A (MW = 2.69; SD = 1.03) the interaction activity of re-planning is significant higher than in configuration B (MW = 1.67; SD = 0.98; Z= -1.895; p=.045). The error rate however is significant higher in configuration A (MW = 2.15; SD = 0.90), as compared to configuration B (MW = 0.62; SD = 0.96; Z= -2.831; p=.001). Additionally, in configuration A (MW = 68.46; SD = 51.31) significant more clicks are made than in configuration B (MW = 23.00; SD = 21.61; Z= -3.059; p=.000). The interaction time in configuration A (MW = 112.60; SD = 58.38) is significantly higher than in configuration B (MW = 50.03; SD = 44.54; Z= -2.432; p=.006). Compared to the baseline data (MW = 17.15; SD = 8.46), in configuration A (MW = 38.79; SD = 17.13) the subjective workload is significant higher (Z= -3.059; p=.000). In configuration B (MW = 37.86; SD = 13.00) the workload is also higher than in the baseline (Z= -3.182; p=.000), but between configuration A and B are no significant workload differences (Z= -0.105; p=.473). The situation awareness after re-planning in configuration A and B (Z= -0.735; p=.236) is equal. Figure 5 presents the results of the number of clicks and the interaction time for the re-planning phase.

The evaluation shows that in general there are differences in the subjective ratings between the two configurations. Furthermore, configuration A is less accepted than configuration B in the subscales system behavior (configuration A: MW = 3.95; SD = 1.46; configuration B MW = 4.70; SD = 1.12; Z=-2.473; p=.005) and overall system (configuration A: MW = 112.60; SD = 58.38; configuration B MW = 112.60; SD = 58.38; Z=-2.665; p=.002). The operators' trust in automation in configuration A (MW = 4.44; SD = 1.08) is lower than in configuration B (MW = 5.04; SD = 0.77; Z=-2.518; p=.004), because the system acts like the operator requests. This might be a reason why the overall system itself is more accepted.

Conclusions

This paper describes an approach to cognitive agent feedback and its experimental evaluation concerning human delegation behavior. The developed cognitive agent is designed to support the human operator to execute supervisory control functions in a highly automated technical system.

While no effects of the provided feedback could be identified for the initial planning phase of the mission, the re-planning phase is positively affected by the provided agent feedback. More feedback information requires more processing time of the human operator while monitoring the system, but raises overall mission efficiency (lower error rate) and lowers the operator's interaction and failure rate during re-planning.

References

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19* (6), 775–779.

Billings, C. E. (1997). *Aviation Automation: The Search for a Human-Centered Approach*. Mahwah, NJ: LEA

Borchers, A. (2014). *Konzeptionierung und Durchführung einer Versuchsreihe zur Evaluierung eines kognitiven Agenten in Bezug auf die Planung und Kommandierung von Aufklärungsmissionen. [Configuration and performance of a research series, to evaluate a cognitive agent, while planning and intervention of a reconnaissance mission] (*Master Thesis). Universität der Bundeswehr, München, Germany.

Clauss, S., Kriegel, M. & Schulte, A. (2013). *UAV Capability Management using Agent Supervisory Control*. In: AIAA Infotech@Aerospace Conference 2013. Boston, Massachusetts, USA. 19-22 August 2013.

Clauss, S. & Schulte, A. (2014). *Implications for operator interactions in an agent supervisory control relationship*. In: International Conference on Unmanned Aircraft Systems (ICUAS). Orlando, FL. 27-30 May 2014

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Human Factors and Ergonomics Society Annual Meeting, 32* (2), 97–101.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. NL: Elsevier.

NASA (1986). Task Load Index (TLX): Computerized version (Version 1.0). Moffett Field, CA: Human Research Performance Group, NASA Ames Research Center.

Leana, C. R. (1986). Predictors and Consequences of Delegation. *TAMJ, 29* (4), 754–774.

Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *JoHFES, 46* (1), 50–80.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 10* (35), 1243–1270.

Miller, C. A. (2002). *Definitions and dimensions of etiquette*. North Falmouth, MA: AAAI Fall Symposium on Etiquette for Human-Computer Work.

Onken, R. & Schulte, A. (2010). *System-ergonomic design of cognitive automation: Dual-mode cognitive design of vehicle guidance and control work systems*. Berlin, Heidelberg: Springer-Verlag.

Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *HF, 2* (39), 230–253.

Sheridan, T. B. (1992). *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: MIT Press.

Theissing, N., & Schulte, A. (2014). *Flight Management Assistance through Cognitive Automation Adapting to the Operator's State of Mind*. In: 31st EAAP Conference. Valetta, Malta. 22-26 September 2014.

Uhrmann, J. & Schulte, A. (2011). *Task-based Guidance of Multiple UAV Using Cognitive Automation*. In: The Third International Conference on Advanced Cognitive Technologies and Applications. COGNITIVE 2011. Rome, Italy. 25-30 September 2011.

Werner, J. (2014). *Konzeptionierung und Durchführung einer Versuchsreihe zur Evaluierung eines kognitiven Agenten in Bezug auf die Überwachung von Aufklärungsmissionen. [Configuration and performance of a research series, to evaluate a cognitive agent, while monitoring a reconnaissance mission] (*Master Thesis). Universität der Bundeswehr, München, Germany.

Wiener, E. (1988). Cockpit Automation. In E. Wiener (Ed.), *Human Factors in Aviation* (pp. 433–462). CA: AP.

AN ECOLOGICAL APPROACH TO THE SUPERVISORY CONTROL OF UAV SWARMS

Chistian Fuchs, Clark Borst, Guido de Croon, René van Paassen, and Max Mulder

Faculty of Aerospace Engineering, Section Control and Simulation, Delft University of Technology, 2629 HS Delft, The Netherlands

Advances in miniaturized computer technology have made it possible for a single Unmanned Aerial Vehicle (UAV) to complete its mission autonomously. This also sparked interest in having swarms of UAVs that are cooperating as a team on a single mission. The level of automation involved in the control of UAVswarms will also change the role of the human operator. That is, instead of manually controlling the movements of the individual UAVs, the system operator will need to perform higher-level mission management tasks. However, most ground control stations are still tailored to the control of single UAVs by portraying raw flight status data on cockpit-like instruments. In this paper, the ecological interface design paradigm is used to enhance the human-machine interface of a ground control station to support mission management for UAV swarms. As a case study, a generic ground-surveillance mission with four UAVs is envisioned. A preliminary evaluation study with 10 participants showed that the enhanced interface successfully enables operators to control a swarm of four UAVs and to resolve failures during mission execution.

The use of Unmanned Aerial Vehicles (UAVs) has grown rapidly over the past years. Advances in the fields of materials and computer technology provided the means to develop UAVs for a multitude of civil applications, such as search and rescue operations and wild life monitoring and protection. While the reasons to use a single UAV are manifold, it is often advantageous to use several UAVs that are operating as a team, for example, to execute tasks at different locations simultaneously or observe a larger area in a shorter time.

Current systems and legislation, however, still require at least one operator, if not more, to be in control of a single UAV. As a result, a ground control station is tailored to the control of a single UAV by portraying raw flight status data on cockpit-like instruments. For a UAV swarm, the number of instruments would then simply multiply, making the control of UAV swarms highly labor intensive and difficult in terms of extracting higher-level mission management information. Thus, some form of automation support and interface enhancements would be required to successfully control UAV swarms.

Whereas the majority of UAV swarming research is focused on improving or increasing the degree of automation (Prinet, Terhune, & Sarter, 2012; Cummings & Mitchell, 2006), the work described in this paper focuses on improving the visual representation in an existing ground control station to support higher-level mission management. By utilizing Ecological Interface Design (EID) principles (Vicente & Rasmussen, 1992), the enhanced interface will make the connections between low-level state information and higher-level mission management more salient. The resulting interface is expected to give operators a better understanding of the system and enable them to creatively solve arising problems, without being limited to prescribed solutions.

In this preliminary research, the scope of the work domain is limited to a simplified ground-surveillance mission consisting of four UAVs, where the emphasis is put on how the lower-level system constraints (e.g., the UAV battery levels and the wind condition of the environment) affect the higher- level joint mission plan of the swarm. To study the effect of the visualizations on human performance, a human-in-the-loop evaluation study is performed to gather feedback and test how well operators can control a UAV swarm when unexpected problems are introduced that jeopardize the mission's success.

Work Domain Analysis

The scope of the work domain analysis entails a generic ground-surveillance mission consisting of four UAVs. All four UAVs are assumed to possess autonomous navigation capabilities and be able to perform individual missions comprising of different mission elements. How and by what technologies these capabilities are achieved is out of the scope for this analysis, however.

The results of a WDA can be summarized in an Abstraction Hierarchy (AH). This hierarchy describes the system at different levels of abstraction – ranging from the functional purpose of the entire system at the top to the physical form of individual components at the bottom. Importantly, it also shows how different elements relate to each other. According to Vicente and Rasmussen (1992), the AH is a psychological-relevant way to organize and structure information in order to facilitate top-down and bottom-up reasoning about the system. Thus, a WDA and the AH should be considered as powerful critical thinking tools to help an interface designer make informed decisions about *what* to put on the interface and how all constraints relate to each other. It does not, however, inform the designer *how* to visualize the constraints on the interface. Given the scope of this work domain, the resulting AH for this case study is shown in Figure 1.

The resulting AH (Figure 1) clearly indicates that the individual mission of each single UAV simply adds up to the global mission of the swarm. Further, to enhance, or, improve, the human-machine interface of a typical UAV ground control station, it would be wise to first study how well the work domain elements, found in the AH, are represented in such interfaces. An analysis of two popular ground control stations indicated that a typical UAV interface depicts low-level state information in the form of raw numbers and/or in the form of cockpit-like flight status instruments, but fails to integrate that into higher-level system functionalities, such as the expected endurance and range of the UAV, that ultimately propagates upward into the expected ground coverage required to complete the surveillance mission. Thus, the opportunity to improve such a UAV interface would be to make the higher-level system functionalities explicit by means of visualizations that enable a system operator to link higher-level system functionalities to lower-level system properties (i.e., support top-down problem-solving activities) and vice versa (i.e., bottom-up reasoning and problem-solving activities).

To visualize the constraints and their dynamics, capturing the laws of physics governing them is necessary. Here, the equations describing this swarming domain consist primarily of aircraft performance equations for fixed-wing, propeller-type aircraft with electric propulsion.

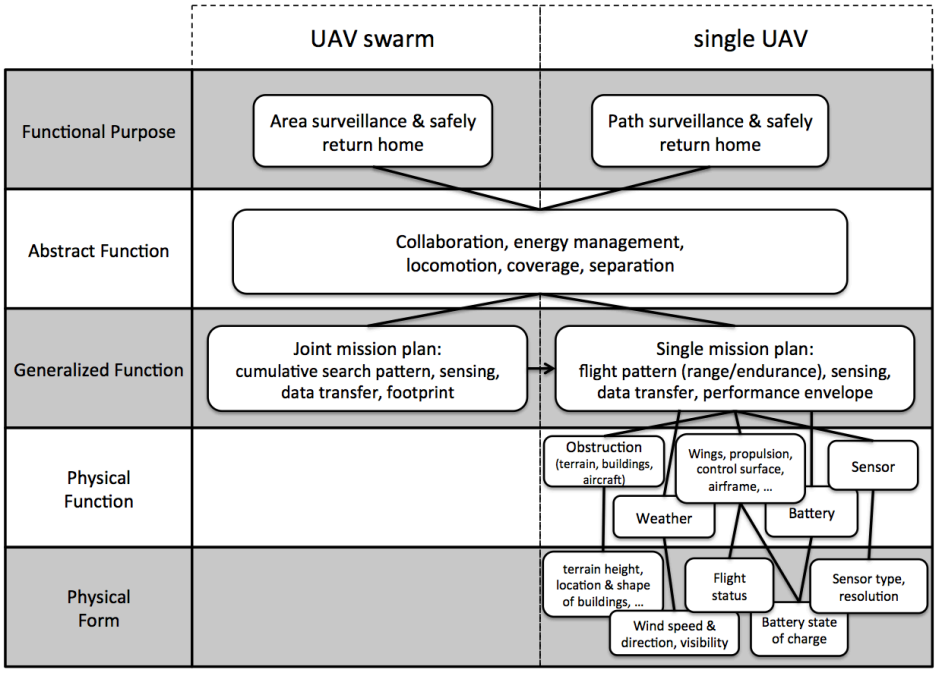


Figure 1. Preliminary abstraction hierarchy (with means-ends links) for a generic ground-surveillance mission of a UAV swarm.

Ecological Ground Control Station

Combining the WDA and the mathematical foundation of the UAV control problem, a set of visualizations is created to enhance the UAV ground control station. As there is no predefined procedure or recipe to follow to create the visual forms of the constraints discovered in the WDA, this part of the ecological approach is sometimes referred to as overcoming the creative gap. Here, the basis for all visualizations is a depiction of the *required* system behavior (e.g., required coverage, required power and energy, required battery state of charge, etc.), the *expected* system behavior (e.g., predicted coverage, predicted power, predicted battery level, etc.), and the *current* state of system behavior (e.g., current coverage, current power, current battery level, etc.). It is expected that such visualizations would help the operator to identify deviations from the mission, trace back the cause of the deviation (e.g., a low battery level in a single UAV), and formulate and implement alternative solutions to complete the mission. A screen capture of the proposed enhanced interface is shown in Figure 2. Besides coloring the waypoints, flight segments, and predicted/expected coverage according to the current and predicted energy state of the UAVs, the most notable addition the state of charge indicator for each UAV (Figure 2, fleet overview) and how it connects to the mission plan view.

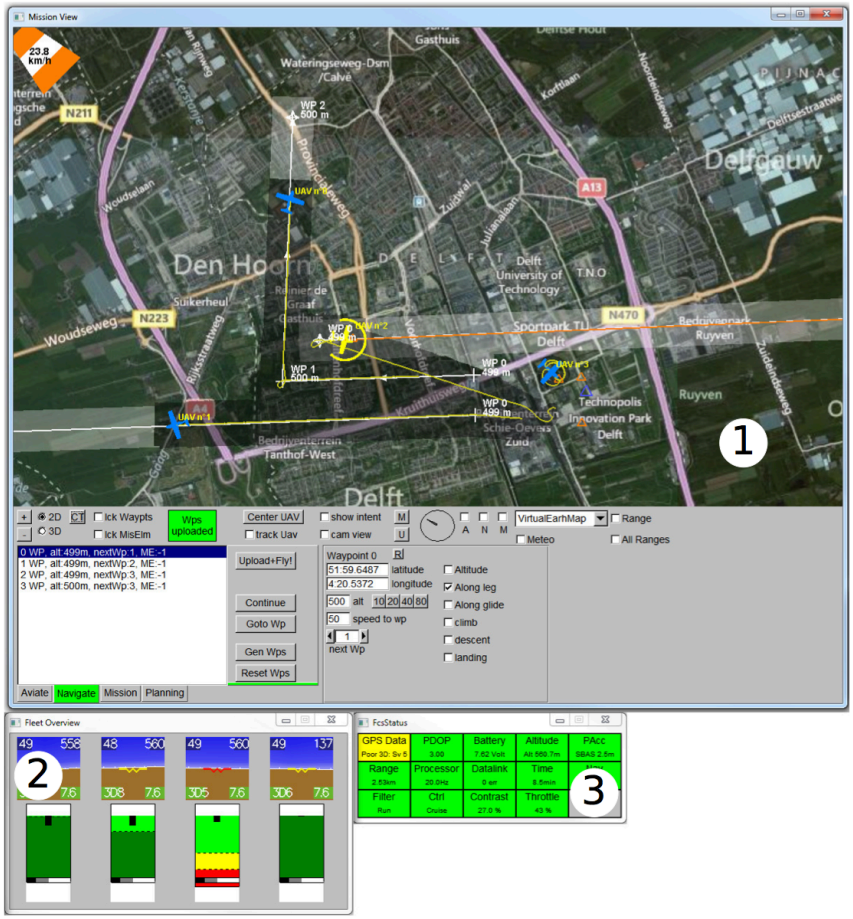
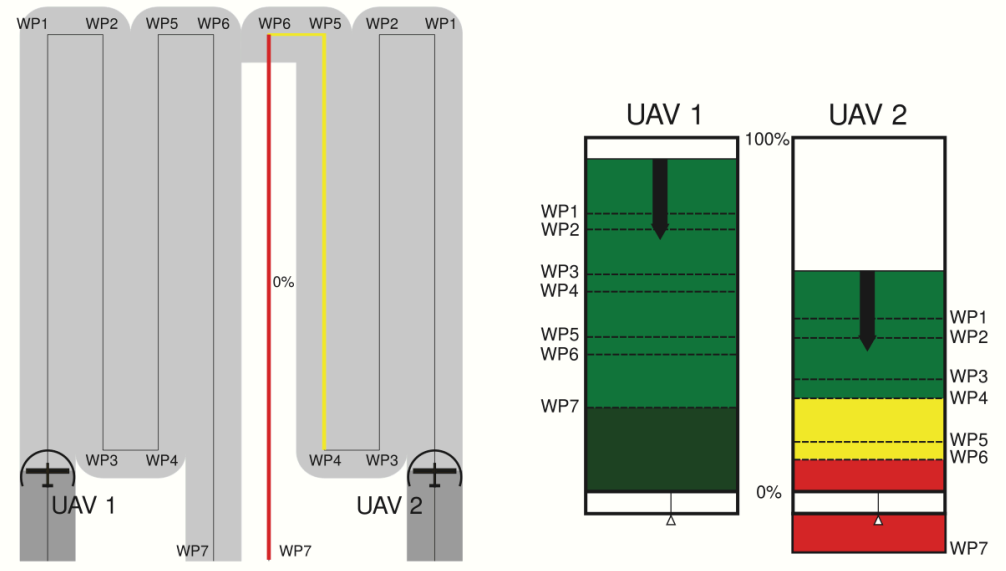


Figure 2. Ground control station interface with ecological additions. Depicted are the mission view (1) , fleet overview (2) and flight control system status window (3).

To visualize the abstract function of coverage, a shaded area around the flight trajectories of all UAVs is used, as shown in Figure 3(a). By using different shades, it is possible to show different states of coverage. Areas that are expected to be covered are shaded lightly and areas that have already been covered are shaded dark. Those areas that cannot be covered (e.g., a UAV cannot complete its flight plan and return to home, because of a low

battery level) leave a "hole" in the shading, e.g., between waypoints 6 and 7 of UAV 2 in Figure 3(a). This would give the system supervisor a clear cue about the predicted mission accomplishment of a single UAV, and thus also the mission accomplishment of the entire swarm. The size of the shaded area depends on the altitude of the waypoints that define the flight trajectory, i.e., a larger area will be covered (and thus shaded) at a higher altitude of the UAV. This thus represents the means-ends link between the flight status of the UAV and the higher-level coverage goal of the system. However, a higher altitude also means less surveillance accuracy when the camera has a fixed resolution. In this prototype, however, this relationship has not yet been modeled. The link between the battery's state of charge (SOC) and coverage is that no shading will be applied when the expected SOC at a waypoint is zero and the waypoint can therefore not be reached. This gives the operator a clear cue that something is amiss and further fault diagnosis is required.



(a) Stylized map view showing the coverage shading and waypoint coloring. For illustration purposes, the exact position at which energy for UAV 2 will run out is marked with "0%".

(b) State-of-charge indicator. UAV 1 still has energy at waypoint 7, while UAV 2 needs additional energy, as shown by the additional red coloring below the 0% marker.

Figure 3. Side by side view of a stylized map view and the state-of-charge indicator for two UAVs. Waypoint numbers (WP1 - WP7) in both depictions correspond to each other. UAV 1 is shown on the left and UAV 2 is shown on the right.

Working with the Interface

The envisioned usage of the ecological interface developed for this study is as follows. If the goal is to surveil a particular area on the Earth's surface, the operator can setup the individual flight plans of the UAVs by positioning waypoints so as to create a cumulative search pattern that fully covers the target area. Entering and dragging waypoints by direct manipulation can be regarded as skill-based behavior, whereas comparing the surveillance area with the expected cumulative coverage patterns would be classified as rule-based behavior (driven by "if-then" rules). After creating the flight plans of the UAVs, the plans can be uploaded to individual UAVs, and each UAV will then automatically fly the intended trajectories. During flight, the operator can monitor the progress of the surveillance mission by comparing the expected coverage with the current (completed) coverage. As such, the operator can stay at higher levels of (control) abstraction and can use rule-based behavior to monitor the mission. If everything is working according to plan, the operator will most likely remain at this level. Whenever a problem would arise, it is expected that the operator will first be alerted by observing a gap in the expected coverage. This would then trigger problem-solving activities to replan the UAV trajectories so as to fill the gap in coverage. The gap in coverage can be caused by many things, such as a higher battery-depletion rate than expected, a changed

wind condition that requires more energy to fly the ground-referenced trajectory and still return safely to home, a failed data transmission (data link problem) to the ground station, or perhaps a combination of these events. In case a problem is identified with a UAVs battery level, such as shown in Figure 3, the operator could alter the flight plan of another UAV (e.g., by choosing a UAV with an excess in battery charge after it has completed its own single mission plan) to fill the coverage gap. For instance, the position of the waypoints and/or the altitude settings can be manipulated to have another UAV successfully take over the mission of a failing UAV. Considering Figure 3, the operator could let UAV2 fly back to home upon reaching its WP4, and change the positions and altitudes of WP6 and WP7 of UAV 1 to make up for the gap in coverage. Of course, upon manipulating the waypoints the operator should ensure that the new flight pattern is feasible by observing the required energy and expected battery power at the new waypoints. As such, the expected nominal strategy to resolve a mission problem would be to delete the problematic waypoints and increase the altitude of the remaining waypoints, while sticking to the general search pattern of the predefined flight plans.

Evaluation Study

To observe how operators would use the ecological enhancements and interface features, an exploratory evaluation study was performed. The focus of this evaluation study was to observe a user's problem-solving activities during mission management of a UAV swarm, consisting of four UAVs, in the presence of several system failures. Ten subjects – consisting of four faculty employees, who had previous experience with UAVs, and six aerospace students – were asked to perform a mission with five different initial conditions. The objective of the mission was to survey the town of Nootdorp (nearby the city of Delft, in The Netherlands) by loading and maintaining a predefined flight plan. This flight plan was equal to the one shown in Figure 3, but extended to four UAVs. Since pairs of UAVs are converging, this flight plan makes it easy to compensate for failures by a single UAV. Further, coverage of a predefined area had to be perfect and there should be no waypoint from which a UAV could not return to base. Finally, possible collisions between UAVs could be ignored under the assumption that each UAV has an autonomous sense-and-avoid capability.

Five test scenarios were defined that covered failures induced internally at the battery and externally by the wind condition. On top of that, they covered failures at a single UAV and at multiple UAVs. To solve problems during the mission, it was possible to change the number and position of waypoints. Participants were therefore not constrained to only use the predefined flight plan but could chose any order of waypoints. However, the altitude of waypoints was limited between 200 m and 500 m. Based on the flight plan and the definition of the scenarios in Table 1, the expected solution strategies are summarized as follows:

1. *Scenario 1: Delete problematic waypoints of all four UAVs and increase the altitude of the remaining waypoints to 500 m*
2. *Scenario 2: No solution required*
3. *Scenario 3: Delete problematic waypoints of UAV4 and increase altitude of UAV3's waypoints to 500 m*
4. *Scenario 4: No solution possible (with nominal strategy)*
5. *Scenario 5: Delete problematic waypoints of UAV2 and increase altitude of UAV 1's waypoints to 500 m*

After each run, participants had to fill out a questionnaire. The first part of this questionnaire contained open questions about the participants' decision process. The second part contained a list of the improvements made to the interface that had to be rated on a Likert-scale from one (bad) to ten (good), according to their perceived usefulness.

Results

Participant feedback

The interface items that were considered very useful were the predicted coverage, the coloring of the waypoints, and the coloring of the lines connecting the waypoints. This is also in line with the observation how the participants solved problems encountered in the scenarios. Feedback to the open questions, as well as the audio recordings, reveal that participants found and solved problems at a high level of abstraction. Specifically, the coloring was used to realize that a problem was present, while a solution was found using the coverage shading. Originally it was expected that problems are found at the abstract function of coverage. However, adding a bright

red line to the map provides a much stronger cue than removing a light shading. The interface features that were considered somewhat useful were the current SOC and the expected SOC at future waypoints. It was observed that the participants used the SOC indicator for two purposes: When the map was not centered at the search area, so that the waypoints were not visible on screen, participants used the SOC indicators to find potential problems. Most of the time it was used to match the flight plans visible on the map with the corresponding UAVs. This reveals a considerable problem with how the joint mission plan is visualized in the interface. By showing all flight plans simulateously, without further distinction between flight plans, operators were forced to use alternative means to identify the problematic UAV. Incidentally, this is the number one feedback given by participants. Thus it appears that the means-ends relationship between the UAV icon and its corresponding flight plan is not made explicit enough in the interface. This problem will likely be amplified for a swarm consisting of more than four UAVs.

Mission success

Out of 40 individual runs, eight were not finished successfully. Of those eight failures, four missions arguably failed due to unnecessary mistakes made by the participant , such as not uploading the flight plans or missing a small part of the search area. Most surprisingly, the envisioned unsolvable scenario 4 was solved six out of ten times. Participants did so by adopting a different strategy than anticipated, which was to delete the problematic waypoints and increase the altitude of the remaining waypoints, while sticking to the general pattern of the predefined flight plan. Instead, they also changed the order of UAVs within the search pattern – a simple, but unanticipated solution strategy. This result clearly demonstrates the power of a constraint-based interface, as it supports creative problem-solving activities.

Conclusion

Following an ecological approach to interface design, the human-machine interface of an existing ground control station was enhanced to support mission management and fault diagnosis of a UAV swarm. These improvements visualize how low-level system properties, such as battery level, wind speed, and wind direction propagate to a higher-level system goal of achieving coverage in a generic ground-surveillance mission. An evaluation study showed that operators could successfully use these new interface elements to control a swarm of four UAVs and solve problems during mission execution. The results of the evaluation study showed that operators had a better system understanding and that it promoted creative problem-solving activities to scenarios that could not have been solved by a predetermined strategy. However, the results also showed that the current interface still required control actions to be performed per single UAV, making it labor intensive to change mission parameters for swarms consisting of more than four UAVs.

References

Prinet, J.C., Terhune, A., & Sarter, N.B. (2012). Supporting Dynamic Re-Planning In Multiple Uav Control: A Comparison of 3 Levels of Automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56(1):423–427. ISSN 1541-9312.

Cummings, M.L., and Mitchell, P.J. (2006). Automated Scheduling Decision Support for Supervisory Control of Multiple UAVs. Journal of Aerospace Computing, Information, and Communication, 3(6):294–308. ISSN 1542-9423.

Vicente, K.J., and Rasmussen, J. (1992). Ecological interface design: theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(4):589–606.

EFFECT OF CONTROL LATENCY ON UNMANNED AIRCRAFT SYSTEMS DURING CRITICAL PHASES OF FLIGHT

Carolina M. Zingale, Ph.D.
Federal Aviation Administration, Human Factors Branch
FAA William J. Hughes Technical Center, Atlantic City International Airport, NJ 08405

Eric G. Taylor, Ph.D.
T.G. O'Brien & Associates, Inc.
FAA William J. Hughes Technical Center, Atlantic City International Airport, NJ 08405

Unmanned Aircraft Systems (UAS) are controlled remotely via terrestrial or satellite-based radio link rather than by a pilot in the cockpit. The remote nature of the transmission results in latencies (time between pilot input and feedback indicating aircraft response) that are typically longer than those in manned aircraft. Researchers from the FAA Human Factors Branch conducted a simulation to investigate the effect of control latencies during takeoff and landing scenarios in UAS with low levels of automation. We evaluated one of four latencies (180, 494, 750, 1026 ms) in each test scenario. Half of the scenarios included crosswinds. Data obtained from 11 UAS pilots indicated that as latency increased aircraft performance and pilot ratings of aircraft handling were negatively affected (e.g., more deviations from pattern). Overall, control latencies above 494 ms adversely affected pilot and aircraft performance relative to baseline.

Unmanned aircraft systems (UAS) are increasingly identified for use in diverse activities such as aerial photography, package delivery, surveillance, and search and rescue. The Association for Unmanned Vehicle Systems International (Jenkins & Vasigh, 2013) predicted that the UAS industry will generate $10 billion of annual revenue once UAS are integrated into the National Airspace System (NAS). To integrate UAS into the NAS, the FAA needs to determine the characteristics that define acceptable UAS performance to ensure that the NAS maintains the highest levels of safety and efficiency. UAS are piloted remotely via terrestrial or satellite-based radio link that result in control latencies that are typically longer than those in manned aircraft. UAS control latencies range from hundreds of milliseconds to several seconds (Walsh, 2009), whereas manned aircraft control latencies are typically less than 150 milliseconds (e.g., Berry, 1985). Control latencies have a substantial impact on precision motor control tasks (e.g., Chen, Haas, & Barnes, 2007) and this effect is difficult or impossible to overcome (Taylor & Zingale, 2014). Hence, it is essential for the FAA to know the specific impact that latencies of various magnitudes may have on UAS operations.

Methods

Eleven Air Force pilots, with experience in UAS launch and recovery (takeoff and landing) operations, participated in this simulation at the FAA William J. Hughes Technical Center's NextGen Integration and Evaluation Capability (NIEC) Laboratory. The pilots used a UAS simulator with a stick and rudder control system to complete short (7 minute) takeoff and landing scenarios under four different control latencies (180, 494, 750, 1026 ms). The latencies included the "TT95" values (time before which 95% of transactions are completed) reported by Walsh (2009) for line-of-sight (494 ms) and beyond-line-of-sight (1026 ms) operations. We included a baseline latency of 180 ms to correspond to the latency of a comparable aircraft with no wireless control link. Half of the scenarios included crosswinds. The simulator heads-up displays showed an out-the-window view from the UAS and, separately, showed an overview sectional map that depicted the traffic pattern overlays and a restricted operating zone (ROZ) outside the patterns. The simulator recorded pilot control inputs and aircraft performance outputs, positions, and attitudes. An FAA pilot who was part of the research team acted as the sensor operator (SO) to provide call-outs as requested by the pilot (e.g., stating current altitude) and to record the time at which a pilot indicated a need to execute a go-around.

Experimental Design

We had three primary independent variables: latency, phase of flight (takeoff vs. landing), and presence or absence of crosswinds (0 or 14 knots). We ran half of the test scenarios with left closed turns and the other half with right closed turns. We randomized test orders for each participant and counterbalanced test orders across participants.

Procedure

Before each scenario, the researchers informed the pilot of the control latency (180, 494, 750, or 1026 ms), whether or not crosswinds would be present, and the traffic pattern direction. For all scenarios, pilots were instructed to stay within +/- 100 feet of the target altitude of 1,000 ft (+/- 10 knots of the target velocity of 105 knots) and within +/- 5 degrees heading as depicted by the pattern overlay. For takeoff scenarios, pilots were instructed to take off from the assigned runway, to climb to the traffic pattern altitude, and to make the appropriate turns. For landing scenarios, pilots started in the air at the mid-way point of the downwind leg and were instructed to complete their approach along the indicated pattern and come to a full stop on the runway. Pilots were told to *always land the aircraft* unless otherwise advised by their SO. In the event they would have executed a go-around, the pilots were instructed to alert the SO, who recorded this information. After each scenario, the pilots completed questionnaires that included the Cooper-Harper (CH) Handling Qualities Ratings Scale (Cooper & Harper, 1969) to provide an assessment of aircraft performance. Ratings on the CH scale range from 1 (*satisfactory*) to 10 (*uncontrollable*).

Data Analysis

We analyzed data on aircraft and pilot performance, including location of UAS during flight relative to pattern and runway centerline; variability in pilot command entries; deviation between target and actual UAS altitude; frequency of go-around requests; aircraft force at touchdown; and questionnaire ratings. We analyzed data from the takeoff and landing scenarios separately using Bayesian hierarchical linear modeling. We chose this framework because our design was unbalanced and to model outliers using robust estimation techniques (see Kruschke, 2010). Our predictor variables were latency, presence of crosswinds, and latency/crosswind interaction. To obtain credible values for our predictors, we estimated the "posterior" distribution of their regression coefficients via Markov Chain Monte Carlo. We report the "high-density interval" (HDI) from the posterior, or the range of most credible coefficients. Formally, the HDI is the interval corresponding to 95% of the posterior probability mass. Bayesian analysis does not entail the computation of *p*-values, but one may consider an effect "significant" if the HDI does not include 0. To provide maximum information pertaining to latency, we report the predicted impact of latency for conditions with crosswinds and without crosswinds separately. We do not provide HDIs for the effects of crosswinds because winds were not the focus of the report.

Results

We measured aircraft deviation from the flight path and runway centerline to evaluate performance while in the pattern and over the runway, respectively. Figure 1 presents the mean aircraft deviations from the pattern overlay for the takeoff scenarios (left) and the landing scenarios (right).

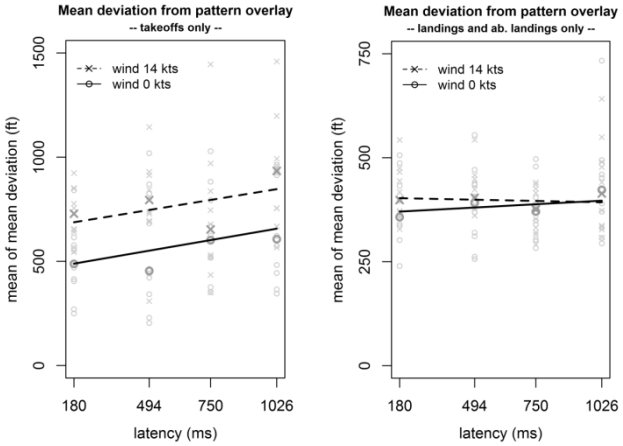


Figure 1. Summary of mean deviations from the pattern during takeoffs (left) and landings (right). Light grey points indicate means for each participant. Dark grey points indicate medians across participants. Black lines indicate the average regression fit from Bayesian analysis.

For takeoffs, the effect of latency was significant with and without crosswinds. Credible values for the latency effect in the presence of crosswinds were between 0.0321 and 0.3425 (95% HDI): an increase between 32.1 and 342.5 feet per second of added latency. Credible values for the latency effect in the absence of crosswinds were between 0.0423 and 0.3501 (95% HDI): an increase between 42.3 and 350.1 feet per second of added latency. For landings, the effect of latency was not significant for either crosswind condition.

We evaluated pilot adherence to the target altitude of 1,000 ft (+/-100 ft) while in the pattern. Figure 2 presents the aircraft deviations from the target altitude for takeoffs (left) and landings (right). For takeoffs, only the effect of latency in the presence of crosswinds was significant, with credible values between 0.0015 and 0.0428 (95% HDI): an increase between 1.5 and 42.8 feet per second of added latency. For landings, the effect of latency was not significant for either crosswind condition.

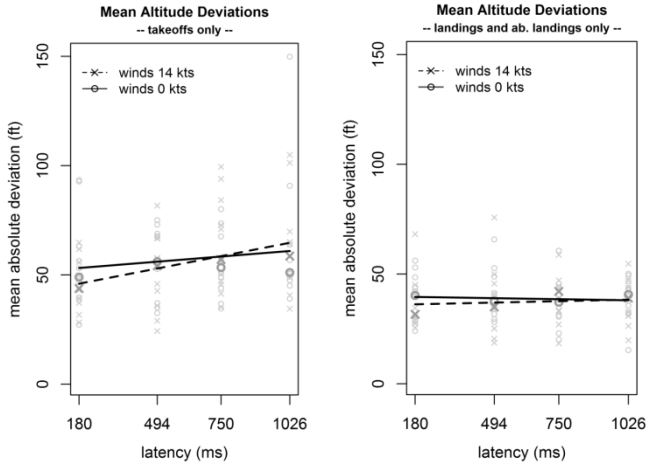


Figure 2. Summary of mean aircraft altitude deviations during takeoffs (left) and landings (right). Light grey points indicate means for each participant. Dark grey points indicate medians across participants. Black lines indicate the average regression fit from Bayesian analysis.

We measured the number of times the participants indicated they would have executed a go-around for each latency and crosswind condition. Figure 3 presents the mean proportion of indicated go-arounds for each latency condition. In this analysis, we used a logistic link function to model the raw binary responses. Only the effect of latency in the absence of crosswinds was significant, with credible values between 0.0003 and 0.0032 (95% HDI): an increase of approximately 17% more go-arounds from 0 to 1 seconds of latency.

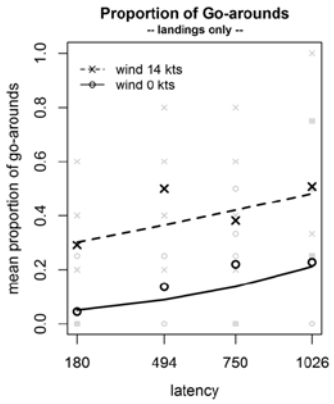


Figure 3. Summary of mean proportion of indicated go-arounds. Light grey points indicate means for each participant. Black points indicate means across participants. Black lines indicate the average regression fit from Bayesian analysis.

For landing scenarios, we also evaluated force of the unmanned aircraft (UA) at touchdown by analyzing the weight applied to the UA landing gear. We computed the maximum weight across the first five seconds after the initial point of touchdown. An example of a rough landing showed a weight of ~7,000 lbs applied at the initial point of touchdown, then another weight of ~15,000 lbs shortly after—whereas, an example of a smooth landing showed a weight of ~2,500 lbs applied at the initial point of touchdown. On average, maximum force at touchdown values for each pilot ranged, approximately, from 5,000 to 15,000 lbs. Only the effect of latency in the absence of crosswinds was significant, with credible values between 0.4510 and 5.6004 (95% HDI): an increase between 451.0 and 5,600.4 lbs per second of added latency.

For each pilot, we calculated the mean and maximum pilot CH rating for each combination of latency and crosswind conditions. We considered the maxima in addition to the means because allowable control latencies for UAS must ensure acceptable performance across a range of performance levels, not merely the average. The effect of latency, with and without crosswinds, was significant for both the mean and maximum ratings (see Figure 4). For mean ratings, credible values for the latency effect in the presence of crosswinds were between 0.0028 and 0.0052 (95% HDI): an increase in 2.8 to 5.2 points on the CH scale per second of added latency. Credible values for the latency effect in the absence of crosswinds were between 0.0022 and 0.0042 (95% HDI): an increase in 2.2 to 4.2 points on the CH scale per second of added latency. Maximum ratings were 1-2 points higher than mean ratings. For maximum ratings, credible values for the latency effect in the presence of crosswinds were between 0.0029 and 0.0050 (95% HDI): an increase in 2.9 to 5.0 points on the CH scale per second of added latency. Credible values for the latency effect in the absence of crosswinds were between 0.0027 and 0.0049 (95% HDI): an increase in 2.7 to 4.9 points on the CH scale per second of added latency. Critically, latencies in excess of 494 ms in conditions with 14 knot crosswinds resulted in maximum ratings in the “inadequate” handling qualities range.

Figure 4. Summary of mean and maximum Cooper-Harper ratings. Light grey points indicate individual means for each participant. Black points indicate means across participants. Black lines indicate the average regression fit from Bayesian analysis.

We also considered objective measures of control difficulty by evaluating two measures of pilot command input variability for yaw, pitch, and roll per scenario: (1) the standard deviations of the control surface deflections to capture the frequency and magnitude of command use and (2) the maximum input deflections to capture the most extreme inputs. The results from these analyses revealed that the participants used significantly more variable and extreme yaw inputs when aircraft were over the runway in conditions with higher control latencies (see Figure 5). The effect of latency on standard deviations of control surface deflections was significant with and without crosswinds. Credible values for the latency effect in the presence of crosswinds were between 0.00007 and 0.0025 (95% HDI): an increase between 0.07 and 2.5 degrees per second of added latency. Credible values for the latency effect in the absence of crosswinds were between 0.0006 and 0.0030 (95% HDI): an increase between 0.6 and 3.0 degrees per second of added latency. Likewise, the effect of latency on maximum input deflections was significant with and without crosswinds. Maximum yaw inputs indicated that the credible values for the latency effect in the presence of crosswinds were between 0.00005 and 0.0050 (95% HDI): an increase between 0.05 and 5.0 degrees per second of

added latency. Credible values for the latency effect in the absence of crosswinds were between 0.0015 and 0.0076 (95% HDI): an increase between 1.5 and 7.6 degrees per second of added latency.

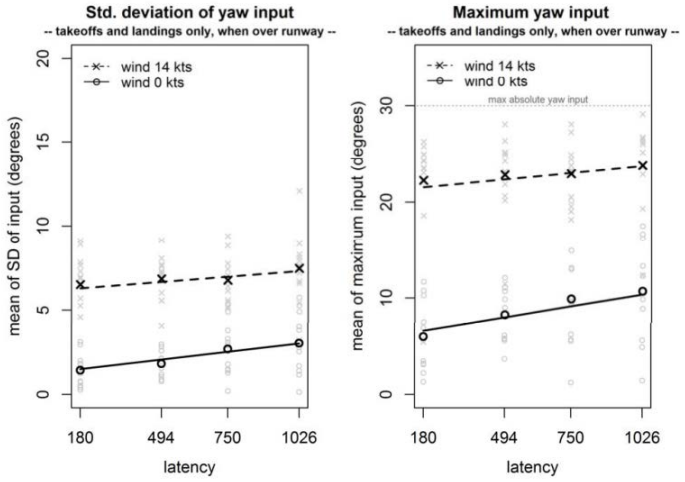


Figure 5. Summary of yaw input standard deviations (left) and maximum inputs (right) for aircraft over the runway in takeoff and landing conditions. Light grey points indicate means for each participant. Black points indicate means across participants. Black lines indicate the average regression fit from Bayesian analysis.

Discussion

Our simulation found that longer control latencies had an adverse effect on aircraft and pilot performance during takeoffs and landings in UAS with low levels of automation. These findings are consistent with prior research in aircraft simulators and in generic tracking tasks which show that latency impacts performance and subjective ratings of aircraft handling qualities. To our knowledge, however, this is the first study to directly measure the impact of latency on specific variables (e.g., go-around propensity and deviation from predetermined flight paths, which are directly relevant to UAS integration into the NAS).

We found that pilots indicated a need to execute more go-arounds with higher latencies. Nearly as many go-arounds were requested with 494 ms of latency as with 750 ms and 1026 ms. Force at touchdown as well as the number and extent of yaw inputs increased significantly as a function of latency. CH ratings increased notably with higher latencies, and maximum pilot CH ratings were in the “inadequate” range of the scale in conditions with latencies exceeding 494 ms when crosswinds were present. We also found that aircraft deviated more from the pattern and from the designated altitude with higher latencies during takeoffs. Interestingly, deviation from the runway centerline on landing was not significantly impacted by latency. We speculate that takeoffs in our simulation may have required more effort because the patterns called for the pilots to execute two turns almost immediately after takeoff. Thus, pilots may have been rushed to set their desired levels of power and trim commands. In contrast, pilots had more time during landing to prepare for final descent because the downwind leg allowed a long approach. Indeed, we designed our landing scenarios to incorporate a long approach based on feedback from UAS subject matter experts.

Acknowledgments

This research was sponsored by the Federal Aviation Administration’s Unmanned Aircraft Systems Integration Office (AFS-80) and conducted internal to the FAA with support from T.G. O’Brien & Associates, Inc. to investigate the integration of UAS into the NAS. We thank Sabrina Saunders-Hodge (ANG-C2), Karen Buondonno (ANG-C32), and John Warburton (ANG-C32) for their leadership within the FAA UAS Matrix Team; FAA sponsors (AFS-80), Ben Walsh, Kerin Olsen, and Ken Fugate for their support and direction; FAA UAS subject matter experts, John Steventon, and Marcello Mirabelli (AFS-80) for helping to develop and execute the simulation; Phil Maloney (ANG-C31) and Jean-Christophe Geffard (General Dynamics Information Technology, Inc.) for preparing and monitoring the simulator; Dan Fumosa (General Dynamics Information Technology, Inc.) for

configuring and troubleshooting data recording equipment; and the Department of Defense (DoD) for making UAS pilots available.

References

Berry, D. T. (1985*). In flight evaluation of pure time delays in pitch and roll* (Technical Memorandum: NASA-86744).

Chen, J. Y. C., Haas, E. C., & Barnes M. J. (2007). Human performance issues and user interface design for teleoperated robots. Systems, man, and cybernetics. Part C: Applications and reviews. *IEEE Transactions, 37*(6), 1231–1245.

Cooper, G. E., & Harper. R. P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (NASA TN D-5153). Moffett Field, CA: NASA.

Jenkins, D., & Vasigh, B. (2013). *The economic impact of unmanned aircraft systems integration in the United States.* Association for Unmanned Vehicle Systems International. Retrieved from http://qzprod.files.wordpress.com/2013/03/econ_report_full2.pdf

Kruschke, J. K. (2010). Doing Bayesian data analysis. Oxford, England: Academic Press.

Taylor, E. G., & Zingale, C. M. (2014). *Effect of control latency on tracking and handling qualities ratings: A review, meta-analysis, and implications for unmanned aircraft systems*. Manuscript submitted for publication.

Walsh, B. (2009). *UAS control and communication link performance – Latency* (SC203-CC0009_UAS). Washington, DC: RTCA, Inc.

ATTENTIONAL NARROWING:
A FIRST STEP TOWARDS CONTROLLED STUDIES OF A THREAT TO AVIATION SAFETY

Julie Prinet
Nadine Sarter

Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI

Attentional narrowing - the involuntary restriction of attention to a small set of data or one task/goal - is a major concern in many complex, high-risk domains. Research into this phenomenon is much needed but hampered by the difficulty of inducing it reliably in a controlled experimental setting. The present study tested the effectiveness of loud noise and high task demand for achieving this goal. Seven participants performed a visual search task in the context of a simplified air traffic control simulation. Performance and eye tracking data were recorded. Eye tracking metrics showed a narrowing of participants' visual attentional field under high demand; however, noise did not have a significant effect on attention allocation. The findings from this study represent an important step towards controlled studies of attentional narrowing. They also highlight the promise of eye tracking for detecting, in real time, breakdowns in attentional processes.

Attentional narrowing is a major and growing concern in many complex high-risk domains, such as aviation. Attentional narrowing refers to the "involuntary allocation of attention to a particular channel of information, diagnostic hypothesis, or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks" (Wickens, 2005). The phenomenon has been identified as a contributing factor to all controlled-flight-into-terrain (CFIT) accidents in the US Air Force (Shappell & Wiegmann, 2003). It also played a role in commercial aviation accidents such as the 1972 crash of Easter Airlines Flight 401 in which 103 people lost their lives. In this case, the failure of a landing gear indicator light during approach led all three pilots to focus on diagnosing the problem. As a result, they failed to notice that the autopilot had been disengaged inadvertently, which resulted in the aircraft gradually descending and ultimately crashing into the Everglades (NTSB, 1973).

Despite the potentially catastrophic consequences of attentional narrowing, empirical research into this phenomenon has been hampered by the difficulty of inducing it reliably in a controlled experimental setting. To date, two factors have been identified as likely triggers of attentional narrowing: (1) high motivational intensity and (2) arousal (Friedman & Förster, 2010; Harmon-Jones, Price & Gable, 2012). High motivational intensity refers to either a strong desire to approach and acquire an object associated with positive affect or to avoid negatively loaded stimuli that trigger, for instance, fear or disgust. Arousal, on the other hand, is defined as an energetic state of the organism (Bourne, 2003) which is affected by factors such as high task demand, loud noise, and extreme heat or humidity (Bursill, 1958; Hockey, 1970; Mackworth, 1965). In the presence of high motivational intensity and arousal, attention appears to narrow towards salient stimuli or stimuli that are perceived to be of high importance or priority (Bacon, 1974; Dirkin, 1983). It is important to note that, in most studies, this effect on attention allocation was inferred from performance data. Very few studies (such as Reimer, 2009) have examined the relationship between observed performance decrements and a person's allocation of visual resources. A promising, non-invasive approach to trace those changes is the use of eye tracking, which provides high-resolution data on eye movements and monitoring strategies in real time (Duchowski, 2007).

The goals of this pilot study were to (1) determine the effectiveness of two arousal-related stimuli – intermittent and aperiodic loud noise containing speech (based on Szalma and Hancock (2011)) and high task demand (e.g., Murata, 2004; Rantanen, 1999) – for inducing attentional narrowing, (2) trace how participants' allocation of visual attention is affected by those two factors and (3) identify eye tracking metrics that can capture, early and in real time, a narrowing of the attentional field. The application domain for this study was Air Traffic Control (ATC), a workplace that imposes high task demands and where breakdowns in attention allocation can have catastrophic consequences.

Methods

Participants

The participants in this study were 7 graduate students from the University of Michigan. Their average age was 25.1 years (SD = 4.7). Participants reported normal or corrected-to-normal vision. None of the participants had prior experience with ATC tasks.

Apparatus

The study was conducted using a simplified ATC simulation that was displayed on a 20-inch monitor, placed approximately 24 inches from the participants. Green aircraft icons were presented against a black background (see Figure 1). They were moving across the screen following a straight line, either horizontally or vertically, at a constant speed of 0.15 inches per second. The aircraft speed (shown in yellow) and altitude (shown in white) were presented in a data block to the lower right of the aircraft icon (see Figure 2).

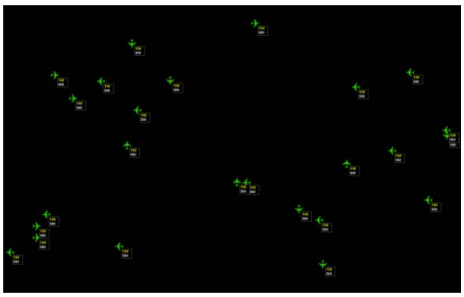


Figure 1. ATC simulation display

Figure 2. Aircraft icon and data block showing airspeed (150 kts) and altitude (FL (flight level) 500)

Tasks

All aircraft were assigned an airspeed of 150 knots and an altitude of 50,000 feet (FL 500). Occasionally, airspeed or altitude deviations occurred (airspeed values in the data block changed to 165 kts, 185 kts or 200 kts; altitude values changed to FL 370, FL 435 or FL 465). Participants were asked to monitor for these deviations and to return the airspeed or altitude to their assigned values as quickly as possible. To do so, they had to left-click the aircraft and choose the appropriate correction from three displayed options (in the case of airspeed deviations, the options were 15 kts, 35 kts and 50 kts; in the case of an altitude deviations, the options were 3,500 feet, 6,500 feet and 13,000 feet). Speed and altitude deviations lasted 6 seconds; if no correction was made during that time, the parameter automatically returned to its prescribed value. If the participant chose the appropriate correction value, the box around the data block turned green for 2 seconds; in case of an inappropriate correction, it turned red.

Experimental Conditions

(1) Noise manipulation. In the loud noise condition, a combination of white noise as well as aviation-related non-speech and speech alerts were presented via headset, at an average amplitude of 95 dBA. The various warnings were presented in an intermittent aperiodic fashion. In the no noise condition, participants were wearing a headset but no noise was presented.

(2) Task demand manipulation. The task demand was varied using the number of aircraft on the screen and the frequency at which speed or altitude deviations occurred. In the low task demand condition, 25 aircraft were presented on the screen, and an altitude or speed deviation occurred every 6 seconds. In case of high task demand, 80 aircrafts were presented, and an altitude or speed deviation occurred once per second.

Experiment Design and Procedure

The study employed a 2 (task demand: low or high) x 2 (noise level: no noise or loud noise) within-subject design. The order in which participants were presented with the two noise conditions was counterbalanced.

Participants were given a 10-minute training session to familiarize themselves with the ATC simulator. Next, they were asked to complete a 12-minute practice scenario during which they were asked to report observed airspeed or altitude deviations as fast as possible, and to apply the accurate correction to the data block. Then, the eye tracker was calibrated, and participants completed the two 12-minute experimental scenarios. Task demand was varied within each scenario, while noise was varied between the scenarios. Each scenario started with a 3-minute low task demand phase, followed by 6 minutes of high task demand, and ended with another 3-minute low task demand period. Participants were offered to take a 5-minute break between the two scenarios. Eye tracker calibration was repeated before each scenario. The entire session took approximately 1.2 hours to complete.

Dependent Measures

Performance data. The performance measure was the detection rate for speed and altitude deviations (expressed as the ratio of detected to total number of deviations), calculated for each of 9 screen sectors and for the overall display.

Eye tracking data. Eye tracking data was recorded using an ASL Eye-Trac D6 infrared-based, desktop-mounted eye tracker which samples at 60Hz. Eye tracking data consists of a series of fixations, or stable points of regard during which information processing occurs (Findlay, 2004), and saccades, or rapid eye movements between fixations during which no processing occurs (Yarbus, 1967). The following eye tracking metrics were calculated from the raw data: (1) the number of fixations on each of the 9 sectors (which can indicate problems with searching for information (Habuchi, Kitajima, & Takeuchi, 2008) and reveals the spread of attention across the screen), (2) the mean fixation duration (which can reflect difficulties with extracting information (Jessee, 2010)), (3) and the mean saccade length (which can provide information on the efficiency of the search; Goldberg & Kotval, 1999). Due to calibration issues, this data is available for only 4 of the participants.

Results

For the data analysis, the screen was divided into 9 sectors of equal size (see Figure 3). The various measures were then calculated for the overall screen and for the individual sectors.

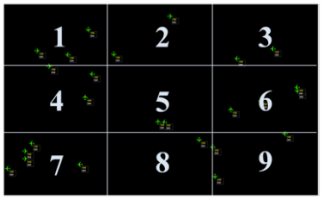


Figure 3. Division of the screen into 9 sectors

Performance Data

The average detection rate for speed and altitude deviations across all participants and conditions was 26.2%. Performance was significantly lower with high task demand, as compared to the low task demand condition (10.7% and 41.7%, respectively; $F(1,26) = 177.2$, $p<0.001$). The performance decrement in the high task demand condition was uniform across sectors (see Figure 4).

27%	20%	20%
39%	31%	31%
25%	20%	27%

Figure 4. Average detection rate for each sector as a function of task demand (expressed as the ratio 'detection rate with high task demand/detection rate with low task demand')

The overall detection rate did not differ significantly between the noise and no-noise conditions (26.4% and 26.0%, respectively). However, detection performance for individual sectors varied somewhat as a function of noise: in the presence of noise, performance for sectors 1, 2, 5, 6 and 7 slightly improved; it tended to decrease for sectors 3 and 9 and remained the same for sectors 4 and 8 (see Figure 5).

116%	118%	68%
98%	114%	139%
117%	99%	86%

Figure 5. Average detection rate for each sector as a function of noise (expressed as the ratio 'detection rate with high noise/detection rate with no noise')

Eye tracking

Number of fixations. Overall, there was a trend towards fewer fixations in case of high task demand, as compared to low task demand (319.5 and 440, respectively). However, when calculated for individual sectors, the number of fixations in central sector 5 increased by 21% with high task demand while most of the other sectors showed a slight decrease in fixations. Noise did not affect the number of fixations.

89%	90%	73%
99%	121%	108%
73%	85%	92%

Figure 6. Average mean number of fixations on each sector as a function of task demand (expressed as 'number of fixations with high task demand/number of fixations with low task demand')

Mean fixation duration. The mean fixation duration was slightly higher with high task demand, compared to the low task demand condition (1.02 seconds and 0.70 seconds, respectively). The sectors that showed the strongest increase in the mean fixation duration (110-123%) were sectors 1, 2, 3 and 5 (see Figure 7). Noise did not affect mean fixation duration.

110%	123%	115%
102%	123%	100%
96%	97%	92%

Figure 7. Average mean fixation duration on each sector as a function of task demand (expressed as 'mean fixation duration in the high task demand condition/mean fixation duration in the low task demand condition')

Mean saccade length. There was a trend towards shorter mean saccade lengths with high task demand, as compared to low task demand (65.3 pixels and 75.0 pixels, respectively), but was not affected by noise. However, there was an interaction between noise and task demand such that a decrease in mean saccade length was less pronounced for high task load in the presence of noise (see Figure 8).

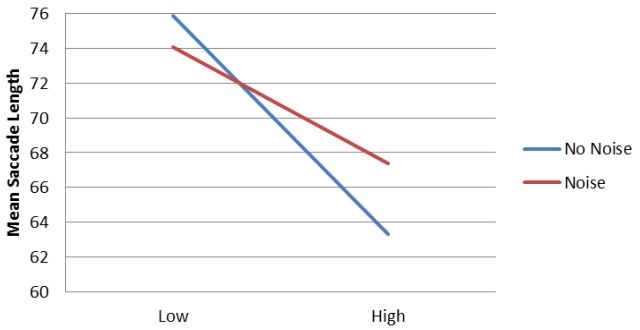
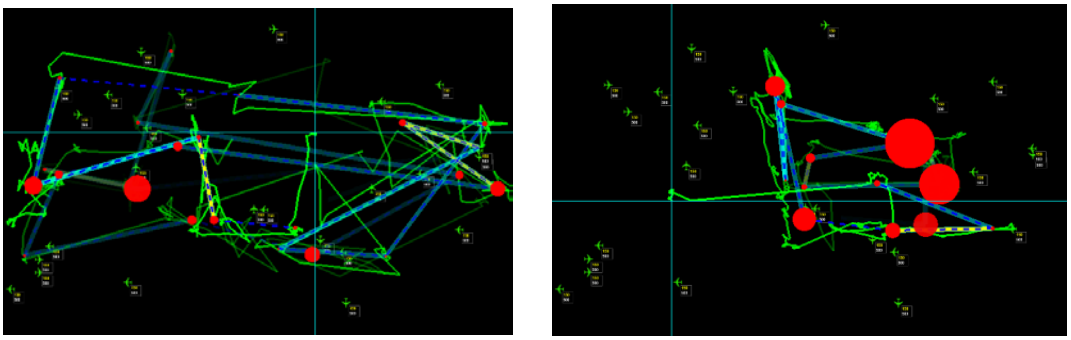


Figure 8. Interaction effects of noise and task demand on mean saccade length.

Figure 9 illustrates the above mentioned changes for one participant's scan pattern in the (a) no noise/low task demand condition and the (b) noise/high task demand conditions. In the latter case, fixations are more closely spaced and longer fixation durations are observed.



(a) (b)

Figure 9. Example of one participant's scan pattern over a period of 20 seconds in (a) the no-noise, low task demand condition and (b) the noise and high task demand condition (the size of the red circles represents fixation duration)

Discussion

The present study examined the effectiveness of two factors – loud noise and high task demand – for inducing attentional narrowing in the context of a simulated simplified ATC task. Results showed that task demand, but not noise, significantly affected both participants' performance and their attention allocation. In the high task load condition, significantly fewer airspeed and altitude deviations were detected. The eye tracking data reveal that this performance decrement resulted from a narrowing of the visual attentional field. The number of fixations in the central sector increased at the expense of more peripheral fixations, and longer fixation durations and shorter mean saccade lengths were observed. In combination, these effects resulted in a slower and more confined visual scan during high task load.

There are several possible reasons why noise may not have affected attention allocation and performance. First, loud noise has been shown to degrade performance on complex tasks (i.e. multi source tasks or tasks with a high signal rate) but it benefits the performance of simple tasks (Hockey, 1970). The task of detecting altitude and speed deviations may not have been sufficiently complex to be affected by noise. Also, noise is known to increase arousal which, in turn, is linked to performance by an inverted U-shaped curve (Yerkes & Dodson, 1908). The Yerkes Dodson law states that the highest level of performance is reached at an intermediate level of arousal. The noise level that participants experienced in this study may not have been sufficient to raise their level of arousal to the point where a decrease in performance would be observed. Finally, the fact that the alerts in the noise condition were not associated with actual threats may have reduced their effectiveness.

In conclusion, the findings from this study represent an important first step towards enabling controlled studies of attentional narrowing. High task demand is an effective manipulation for inducing the phenomenon, and

the eye tracking metrics proved useful for gaining insight into underlying attentional processes and for detecting and possibly counteracting the phenomenon in real time.

Acknowledgments

We would like to thank Joseph Phillips for building the simulator used in this study.

References

Bacon, S. J. (1974). Arousal and the range of cue utilization. *Journal of Experimental Psychology*, *102*(1), 81.

Bourne, L. E., & Yaroush, R. A. (2003). Stress and cognition: A cognitive psychological perspective. *Unpublished manuscript, NASA grant NAG2-1561*.

Bursill, A. E. (1958). The restriction of peripheral vision during exposure to hot and humid conditions. *Quarterly Journal of Experimental Psychology*, *10*(3), 113-129.

Dirkin, G. R. (1983). Cognitive tunneling: use of visual information under stress. *Perceptual and Motor Skills*, *56*(1), 191-198.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (Vol. 373). Springer Science & Business Media.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.), The Interface of Language, Vision and Action: Eye Movements and the Visual World. New York: Psychology Press.

Friedman, R. S., & Forster, J. (2010). Implicit affective cues and attentional tuning: An integrative review. Unpublished manuscript, University of Albany.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Habuchi, Y., Kitajima, M., & Takeuchi, H. (2008, March). Comparison of eye movements in searching for easy-to-find and hard-to-find information in a hierarchically organized information structure. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 131-134). ACM.

Harmon-Jones, E., Price, T. F., & Gable, P. A. (2012). The influence of affective states on cognitive broadening/narrowing: considering the importance of motivational intensity. *Social and Personality Psychology Compass*, *6*(4), 314-327.

Hockey, G. R. J. (1970). Effect of loud noise on attentional selectivity. *The Quarterly Journal of Experimental Psychology*, *22*(1), 28-36.

Jessee, M. S. (2010, September). Ocular activity as a measure of mental and visual workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 18, pp. 1350-1354). SAGE Publications.

Mackworth N H, (1965) ``Visual noise causes tunnel vision" Psychonomic Science 3 67 ^ 68

Murata, A. (2004). Foveal task complexity and visual funneling. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 135-141.

National Transportation Safety Board. (1973). *Aircraft accident report: Eastern Air Lines L-1011, N310EA, Miami, Florida, December 29, 1972*. Washington D.C: Author.

Rantanen, E. M., & Goldberg, J. H. (1999). The effect of mental workload on the visual field size and shape. *Ergonomics*, *42*(6), 816-834.

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*(1), 13-19.

Shappell, S. A., & Wiegman, D. A. (2003). *A human error analysis of general aviation controlled flight into terrain accidents occurring between 1990-1998* (No. DOT/FAA/AM-03/4). Federal Aviation Administration Oklahoma City OK Civil Aeromedical Inst.

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychological bulletin*, *137*(4), 682.

Wickens, C. D. (2005). Attentional tunneling and task management. Proceedings of the 13th International Symposium on Aviation Psychology (pp. 620–625). Oklahoma City, OK, April 18–21.

Yarbus, A. L. (1967). Eye movements and vision. New York: Plenum Press.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit- formation. Journal of comparative neurology and psychology,18(5), 459-482.

THE COGNITION OF MULTI-AIRCRAFT CONTROL (MAC): PROACTIVE INTERFERENCE AND WORKING MEMORY CAPACITY

Kelly Amaddio, Michael Miller, Ph.D. and John Elshaw, Ph.D.
Air Force Institute of Technology, Dayton, Ohio
Victor Finomore, Ph.D.
United States Air Force Academy, Colorado Springs, Colorado

As the number of U.S. Air Force missions requiring UAVs has rapidly increased without commensurate increases in manpower, systems which permit a single operator to supervise and control multiple, highly-automated aircraft are being considered. The operator of such a system may be required to monitor and respond to voice communications for multiple UAVs, each of which can have aircraft specific call signs, which may impose excessive requirements on constrained operator attention, working memory, and cognitive processing. The current research investigates the cognitive load (number of aircraft call signs) an individual can handle and explores the effect of proactive interference (PI) within this application. The results indicate a reduction in performance as the number of call signs are increased from 5 to 7 in the presence of PI. Interestingly performance with 5 call signs without PI is lower than performance with 5 call signs in the presence of PI.

The United States military is currently involved in many conflicts and activities worldwide. As these wars continue and budget pressures forces the decrease of military personnel, technology is relied upon as a force multiplier. Unmanned Aerial Vehicles (UAV) have become increasingly important in recent years as they significantly enhance the gathering of Intelligence, Surveillance and Reconnaissance (ISR) without risking bodily injury to the operators. As a result, the number of UAV sorties has increased exponentially in recent years despite the limited number of pilots available to control them. As a result, new concepts of operation are under consideration wherein a single pilot might control multiple aircraft during certain phases of flight. For example, transit operators may be employed to simultaneously pilot multiple semi-autonomous aircraft between an airbase and the battlespace. If pilots are going to be operating multiple aircraft at once, they will have to monitor and respond to a large throughput of radio communications. Additionally, there is a concern that proactive interference (PI), when previously stored information prevents the learning of new information, may occur when pilots transfer aircraft to other pilots, but still hear the previous aircraft specific radio calls. Several principles related to working memory, interference, and attention are important to the analysis of this issue. The following study is a cognitive laboratory experiment aimed at evaluating cognitive load and the effects of PI.

The ability of an operator to listen to and respond appropriately to radio traffic which contains references to the call signs of the aircraft they are controlling, as well as other entities, is likely to be constrained by their available working memory. Working memory is involved in storing and manipulating information for short-term use in tasks like reasoning and comprehension (Baddeley & Hitch, 1974). A common model of working memory that has been proposed by Baddeley (2000) contains a set of subsystems, including the central executive, which controls attention between the visuospatial sketchpad, episodic buffer, and phonological loop subsystems. The visuospatial sketchpad manipulates visual images while the phonological loop is responsible for storing and replaying words and sounds. The episodic buffer temporarily stores and integrates multimodal information and relays information between the visuospatial sketchpad and phonological loop. The auditory component of this model is important to the current study because participants are asked to listen and respond to a select series of aircraft radio calls.

Although significant research has been conducted on visual working memory, auditory working memory has garnered less attention. Considering this, Kumar et al., 2013 attempted to test auditory working memory over a continuous scale by using sequences of tones in different lengths where participants were asked to adjust a dial to replicate a specific tone that they heard. The findings indicate that increasing the number of tones held in working memory reduced the precision of the memories, much like what is found in visual working memory (Alvarez & Cavanagh, 2004).

Working memory is usually measured by span tasks that require the individual to simultaneously process and remember verbal information, usually words, letters, or numbers. The current study uses a more functional measurement of working memory by requiring the individual to remember a set of words and respond to them when they are spoken in the form of radio calls. They also have to perform this task in the presence of distracting, and

sometimes interfering information. This increases their cognitive load, which is considered a measure of the mental effort used to maintain information in working memory (Sweller, 1988), implying that working memory is limited by the amount of information it can hold and process. Miller's (1959) article provides the rule of thumb for information processing capacity: people's ability to process and remember limits them to 7± 2 items. Although the current study only requires participants to recognize call signs (instead of recalling them), the temporal complexity of the task and presence of distracting information causes us to hypothesize that individuals will be able to effectively attend to a similar number of call signs.

One of the primary functions of working memory is to navigate the effects of PI (Kane & Engle, 2000) where timely information replaces less recent information to reduce the likelihood of confusion. Therefore, effective working memory will suppress memory of outdated information to prevent it from interfering with the encoding of new information. PI has been shown to affect performance on working memory tasks. May, Hasher, and Kane (1999) found that performance on a working memory span test was improved when measures were taken to prevent PI (e.g., temporally separating trials). Kane and Engle (2000) found that individuals with low working memory spans showed greater susceptibility to PI under low cognitive load conditions, but under high cognitive load conditions, both high and low working memory span individuals showed equal levels of PI. Engle and Oransky (1999) propose that controlled attention is the mechanism by which working memory functions. They describe controlled attention as "an ability to effectively maintain stimulus, goal, or context information in an active, easily accessible state in the face of interference, to effectively inhibit goal-irrelevant stimuli or responses, or both" (Kane, Bleckley, Conway, & Engle, 2001, p.18). Neurological evidence shows that different information (sensory, semantic, etc.) is stored in different areas of the brain (Postle, 2006) suggesting that working memory should be seen as directing attention towards different memory codes stored in long term memory. Although these models of working memory have different implications for the design of interfaces to support MAC, they all support the view that the operator's attention must be divided between the visuospatial tasks necessary to control the aircraft, processing of audio call signs, and the integration of this information.

The current literature has shown that while working memory tests have been applied in numerous laboratory environments, they have not been applied to understand individual differences in real-life applications of working memory. This study will provide a more functional test of working memory by measuring participants' performance (in terms of accuracy and response time (RT)) on a multiaircraft control task in the presence of distracting information. It is predicted that higher cognitive load (created by the addition of more call signs and the presence of PI) will decrease performance.

Method

Participants

Twenty one (5 female and 16 male) volunteers with ages between 22 and 44 (M = 27.75, SD = 4.96) participated in the study. Participants were required to have a visual acuity of 20/30 or better, determined using a Logarithmic Near Visual Acuity Chart ("New ETDRS" Charts, 2011) and normal color vision, determined using isochromatic plates(Ishihara, 1980). There was no educational requirement, although most participants were graduate engineering students. Participants were recruited through e-mail. A participant number was assigned to each consenting participant's data and no personally identifiable information was retained per Institutional Review Board Protocol.

Apparatus

The experiment was conducted in a 6ft x 6ft cubicle in a quiet laboratory to minimize distractions. The experimental setup consisted of Bose AE2w headphones and a laptop to present the call signs using the Multi-modal Communication (MMC) software (Finomore, Popik, Castle, & Dallman, 2010). Participants were also given a wireless ten-digit number keypad, a clipboard containing a number grid with four rows and three columns, and a clipboard containing the list of call signs. The list of call signs was provided to the participants to remember before the experiment began and attached to the left wall to the cubicle slightly above eye level once the participants indicated their comfort with the call signs. The placement was selected to require the participant to actively turn their head to view the list.

The Multi-modal communication program (MMC) is an Air Force Research Laboratory developed multi-modal, network-centric communication management suite developed to aid Command and Control operators in increasing communication intelligibility and reduce mental workload. This tool combines several features designed to improve the performance of the users, including spatial audio, speech transcription, data capturing and playback,

chat messages, and automatic keyword highlighting (for full description of the MMC tool see Finomore et al., 2012). Additionally, this tool has been used extensively as a research tool to evaluate a variety of communication effectiveness questions (Blair, Rahill, Finomore, Satterfield, Shaw, & Funke, 2014; Finomore, et al. 2010; Finomore, Stewart, Singh, Raj, & Dallman, 2012; Finomore, Satterfield, Sitz, Castle, Funke, Shaw, & Funke, 2012; Santana, Langhals, Miller & Finomore, 2013). This experiment utilized monaural sound, a chat window to prompt the participant to enter the appropriate code, and the logging function to record the participants' inputs.

Experimental Procedure

In the design of this experiment, a few assumptions were made regarding the operational components of the UAV control task. Each aircraft was assumed to have a unique call sign and individuals having different voices made radio calls for any of the call signs (one voice was not reserved for each call sign) as is typical in current operational environments. It was also assumed that the workload level was high enough where the participants had to intentionally process the radio calls but not so high that they could not listen to all of the radio calls. Therefore, radio calls were made every five seconds. This differs from the operational environment, which would contain variable levels of workload. Additionally, there were no secondary tasks to accomplish while participants were completing the auditory task, despite the fact that the operators in an operational environment will be responsible for other tasks like navigation, communication, and aircraft monitoring. This simplification of the environment made it possible to assess the ability of the operators to perform this auditory task under near ideal circumstances.

Upon arrival, participants were randomly assigned to one of two groups based on their participant number. They were given a quick explanation of the software and task, and then given a three minute practice trial where they were responsible for three call signs. This practice trial was designed to minimize the possibility of a learning effect. Although a hearing test was not administered, participants were encouraged to set the volume of the radio calls to their comfort level during this warm-up period.

Based on their group, participants were asked to attend to either 5 or 7 call signs (out of 13 possible call signs) during each of four 8-minute experimental trials. The trials were counterbalanced to offset a potential learning effect. Participants in Group 1 were assigned five call signs for the first two trials and seven for the second two trials. Participants in Group 2 were assigned seven call signs for the first two trials and five call signs for the second two trials. Each 8-minute trial contained 100 radio calls that were evenly spaced 5 seconds apart. Approximately 50 radio calls were critical and an equal number were distracters. The participants did not know what the ratio was, however. During the second and fourth trials, 20 of the distracters were selected to induce PI as they were among the critical call signs in the previous trial. The order of the radio calls and calls signs was randomized. Table 1 presents the trials and the critical and PI call signs for participant Group 1. The scenarios will be referred to as 5-NP (5 call signs, no PI condition), 5-PI (5 call signs, PI condition), 7-NP (7 call signs, no PI condition), and 7-PI (7 call signs, PI condition).

Table 1.
Call signs experienced by the first participant group during each trial. Call signs which were employed to induce PI during Trials 2 and 4 are shown in Bold-Italics for Trials 1 and 3.

Participant Group	**Trial 1** (5-NP)	**Trial 2** (5-P)	Break (15 minutes)	**Trial 3** (7-NP)	**Trial 4** (7-P)
1	Laker	Laker	Working memory	Charlie	Charlie
	Hopper	Hopper	capacity test	Gringo	Gringo
	Arrow	Arrow	followed a break	Laker	Laker
	Charlie	Tiger		Raptor	Raptor
	Gringo	Eagle		Viking	Viking
				Arrow	Thunder
				Tiger	Cobra

The participants were instructed to listen for the commands that contained their call sign. Each radio call began with the word "Ready", which was proceeded by a call sign and a command containing a grid coordinate; for example, "Ready Charlie go to blue one now." The color indicates a column in the grid and the number represents a row in the grid. The grid location would then contain a number. For critical call signs, the participants then found the space on the grid that corresponded with the command, and typed the number from the grid location into the

MMC chat window. For example, when the participant heard "Ready Charlie go to blue one now," if the participant was responsible for "Charlie" during that trial (Charlie would be on their list of call signs), they would be expected to find the "blue 1" spot on the grid and type the two digit number in that grid location on the keypad. If the participant heard a call sign that was not on their list, they were instructed to type a zero into the chat window. Also, if for some reason they were not sure whether they were responsible for a specific call sign, they were instructed to type a zero. The randomized numbers on the grid were between 10 and 99. Participants were given as much time as they needed to memorize the call sign list before every trial and were instructed to only look at the list of call signs if they forgot them during the trial. The number of times they looked at the call sign list was recorded by the investigator for every trial.

To keep the participants from habituating to certain experimental conditions (call signs and voices), certain measures were taken. First, the list of critical call signs on the clipboard were shuffled for each trial so that they were not in the same order for sequential trials, making it harder to memorize. All trials contained different orders of radio calls, different call signs, and called for different grid locations. Additionally, a new number grid was used for each trial. Finally, a variety of voices made radio calls for every call sign so that the participant could not ignore or attend to a certain call sign based on the speaker. During the experiment, the participant could hear up to 12 different individual's voices and up to 13 different call signs.

Performance Measures

Data was collected during all trials using the logging function in MMC. After each trial, participants were asked to respond to two 5-point Likert Scale questions: one regarding their workload level (Tattersall & Foord, 1996) and the other regarding the perceived difficulty (1= very easy, 2 = easy, 3 = neutral, 4 = difficult, 5 = very difficult). After the last trial, participants were asked to self report the number of call signs they believed they could reliably monitor.

Numerical responses to the MMC task provided by the participants were evaluated for accuracy and RT. For each trial, the accuracy score was calculated by dividing the number of correct responses by the total number of radio calls and multiplying by 100%. Additionally, a PI accuracy percentage correct score was determined by adding the number of correct responses given for the PI call signs divided by the total number of radio calls expected to induce PI for 5-PI and 7-PI conditions. Finally, the average of the participant's RTs were calculated for each trial as the average of the amount of time lapse between the time when the radio call was spoken and the time the participant pressed enter after typing their numerical response. This score did not account for RTs for correct and incorrect responses.

Results

A two-factor repeated measures ANOVA revealed that there was a significant main effect of the number of call signs as well as the interaction between the number of call signs and the presence of PI on accuracy scores on the MMC task ($F(1, 19) = 7.631$, $p = 0.012$), as shown in the left panel of Figure 1. The interaction was further analyzed by applying a single factor repeated measures ANOVA. This analysis revealed that the accuracy scores were significantly different across trials ($F(2.28, 43.31) = 4.307$, $p = 0.016$, partial eta squared = 0.19). Post hoc tests using the Bonferroni correction determined that scores in the 5-PI condition ($M = 97.11\%$, $SD = 3.75\%$) were statistically higher than scores in the 5-NP condition ($M = 93.70\%$, $SD = 3.16\%$) and 7-PI conditions ($M = 91.73\%$, $SD = 6.48\%$). The scores for 5-NP, 7-NP ($M = 94.14\%$, $SD = 6.44\%$), and 7-PI were not significantly different from one another. Therefore, we can conclude that the highest scoring condition occurred when the participants were tasked with 5 call signs in the PI condition. A paired samples t-test indicated that PI accuracy scores were not significantly different between 5-PI ($M = 95.29\%$, $SD = 12.63\%$) and 7-PI ($M = 90.25\%$, $SD = 15.27\%$). Additionally, an independent samples t-test showed that accuracy scores were not significantly different based on the order the participants experienced those conditions, indicating that there was not a significant learning effect.

A two-factor repeated measures ANOVA revealed that the number of call signs had a significant effect on RT ($F(1, 17) = 11.786$, $p = 0.003$, partial eta = .409), but there was no significant effect of PI (although it approached significance at $p = .073$) or the interaction on RTs, as shown in the right panel of Figure 1. A repeated measures single factor ANOVA with a Greenhouse-Geisser correction revealed that the RTs across trials were significantly different, ($F(1.7, 28.5) = 8.520$, $p = 0.002$, partial eta = 0.334). Post hoc tests using the Bonferroni correction determined that RTs in 5-PI ($M = 3.338$ $SD = .342$) were statistically significantly lower than RTs in 7-NP ($M = 3.587$, $SD = .405$) and 7-PI ($M = 3.579$, $SD = .430$). The RT for 5-NP ($M = 3.425$, $SD = .316$) was not significantly different from the others.

Additionally, an independent samples t-test indicated that RTs were significantly different based on the order participants experienced the 5 versus 7 call sign condition ($t(76) = 3.034$, $p = .003$) where those experiencing the 5-CS conditions first had a significantly higher RT ($M = 3.601$, $SD = .376$) than those who experienced the 7-CS conditions first ($M = 3.352$, $SD = .349$).

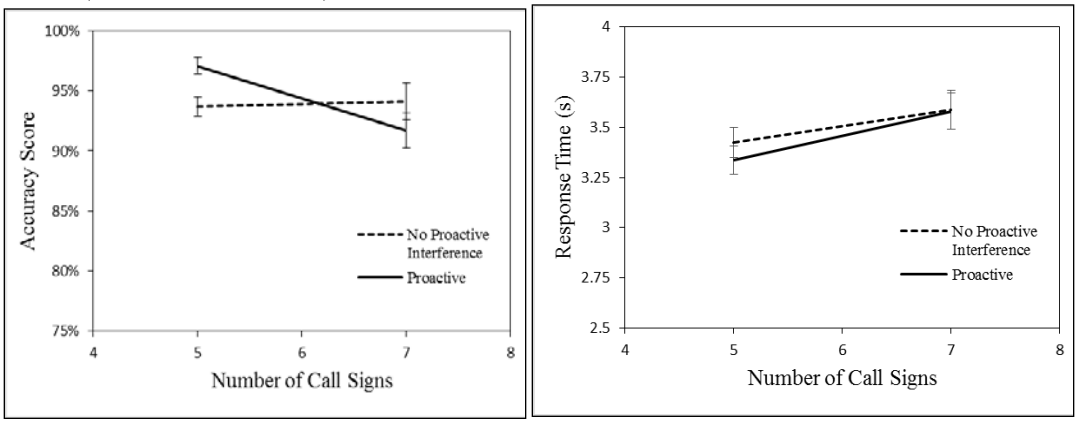


Figure 1. Interaction of number of call signs on accuracy scores for both PI conditions,(left panel) and the interaction of number of call signs on response times for both PI conditions (right panel).

A repeated measures ANOVA determined that there was no significant difference between workload or difficulty measures across all trials. Additionally, when asked "based on your experience today, how many call signs do you think you could monitor comfortably before you would begin missing time critical information?" after all experimental trials, participants responded with a mean of 5.86 ($SD = 1.35$). Responses ranged from 3 to 8 call signs.

Discussion

Overall, the results show that the participants' accuracy and response time was degraded as the number of call signs increased from 5 to 7, as expected. However, the results with respect to proactive interference differed from expected as accuracy and response time were not consistently degraded in the presence of proactive interference. Specifically, with respect to the accuracy scores, the 5 call sign PI condition was the highest scoring even though it was not the lowest taskload condition. A few possible explanations could be offered.

First, the workload-performance curve (similar to the Yerkes-Dodson Law) shows that high and low levels of workload result in low performance, but medium levels of workload result in higher performance (Teigen, 1994) creating an inverted-U shaped relationship. One potential explanation is that the workload was so low that the participants' performance did not reach its optimal level. This, however, was not supported by the reported workload and difficulty scores which did not significantly differ across the experimental conditions.

As it is necessary for the participants to be exposed to a set of call signs before these same call signs can induce proactive interference, another possible explanation stems from the need to present the PI conditions after the NP conditions. The results indicated that RT was influenced by whether the participants experience the 5 or the 7 call sign condition first, potentially indicating that the participants who experienced the 7 call sign condition first underwent a higher rate of learning than the participants who experienced the 5 call sign condition first. It is possible that negative effects of proactive interference were offset by learning effects within the current experiment.

Sampling error could have also contributed to the unexpected outcomes. For most variables, there was data from only 21 participants (due to missing data). Because of this small sample size, irregular data points could have been magnified in the results. Although the trials were kept to a short length, fatigue could have been a factor in this study, as some participants reported feelings of boredom. Additionally, there were a limited number of call signs used in this experiment, with only 13 call signs available for use in the trials. As a result, on trials where participants were supposed to remember 7 call signs, some reported that instead of listening for the call signs on the list, they listened for the ones not on the list since they believed (correctly) that there were fewer of those. Ideally, a new set of call signs would be used on each trial to prevent habituation.

Conclusion

The results of this study provide conflicting evidence about whether higher taskload conditions actually produce lower levels of performance. This study indicated that increasing the number of call signs from 5 to 7 reduced the participants' accuracy and increased their response time. However, the results do not support the hypothesis that performance will be reduced by proactive interference, a result which has multiple potential explanations including learning, workload, and sample bias effects. Further research is recommended which include additional task load levels (more call signs/PI conditions), more participants, less overlap in call signs between conditions, and potentially enhanced training. Data from this research could give insight into a relationship that exists among these variables.

Acknowledgements: The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Air Force, Department of Defense, nor the U.S. Government. The authors would like to thank Colonel Anthony Tvaryanas for motivating the current study.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15(2),* 106-111.

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in Cognitive Sciences, 4(11),* 417-423.

Baddeley, A.D. and Hitch, G. (1974) *Working memory*. In The Psychology of Learning and Motivation (Bower, G.A., ed.), pp. 48-79, Academic Press.

Blair, E. A., Rahill, K. M., Finomore, V., Satterfield, K., Shaw, T., & Funke, G. (2014, September). Best of Both Worlds Evaluation of Multi-Modal Communication Management Suite. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 58, No. 1, pp. 410-414). SAGE Publications.

Engle, R. W., & Oransky, N. (1999). Multi-store versus dynamic models of temporary storage in memory. *The Nature of Cognition*, 515-555.

Finomore, V.S., Popik, D.K., Castle, C.E., & Dallman, R.C. (2010). Effects of network-centric multi-modal communication on a communication-monitoring task. *Proceedings of the Human Factors and Ergonomics Society, 54,* 2125- 2129.

Finomore, V., Stewart, J., Singh, R., Raj, B., & Dallman, R. (2012). Demonstration of advanced multi-modal, network-centric communication management suite. *Proceedings of the Interspeech, 13.*

Finomore, V., Satterfield, K., Sitz, A., Castle, C., Funke, G., Shaw, T., & Funke, M. (2012). Effects of the multi-modal communication tool on communication and change detection for command and control operators. *Proceedings of the Human Factors and Ergonomics Society, 56, 1464-1465*.

Ishihara, S. (1980). *Ishihara's design charts for colour-blindness of unlettered persons*. Kanehara & Company.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130(2),* 169.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(2),* 336.

Kumar, S., Joseph, S., Pearson, B., Teki, S., Fox, Z. V., Griffiths, T. D., & Husain, M. (2013). Resource allocation and prioritization in auditory working memory. *Cognitive Neuroscience, 4(1),* 12-20.

Logarithmic Near Visual Acuity Chart 2000 "New ETDRS" Charts (2011). Precision Vision.

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition, 27(5),* 759-767.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63(2),* 81.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139(1),* 23-38.

Santana, L., Langhals, B., Miller, M., & Finomore, V. (2013). Does supplementary computer generated cueing enhance controller efficiency in a congested communication environment? *Proceedings of the International Symposium on Aviation Psychology, 17,* 226-231.

Sweller, J (1988). "Cognitive load during problem solving: Effects on learning". *Cognitive Science 12(2)*: 257–285.

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740-748.

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. Theory & Psychology, 4(4), 525-547.

VISUAL SEARCH AND TARGET SELECTION USING A BOUNDED OPTIMAL MODEL OF STATE ESTIMATION & CONTROL

Brandon S. Perelman
Michigan Technological University
Houghton, MI
Christopher W. Myers
Air Force Research Laboratory
Wright-Patterson Air Force Base, OH

Visual attention and motor control are tightly coupled in domains requiring a human operator to interact with a visual interface. Here, we integrate a boundedly optimal visual attention model with two separate motor control models and compare the predictions made by these models against perceptual and motor data collected from human subjects engaged in a parafoveal detection task. The results indicate that humans use an optimal motor control policy limited by precision constraints – humans executed ballistic movements using near-optimal velocity (i.e., bang-bang control), but imprecision in those movements often caused participants to overshoot their targets, necessitating corrective action. Motor movements did not reflect response hedging, but rather a perceptual-motor policy permitting ballistic movements to a target only after localization confidence exceeded a threshold. We conclude that a boundedly-optimal perceptual-motor model can predict aspects of human performance visual search tasks requiring motor response.

Introduction

Visual search is conducted in nearly everything we do, from the mundane to dangerous operational domains. For example, while shopping online we search crowded visual displays to find the correct item. Similarly, sensor and radar operators must discriminate targets from noise and foils. These tasks require that an agent identify the target through visual search, and then select that target. In this study, we integrated a boundedly-optimal state estimation model of visual search with models of motor control, and tested them against human performance in a parafoveal detection task (PDT). The long-term goal of this line of research is a model capable of generating performance ceiling predictions and automatic interface evaluations. In the following sections we introduce visual search and oculomotor control, followed by manual motor control.

Visual Search and Eye Movements

Efforts to model visual search fall largely into two categories (see Kowler, 2011 for a comprehensive review). *Map-based* approaches (e.g., Itti, 2006; Itti & Koch, 2000; Pomplun, 2003; Wolfe, 2007) use bottom-up processing to subdivide raw images into saliency or activation maps. These models predict that the agent will produce a saccade to the most salient or active areas of a map derived from bottom-up processes over feature and spatial information (e.g., color, rotation, distance, etc.). Conversely, *visibility models* (Geisler, 2011; Myers, Gray, & Sims, 2011; Myers, Lewis, & Howes, 2013; Najemnik & Geisler, 2008; Baron & Kleinman, 1969) are top-down models of visual attention, predicting that eye movements are made in the service of maximizing information gain. Such models, often labeled *ideal observers*, have demonstrated much success in accounting for human saccades during search. In the present study, we use a visibility model to derive predictions in the PDT. Our visibility model optimally estimates the state of a presented stimulus given known bounds of the human visual system. We refer to this model as boundedly-optimal, and combine boundedly-optimal state estimation with near optimal oculomotor control (i.e., the near-optimal saccadic selectivity, given a boundedly-optimal estimation of the state of the display).

Manual Motor Control

Evidence in the motor literature suggests that motor control reflects a dynamic decision-making process (Freeman, Dale, & Farmer, 2011; Wolpert & Landy, 2012) whereby participants' motor trajectories reflect cognitive phenomena, such as confidence. These effects have been investigated in the Iowa gambling task (Koop & Johnson, 2011), memory tasks (Papesh & Goldinger, 2012), and item selection from an interface (Bailly, Oulasvirta, Brumby, & Howes, 2014). In ambiguous situations where the operator must distinguish among multiple potential targets, high probability locations often attract the operator's cursor even when they are not ultimately selected (Farmer, Cargill,

& Spivey, 2007). Therefore, motor control does not involve merely converting visual information to motor coordinates, as saccades occurring during a motor movement can cause immediate changes in destination and trajectory (Thompson, Byrne, & Henriques, 2014). It is important to distinguish among input devices, as motor control trajectories derived using a mouse (e.g., Bailly et al., 2014; Koop & Johnson, 2011; Papesh & Goldinger, 2012) differ from those generated using touchscreen inputs (e.g., Parhi, Karlson, & Bederson, 2006).

To model participants' velocity profiles, we used two theories of motor control (Kelso, 1982) – an open-loop theory (*bang-bang* control, which produces the optimal acceleration and deceleration between starting and target cursor positions assuming equal rates of acceleration and deceleration), and a closed loop theory (velocity proportional to distance, *vProp*, which begins at maximum acceleration and reduces its speed proportionally with feedback based on the decreasing distance to the target). In the integrated visuomotor model, we represent the decision making process' influence on motor trajectories using closed-loop control that permits the model to dynamically update the cursor destination to the current highest probability target location. Finally, to model the effect of motor control parameters on response times, we implemented a crude version of Fitts' law (Fitts, 1954) whereby the model's ballistic cursor movements were perturbed by noise that scaled with that movement's distance.

Experiment

All participants completed a PDT that required target detection and localization within a pair of items on opposing sides of initial fixation crosshairs (see Figure 1). Participants were instructed to fixate on crosshairs located at the center of the screen (see Stage 1, Figure 1), and on fixation, to click the mouse to initiate stimulus onset. A stimulus consisted of red and green X and O characters in four locations on opposite sides of the crosshairs, but on the same plane (Stage 2, Figure 1). On stimulus onset, the mouse was positioned 334 pixels below the crosshairs, and participants were instructed to respond by clicking on the perceived target location. If the target was not detected, participants clicked on the fixation crosshair (target-absent; Stage 3, Figure 1).

The PDT was programmed using the Psychology Experiment Building Language (Mueller, 2014). The stimulus was presented on a screen (1024 x 768 pixels resolution) viewed from a distance of 20 inches Pairs were separated by 1° of visual angle. Pairs varied in eccentricity from the crosshairs by 8°, 12°, 16°, and 20°. Pairs on each side of the crosshairs were always separated by the same eccentricity. .

Of the 189 ecologically possible displays, 32 were selected that maximized prediction differences between map models and those from our boundedly-optimal state estimation model. Half of these displays contained targets and half did not. Participants completed 640 trials, broken into five blocks of 128 trials. Each block was a randomized set of all 32 displays at each of the four eccentricities.

Modeling

The model was previously fit to a version of the PDT that required no mouse responses, only target-present or target-absent response through keypresses. Best parameter fits were determined by investigating a space of 3,200,000 parameter combinations using www.mindmodeling.org (Harris, Gluck, Mielke, & Moore, 2009). Results indicated best-fitting parameters: saccade threshold = 0.21, response threshold = 0.84, spatial noise = 10, and feature noise = 6. For a detailed description of the boundedly-optimal state estimation model, see Myers et al. (2013). These same parameter values were then used in modeling a version of the PDT requiring point-and-click responses.

To issue point-and-click responses, the model executed motor movements from the starting mouse location to its current target (i.e., highest probability location once the motor threshold is reached. The cursor's destination was perturbed as above, then updated at the model's sampling rate of 25 ms, while unperturbed (direct) trajectories represent the optimal motor response against which to compare human data. Motor control parameters included maximum acceleration and deceleration (2 pixels per sample; though the proper setting of this parameter could be determined empirically using Fitts' Law) and a motor movement initiation threshold (probability of target present or absent >= 0.51). This threshold parameter permitted the model to initiate "early" motor movements, guiding the cursor during stimulus presentation before the model had committed to a particular target location. The cursor update loop ends when the cursor is within the target location's clickable field, at which point the response time is appended with a manual response time, intended to simulate a mouse click, drawn from a gamma distribution (shape = 11.11, scale = 9), which produced a mean manual response time of roughly 100 ms (M = 99.93, SD = 29.77).

Hypotheses and Model Predictions

Using the aforementioned parameters, we ran the model on 25 trials for each of the 128 display combinations, for a total of 3,200 runs, to produce a dataset against which to compare the human subjects' performance. Model velocity profiles using both the bang-bang and vProp algorithms provide a baseline against which to compare participants' performance. Straight paths between the cursor starting and target locations provide optimal motor trajectories against which to compare the human data. Finally, a crude implementation of Fitts' law (see above) permitted the model to generate motor response times for comparison. Because the visual attention model applies feature and spatial noise that increases with distance from the point of fixation, it predicts an effect of eccentricity on all aspects of task performance. Therefore, we expect that,

1. Motor Velocity: We expect that humans will exhibit one of the experimental (bang-bang or vProp) motor velocity profiles given there is evidence for each in the literature, However, because of the static nature of the task and only a required straight movement to reach a location, then bang-bang is a better candidate.
2. Motor Trajectories: Dynamic decision making theory suggests that error in participants' motor response trajectories should increase with increasing difficulty, therefore for humans we expect increased response times and increased motor trajectory error (greater divergence from the optimal trajectory), measured by pathmapping) with increasing eccentricity. Furthermore, we expect less motor control error on trials where participants respond correctly.
3. Response Initiation: The model builds evidence toward a decision more slowly as task difficulty increases, therefore the proportion of trials on which the agent initiates an *early motor movement* > 25 pixels from cursor starting location *during* the 500 ms stimulus presentation window should similarly decrease.
4. Response Accuracy: The model predicts that target identification (distinguishing target present from target absent displays) and localization (determining the specific location of the target) performance should decrease with increasing eccentricity

Data Analysis

Motor data were subdivided into two components – trajectory and velocity. To analyze the velocity profiles, we split human and model trajectories into two halves, then divided the average velocity in the second half of the trajectory by the average velocity in the first, creating a *split-half velocity ratio*. We expected that participants would exhibit one of three velocity profiles, each distinguishable by the velocity ratio. Because the bang-bang algorithm uses maximum acceleration and deceleration before and after a halfway point, its velocity profile predicts a split-half velocity ratio around one (i.e., the average velocity in both trajectory halves are equal). Conversely, vProp predicts a split-half velocity ratio of either greater or less than one, depending upon whether the cursor accelerates or decelerates as it approaches its destination.

To analyze error in the human data, we used the 'pathmapping' package (Mueller & Perelman, 2013) built for the R statistical computing language. This package creates a polygon from two arbitrary paths (i.e., the empirical trajectory, and the optimal trajectory, a straight line from starting position to that target), the area of which is the error in pixels, and holds an advantage over traditional measures of motor error (Koop & Johnson, 2011).

Results

Hypothesis 1: Velocity Profiles and Optimal Control

Across all eccentricities, and for all target locations, the model produced mean split-half velocity ratios of 1.03 (SD = 0.19) using the bang-bang algorithm, and 6.68 (SD = 1.82) with vProp. Participants' trajectories consisted of an initial ballistic trajectory toward the target location, which often carried the cursor past the target location, followed by a corrective trajectory, which brought the cursor back to the target. The ballistic trajectory was operationalized as all sampled trajectory points to the cursor's farthest distance from the target location, with corrective trajectory accounting for remaining points. Split-half velocity ratios indicated that the ballistic trajectory was very similar to bang-bang style movement (M = 0.99, SD = 0.08), whereas the corrective trajectory functioned similarly to a vProp control method (M = 0.85, SD = 0.35), owing largely to the requirement for the participant to change direction of travel nearly 180 degrees back to the target location, and the relatively short travel time. Participants tended to overshoot more distant targets by a margin that increased with eccentricity, $F(3, 3928)$ =

23.70, $p < .001$ (see Table 1), an effect which holds implications for modeling motor performance in higher fidelity. These results indicate that humans exhibit optimal bang-bang control in this task.

Hypothesis 2: Motor Trajectories and Optimal Control

Human motor control error (i.e., divergence from optimal measured via pathmapping) increased with eccentricity, $F(3, 5115) = 74.94$, $p < .001$. Trajectories produced during correct responses were roughly twice as close to optimal as incorrect response trajectories, $t(560.46) = 9.20$, $p < .001$. One potential criticism of this approach is that longer trajectories leave more room for potentially producing error, due to the cursor travel distance. To address this problem, we scaled the error values at each eccentricity by the cursor's distance from the target location. Using these scaled error terms, effects of eccentricity, $F(3, 5115) = 16.7$, $p < .001$, and accuracy, $t(562.49) = 8.72$, $p < .001$, persisted even when controlling for cursor distance of travel.

Hypothesis 3: Motor Movements and Dynamic Decision Making

Humans and the model produced the expected effect of eccentricity on response time during target trials, however only humans exhibited this effect when the target was not present (see Figure 3). Humans were faster by 185 and 198 ms on the target-absent and target-present trials, respectively. The model, given a motor threshold of 0.51, and participants produced fewer early motor movements with increasing eccentricity (see Table 2).

Participants exhibited mean motor velocities that varied with signal detection and correctness. Hits ($M = 9.19$, $SD = 3.02$) and correct rejections ($M = 8.85$, $SD = 3.37$) produced faster mean motor velocities than misses ($M = 8.09$, $SD = 3.37$) or false alarms ($M = 7.69$, $SD = 4.42$). A 2 (Correct vs. Incorrect) x 2 (Trajectory: Ballistic vs. Corrective) factorial ANOVA revealed a significant interaction effect, $F(1, 9037) = 85.98$, $p < .001$, whereby correct answers produced faster ballistic trajectories ($M = 12.68$, $SD = 6.93$) than incorrect answers ($M = 11.08$, $SD = 7.69$), but slower corrective trajectories ($M = 1.55$, $SD = 2.54$) than incorrect answers ($M = 3.47$, $SD = 4.56$). These results indicate that confident response selection produces more precise and expedient motor movements.

Hypothesis 4: Target Identification and Localization

Both the model, $F(3, 3196) = 5.56$, $p < .001$, and participants, $F(3, 5116) = 38.81$, $p < .001$, exhibited target identification performance that degraded with increasing eccentricity. In addition, localization degraded with increasing eccentricity in participants, $F(3, 2556) = 27.57$, $p < .001$, and the model, $F(3, 3196) = 10.36$, $p < .001$ (see Figure 2). To further evaluate model and participant accuracy in target identification, we applied signal detection theory (see Table 2). Humans' target discriminability (D') degraded with increasing eccentricity. The model demonstrated a similar trend, with the exception that the model's D' at 12 degrees was higher than at 8 degrees of eccentricity. Given the aforementioned effect of eccentricity on response accuracy, this difference lies in the false alarm rate and may reflect a tradeoff whereby the model adopts a more conservative strategy than humans.

Conclusions and Future Directions

The human subjects and modeling results, taken together, indicate that a bounded optimal state estimation model of visual attention, coupled with bang-bang motor control, produces similar effects to those seen in humans in the PDT. Specifically, the model predicts decreasing response confidence, and performance, as measured using identification and localization accuracy and response time, with increasing eccentricity from the point of fixation.

One proximal goal of future research is to further validate the motor control system. Analyzing model overshoots and trajectory divergence would permit direct comparison with the human data. Furthermore, it is not currently clear as to whether the differences in RT between humans and the model were due to differences in motor speed or accuracy, and trial-by-trial analysis should elucidate this in future research.

In service of the project goal of predicting human performance, a more accurate implementation of Fitts' Law would provide a more realistic account of the motor data, and should also impact response time distributions. Finally, testing the model against human data in a more complicated task, such as a computer interface, would speak to external validity of this approach for predicting performance in naturalistic tasks.

Tables and Figures

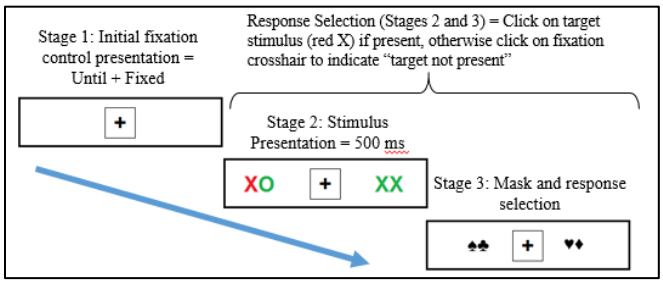


Figure 1. Parafoveal detection task time course and instructions.

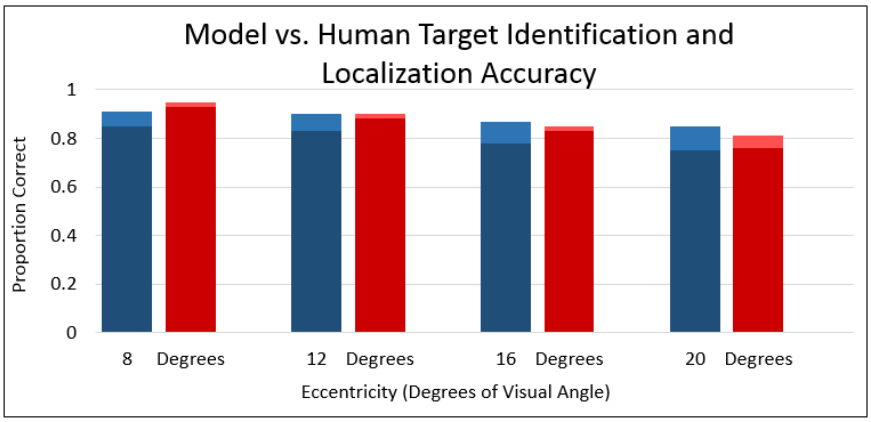


Figure 2. Model and human performance in identifying and localizing targets across all experimental eccentricities in target present trials. Blue and red bars correspond to model and human performance, respectively. Bars indicate the proportion of trials in which the agent successfully identified the target, while darkened portions indicate the proportion of those trials in which they also successfully localized the target.

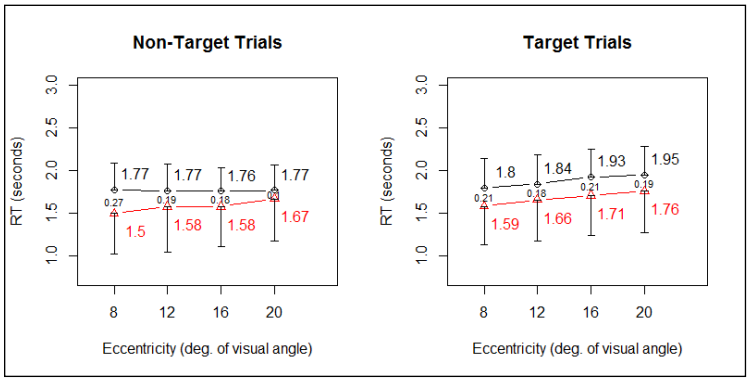


Figure 3. Model (black circle) and human (red triangle) response times in target and non-target trials by inner eccentricity. Difference scores are shown between the two data series. Errors bars reflect standard deviation.

Table 1.
Corrective trajectory split-half velocity ratios and target overshoot by eccentricity.

Eccentricity	Split-Half Velocity Ratio	Target Overshoot (Pixels)
8 Degrees	0.83	39
12 Degrees	0.85	52
16 Degrees	0.86	61
20 Degrees	0.87	73

Table 2.
Proportion of trials where the agent initiated an early motor movement (left), and D' calculations (right) at each experimental eccentricity.

	Early Motor Movements		D' Calculations	
Eccentricity	**Humans**	**Model**	**Humans**	**Model**
8 Degrees	0.17	0.19	3.28	2.40
12 Degrees	0.11	0.15	2.95	2.49
16 Degrees	0.08	0.12	2.62	1.92
20 Degrees	0.08	0.12	1.94	1.61

References

Bailly, G., Oulasvirta, A., Brumby, D. P., & Howes, A. (2014). Model of visual search and selection time in linear menus. In *Proc. of the ACM CHI Conference on Human Factors in Computing Systems*, Toronto, Canada.

Baron S. & Kleinman, D. L. (1969). The human as an optimal controller and information processor. *IEEE Transactions on Man-Machine Systems, 10,* 10-17.

Farmer, T. A., Cargill, S. A., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *J. Mem. Lang., 57,* 570-595.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement". *Journal o Experimental Psychology, 47,* 381–391.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2,* 1-6.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51,* 771-781.

Harris, J., Gluck, K., Mielke, T., & Moore, L. R. (2009). Mindmodeling@Home...and anywhere else you have idle processors. In *Proceedings of the Ninth International Conference on Cognitive Modeling*, Manchester, UK.

Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition, 14,* 959-984.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489-1506.

Kelso, J. S. (Ed.). (2014). *Human motor behavior: An introduction*. Psychology Press.

Koop, G. J. & Johnson, J. G. (2011). Response dynamics: A new window on the decision process. *Judgment and Decision Making, 6,* 750-758.

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research, 51,* 1457-1483.

Mueller, S. T. (2014). PEBL: The Psychology experiment building language (Version 0.14) [Computer experiment programming language]. Retrieved June 2014 from http://pebl.sourceforge.net.

Mueller, S. T. & Perelman, B. S. (2013). Pathmapping: Software for Determining the Divergence and Mapping Between Two Paths. R package version 1.0.

Myers, C. W., Gray, W. D., & Sims, C. R. (2011). The insistence of vision: Why do people look at a salient stimulus when it signals target absence? *Visual Cognition, 19,* 1122-1157.

Myers, C. W., Lewis, R. L., & Howes, A. (2013). Bounded optimal state estimation and control in visual search: Explaining distractor ratio effects. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, Berlin, Germany.

Najemnik, J. & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision, 8,* 4.1-14.

Papesh, M. H. & Goldinger, S. D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychon. Bull. Rev., 19,* 906-913.

Parhi, P., Karlson, A. K., & Bederson, B. B. (2006). Target size study for one-handed thumb use on small touchscreen devices. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, Espoo, Finland.

Pomplun, M. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science, 27,* 299-312.

Thompson, A. A., Byrne, P. A., & Henriques, D. Y. P. (2014). Visual targets aren't irreversibly converted to motor coordinates: Eye-centered updating of visuospatial memory in online reach control. *PLoS One, 9.*

Wolfe, J. M. (2007). Guide Search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford University Press.

Wolpert, D. M. & Landy, M. S. (2012). Motor control is decision-making. *Current Opinions in Neurobiology, 22,* 996-1003.

ANTICIPATORILY CONTROLLED TOP-DOWN PROCESSES INFLUENCE THE IMPACT OF CORIOLIS EFFECTS

Christine M. Talker
University of Graz
Graz, Austria
K. Wolfgang Kallus
University of Graz
Graz, Austria

The impact of the vestibular-induced Coriolis illusion becomes apparent in spatial disorientation and symptoms of motion sickness. Empirical data indicated that anticipatory processes, evolved by experience, influence the sensation of Coriolis illusion. We measured subjective well-being and stress responses of 13 experienced pilots and 13 non-pilots in order to study the influence of anticipatorily controlled top-down attention on the impact of Coriolis effects and to examine the role of experience. Subjective data and psychophysiological data (EDA, ECG) were recorded, reflecting the underlying psychological processes involved. Participants distracted by doing a reaction test (experimental group) gave higher drowsiness ratings and higher dizziness ratings than non-distracted participants (control group) immediately after the Coriolis induction, independently of experience. EDA data showed higher emotional stress responses in the experimental group throughout the psychophysiological sensation unit of 4x10s. Data suggest that anticipatorily controlled top-down processes are of particular importance in Coriolis-provoking environments.

Coriolis illusion is known for its incapacitating effects on a pilot's spatial orientation and/or physical well-being (e.g. dizziness, drowsiness), and hence, pose high safety risks in aviation (Gibb, Ercoline, & Scharff, 2011). Empirical data revealed that anticipatory top-down processes, evolved by experience, attenuate the impact of the Coriolis illusion (Talker, Kallus, Schwandtner, Joachimbauer, & Beykirch, 2014) and can improve a pilot's performance in disorientation-prone flight situations (Koglbauer, Kallus, Braunstingl, & Boucsein, 2011; Tropper, Kallus, & Boucsein, 2009). Gresty, Golding, Le, & Nightingale (2008) emphasized that attentional processes are of vital importance to regain orientation when spatial orientation is threatened. As top-down attention is mostly controlled anticipatorily (Butz & Pezzulo, 2008), the question arises whether the distraction of anticipatorily controlled attention influences the impact of the Coriolis illusion.

In flight, a pilot's awareness of her/his position and the attitude of the aircraft in relation to the gravitational vertical is pivotal for flight safety. A pilot's spatial orientation is threatened by different kinds of sensory illusions. One of the most dangerous vestibular-induced illusions is the Coriolis illusion (Cheung, 2013). Coriolis illusion can emerge from special flight maneuvers, as well as during prolonged turns when the pilot moves the head out of the axis of rotation. Moreover, the Coriolis illusion can provoke symptoms of motion sickness (MS). These evoked effects can severely impair the performance of those affected (Benson, 2002). The most influential theory explaining the occurrence of motion sickness is the Sensory Rearrangement

Theory (Reason & Brand, 1975). The authors emphasized the particular importance of experience by postulating a mismatch of the perceived sensory information with what is expected from previous experience. This proposed expectation process may arise from a mental model which may be generated and "updated" by a continuous anticipation-action-comparison learning process (Hoffmann, 1993; Kallus, 2012). This ongoing match-mismatch comparison of actually sensed multi-sensory information (visual, vestibular and proprioceptive cues) and anticipated multi-sensory patterns memorized from previous experience might lead to the formation of correct anticipations of upcoming (flight) situations. Distracting top-down attention from ongoing (flight) situations might influence anticipatory processes, and hence, the sensation of Coriolis effects. In order to shed light on this issue, we investigated the impact of Coriolis effects in dependence of distraction and examined the role of experience. Well-being and stress responses were investigated by collecting subjective data (ratings, questionnaires and reconstruction interviews) and psychophysiological data (EDA and ECG).

Method

Participants

13 active pilots and 13 non-pilots participated in the study, including two females, each. Pilots were between 20 and 55 years old (M = 41.08, SD = 10.40); non-pilots between 19 and 65 years, M = 34.46, SD = 12.69. The difference, $T(1,24) = 1.453$, $p = .159$ (*n.s.*), did not reach significance. Among the pilots, there were VFR pilots and IFR pilots. The pilots' flight experience ranged from 60 to 2.000 flight hrs; their experience with flight simulators ranged from "no experience" to 50 hrs. The participants in the sample of non-pilots were required not to have any experience in operating an aircraft and to have no or low experience with flight simulators. All participants took part in the experiment voluntarily. They signed an informed consent and were informed that they could quit the experiment whenever they wished, without giving any reasons. Each participant received an expense allowance of 75 Euro at the end of the experiment.

Design and Procedure

Participants were assigned to two groups according to their experience in flight motion (pilots vs. non-pilots). The main part of the experiment was comprised of the evaluation of Coriolis sensations. In the scenario of interest, Coriolis illusion was induced passively, i.e. solely by motion of the simulator cabin. The scenario consisted of a pitch-up motion of the cabin in CAVOK (*Ceiling And Visibility OKay*) weather conditions. The simulator cabin constantly rotated clockwise. The condition of the experimental group included an imperative stimulus indicating the beginning of the testing phase followed by a reaction test. Participants had to push a button once they had heard a particular sound sequence via the headset. The end of the testing phase was indicated by a second stimulus after the Coriolis induction. There was no sound sequence during the Coriolis induction. The control group received two control stimuli indicating the beginning and the end of the testing phase. The investigation took place at the premises of AMST-Systemtechnik GmbH in Ranshofen, Austria, using the AMST motion flight simulator AIRFOX®ASD. The simulator session of the passive maneuvers required approximately 25 minutes in total. At the beginning, participants received detailed information about the

procedure. Immediately after each maneuver, participants evaluated their subjective well-being via headset. After the simulator session, a reconstruction interview was conducted in order to figure out special aspects of Coriolis sensations.

Dependent Variables

Since the main purpose of this paper is to report the data of the passive scenario with distraction, only the key dependent variables of the multilevel approach will be mentioned. As a key symptom of spatial disorientation, participants evaluated the degree of dizziness on a scale from "0" (no dizziness) to "20" (extremely strong dizziness) (adapted from Keshavarz & Hecht, 2011) and drowsiness as a key symptom of MS ("0" = no drowsiness to "20" = extremely strong drowsiness) after each Coriolis maneuver. After the simulator session, participants were interviewed in regard to their sensations and mental pictures during the Coriolis maneuvers by using a post-task reconstruction interview.

During the entire simulator sequence, electrodermal activity (EDA) was recorded with the Varioport Biosignalrecorder (Becker Meditec, Karlsruhe, 2005). The received signal was monitored on an additional Laptop screen using the software Variograf Win32: Rev. 4.76 © G. Mutz 1988 – 2005 (Dipl.-Ing. Becker Meditec; 2005). Baseline measurements of 60 seconds were collected. The recording of the EDA was done using two active (0.5 Volt) non-polarised silver/silver chloride electrodes with a diameter of 22 mm (1 cm^2 measurement area). Signals were recorded from the plantar recording sites of the non-dominant foot as described by Boucsein (1992). Before application, the electrodes were filled with 0.5% non-ionising NaCl paste. The resulting conductance was measured with a resolution of 0.002 μS. The parameters SCL (skin conductance level) and NS.SCRfreq (frequency of non-specific skin conductance responses) were evaluated using the program EDA-Vario, Version 1.94 (Schaefer, 2009).

Statistical Analyses

Statistical evaluation was performed with SPSS 22.0. Subjective data were analyzed using the procedure of a two-factorial ANOVA. Psychophysiological data were evaluated by means of a multivariate analysis of variance (MANOVA) for repeated measurements. A significance level of $\alpha \leq .05$ was adopted for the statistical tests. The assumption of normal distribution was checked by means of the Kolmogorov-Smirnov Test, the premise of variance homogeneity was evaluated by means of Levene Test, and the sphericity assumption was evaluated by means of the Mauchly's Test. Repeated measures effects were based for all variables on the Huynh-Feldt Tests, using corrected degrees of freedom for countering the exceptions from the homogeneity assumption. Due to the explorative character, no correction for type-I-error was conducted. For the EDA parameters (SCL, NS.SCRfreq), baseline corrections were computed. Statistical analyses of EDA parameters were based on 10-second intervals of analyses where time intervals of 4x10 seconds were combined to a psychophysiological unit.

Results

The goal of this experimental scenario was to investigate the impact of Coriolis effects in dependence of distraction and to examine the influence of experience. A two-factorial ANOVA

was conducted with *Distraction* and *Experience* as independent variables and *Drowsiness* as dependent variable. The results revealed a significant between-subject main effect for *Distraction*, $F(1, 20) = 4.992$, $p = .037$, $\eta_p^2 = .200$. Participants who received the imperative stimulus (and did the reaction test) gave higher drowsiness ratings as compared to the control group, immediately after the Coriolis induction. The between-subject main effect for *Experience*, $F(1, 20) = .186$, $p = .671$, $\eta_p^2 = .009$, did not reach statistical significance. There was no significant interactive effect between *Distraction* and *Experience*, $F(1, 20) = .282$, $p = .271$, $\eta_p^2 = .060$. Results of the dizziness ratings revealed higher dizziness ratings of the experimental group as compared to the control group. The between-subject main effect for *Distraction*, $F(1, 20) = 2.522$, $p = .128$, $\eta_p^2 = .112$, and the between-subject main effect for *Experience*, $F(1, 20) = 1.282$, $p = .271$, $\eta_p^2 = .060$, did not reach statistical significance. There was no significant interactive effect between *Distraction* and *Experience*, $F(1, 20) = .634$, $p = .435$, $\eta_p^2 = .031$. However, it has to be noted that, after the simulator session, experienced pilots reported significantly less physical discomfort (e.g. nausea, vertigo), $T = -2.06$, $p = .028$ *(1-tailed sig.),* due to Coriolis induction as compared to non-pilots (Talker et al., 2014).

The baseline-corrected EDA parameters (mean NS.SCRfreq, mean SCL) were analyzed in time intervals of 10 seconds with the four different time intervals as levels of the within-subject factor *Psychophysiological Unit* (Reference, Anticipation/Reaction Test, Coriolis Sensation, Post-Coriolis Sensation), and the two categories of distraction of attention (imperative stimulus vs. control stimulus) as levels of the between-subject factor *Distraction*. A repeated measures multivariate analysis of variance (MANOVA) was conducted with *Psychophysiological Unit* and *Distraction* as independent variables and NS.SCRfreq and SCL as dependent variables. The results revealed a non-significant between-subject main effect for *Distraction*, *Wilks'* $\lambda = .788$, $F(2, 19) = 2.556$, $p = .104$, $\eta_p^2 = .212$, and a highly significant within-subject main effect for *Psychophysiological Unit*, *Wilks'* $\lambda = .157$, $F(6, 15) = 13.392$, $p < .001$, $\eta_p^2 = .843$. There was no significant interactive effect between *Distraction* and *Psychophysiological Unit*, *Wilks'* $\lambda = .722$, $F(6, 15) = .962$, $p = .482$, $\eta_p^2 = .278$.

Based on the responses of the experimental group in the reconstruction interview, psychophysiological data were analyzed in dependence of allocation of attention. A repeated measures multivariate analysis of variance (MANOVA) was conducted with *Psychophysiological Unit* and *Allocation of Attention* as independent variables and NS.SCRfreq and SCL as dependent variables. Results revealed a significant between-subject main effect for *Allocation of Attention*, *Wilks'* $\lambda = .408$, $F(2, 9) = 6.537$, $p = .018$, $\eta_p^2 = .592$. Participants who allocated their attention to the ongoing flight scenario showed less electrodermal responses as compared to participants who allocated their attention to the distracting stimulus. There was a significant within-subject main effect for *Psychophysiological Unit*, *Wilks'* $\lambda = .130$, $F(6, 5) = 5.584$, $p = .039$, $\eta_p^2 = .870$. The interactive effect between *Distraction* and *Psychophysiological Unit*, *Wilks'* $\lambda = .743$, $F(6, 5) = .288$, $p = .919$, $\eta_p^2 = .257$, did not reach statistical significance.

To sum up, participants distracted with the imperative stimulus (experimental group) reported significantly higher drowsiness ratings and higher dizziness ratings immediately after the Coriolis induction than participants distracted with the control stimulus (control group), independently of experience. Psychophysiological data (NS.SCRfreq, SCL) recorded during the simulator session revealed a higher electrodermal activity of the experimental group throughout a

Psychophysiological Unit (i.e. before, during and after the Coriolis induction) as compared to the control group. The analyses of electrodermal responses of the experimental group in dependence of *Allocation of Attention* (scenario vs stimulus) showed significantly less mismatch responses in participants who mainly allocated their attention to the scenario before, during and after the Coriolis induction.

Discussion and Conclusions

In this experiment, we shed light on the role of anticipatorily controlled top-down processes on the sensation of Coriolis illusion and examined the influence of experience. Subjective data revealed that distracting top-down attention by a reaction test led to a higher impact of Coriolis effects, independently of experience. These results extend the findings of Talker et al. (2014) that revealed less impairment of experienced pilots' subjective well-being after a Coriolis session in the flight simulator as compared to non-pilots. While the results of Talker et al. are well in line with the Sensory Rearrangement Theory (Reason & Brand, 1975), the results at hand indicate that top-down attention might be an important influencing factor.

EDA data recorded during the simulator session supported the subjective rating of well-being. The effects on EDA parameters are well in line with modern arousal conceptions like the 4-arousal-model (Boucsein & Backs, 2009). Higher levels of electrodermal responses throughout the Psychophysiological Unit of 4x10 seconds indicated that participants of the experimental group experienced more negatively toned emotions and/or emotional stress in the Coriolis-prone environment as compared to the control group. It can be interpreted that the distraction of attention might have influenced anticipatory processes negatively, so that the expectation of upcoming sensory information matched the actual sensed sensory information to a lower degree and, hence, led to a higher impact of Coriolis effects. Interestingly, participants of the experimental group showed significantly less mismatches when they allocated their attention mainly to the ongoing scenario. In this experiment, the results of subjective data and psychophysiological data suggest that anticipatorily controlled top-down processes are of particular importance in Coriolis-provoking environments.

Acknowledgements

This experiment was part of a project funded by the Austrian Ministry for Transport, Innovation and Technology, TAKE OFF - Technology between sky and earth, 2012. FFG project number 839013. The views of the research reported do not reflect the views of the granting organization.

References

Benson, A. (2002). Motion sickness. In K. B. Pandoff (Ed.), *Medical aspects of harsh environments,* Bd. 2 (pp. 1048-1083). United States: Government Printing.

Boucsein, W. & Backs, R. W. (2009). The Psychophysiology of Emotion, Arousal, and Personality: Methods and Models. In V. G. Duffy (Ed.), *Handbook of Digital Human Modeling* (pp. 35-1 – 35-18). Boca Raton: CRC Press/ Taylor & Francis.

Butz, M. V., & Pezzulo, G. (2008). Benefits of anticipations in cognitive agents. In G. Pezzulo, M. V. Butz, C. Castelfranchi, & R. Falcone (Eds.), *The challenge of anticipation. A unifying framework for the analysis and design of artificial cognitive systems* (pp. 45-64). Berlin: Springer-Verlag.

Cheung, B. (2013). Spatial disorientation: more than just illusion. *Aviation, Space, and Environmental Medicine, 84*, 1211-4. doi: 10.3357/ASEM.3657.2013

Gibb, R., Ercoline, B., & Scharff, L. (2011). Spatial disorientation: decades of pilot fatalities. *Aviation, Space and Environmental Medicine, 82*(7), 717-24. doi: 10.3357/ASEM.3048.2011

Gresty, M. A., Golding, J. F., Le, H., & Nightingale, K. (2008). Cognitive impairment by spatial disorientation. *Aviation, Space and Environmental Medicine, 79*, 105-11. doi: 10.3357/ASEM.2143.2008

Hoffmann, J. (1993). *Anticipation and cognition: The function of anticipations in human behavioral control and perception*. Göttingen: Hogrefe.

Kallus, K. W. (2012). Anticipatory processes in critical flight situations. In A. De Voogt & T. D'Oliveira (Eds.), *Mechanisms in the chain of safety* (pp.97-106). Ashgate.

Keshavarz, B., & Hecht, H. (2011). Validating an efficient method to quantify motion sickness. Human Factors: *The Journal of the Human Factors and Ergonomics Society, 53*(4), 415-426. doi: 10.1177/0018720811403736

Koglbauer, I., Kallus, K. W., Braunstingl, R., & Boucsein, W. (2011). Recovery training in simulator improves performance and psychophysiological state of pilots during simulated and real Visual Flight Rules flight. *The International Journal of Aviation Psychology, 21*(4), 307-327. doi: 10.1080/10508414.2011.606741

Reason, J. T., & Brand, J. J. (1975). *Motion sickness*. London: Academic Press.

Talker, C. M., Kallus, K. W., Schwandtner, J., Joachimbauer, J., & Beykirch, K. (2014, September, 22-26). *The influence of top-down processes on perception in spatial disorientation-prone situations*. Paper presented at the Proceedings of the 31th EAAP Conference. Aviation Psychology: facilitating change(s). With special sessions on Flight deck and ATC., Valetta, Malta.

Tropper, K., Kallus, K. W., & Boucsein, W. (2009). Psychophysiological evaluation of an antidisorientation training for Visual Flight Rules pilots in a moving base simulator. *The International Journal of Aviation Psychology, 19*(3), 270-286. doi: 10.1080/10508410902983912

PROCEDURE USED FOR ESTABLISHING SCREENING TEST CUT-POINTS BASED ON AVIATION OCCUPATIONAL TASK PERFORMANCE

Nelda Milburn, Thomas Chidester, Kevin Gildea, and Linda Peterson
Federal Aviation Administration, Civil Aerospace Medical Institute
Oklahoma City, OK
Carrie Roberts and Deborah Perry
Xyant Technology, Inc.
Norman, OK

Previous research has shown that some individuals with color vision deficiencies (CVD) are capable of performing some aviation occupational tasks as well as those with normal color vision (NCV); implying that passing a screening test with a diagnosis of NCV may not be necessary for all aviation occupations. Our goal was to find *outcome consistency* between performance on occupational tasks and several screening tests; further, to compare those pass/fail outcomes to the Colour Assessment and Diagnosis (CAD) test for aviation certification. The strategy involved establishing a pass/fail cut-point separately for four occupational tasks at the 5th percentile of the NCV group. A scatterplot was constructed displaying the sum of correct screening test trials on the x-axis and the red/green threshold of the CAD test on the y-axis. By defining the markers according to pass/fail status on the occupational tasks, it was easy to evaluate multiple factors to arrive at an appropriate cut-point for each screening test.

All of the Federal Aviation Administration (FAA) approved color vision screening tests have reasonable pass/fail agreement (as measured by kappa scores; Cohen, 1960) for diagnosing normal or deficient color vision when compared to diagnoses with the Nagel anomaloscope (Mertens & Milburn, 1993). However, we have found that some individuals with color vision deficiencies (CVD) are capable of performing some aviation occupational tasks as well as individuals with normal color vision (NCV). Because the FAA allows different several color vision screening tests to ensure that those passing are capable of performing the necessary pilot color-decoding tasks, it is important to match the pass/fail cut-point to performance on aviation color-coded tasks. Furthermore, if an airman fails the initial screening, he/she can request secondary screening involving presentation of signal lights at an airport by FAA personnel, which is time-consuming and more expensive to conduct than in-office clinical screening tests. Therefore, it is prudent to maximize the sensitivity and specificity of the in-office screening tests by appropriately setting the pass/fail cut-points; and, it is also important from a safety standpoint to minimize the "false negative" numbers to ensure that airmen that pass the screening test are capable of accomplishing the requisite color tasks of modern aviation.

Purpose

Our goal was to find a decision point that reliably defined performance on occupational tasks matched to passing cut-points on clinical and precision tests; and further, to compare those cut-points to a valid and reliable criterion measure. The Colour Assessment and Diagnosis (CAD) certification standard was based on performance of aviation-related color tasks and has been successfully in use for pilot selection since 2008 (Barbur, Evans, & Milburn, 2009). The CAD conveniently provides individual threshold values in standardized normal units. By doing so, it essentially quantifies, on a linear scale, one's ability to see color, which in turn can be linked more directly to percent correct performance on specific tasks, unlike traditional normal/deficient test outcomes.

Method

Participants

The CAD test was used for diagnoses of type and degree of color vision deficiency, and it has a high diagnosis agreement for red-green types of color vision deficiencies with the gold-standard, the Nagel anomaloscope (Barbur et al., 2009). However, the Nagel does not diagnose yellow-blue types of deficiencies, and it requires tedious, one-on-one screening, averaging about 20 to 30 minutes to complete. Study participants included 57 males and 38 females, with 89% between the ages of 18 and 31 to match the population of air traffic control applicants (a separate study). Ten adults over 31 years of age with CVD were recruited to equalize the NCV and CVD groups. Most subjects with CVD were male because the congenital deficiency results from a recessive trait on the X chromosome. All subjects met a screening requirement of at least 20/30 near and far visual acuity. The participants included 47 NCV and 48 with CVD, classified by type of deficiency: 16 protan, 20 deutan, 3 tritan, and 9 exhibiting both red-green (RG) and yellow-blue (YB) weaknesses. Table 1 shows participant CAD type classifications and thresholds.

Table 1.
Participant Color Vision Classification and Threshold Values

		CAD Test	
Diagnosis	N	Red/Green Threshold	Yellow/Blue Threshold
Normal	47	.84 - 1.71	.67 - 1.62
Protan	16	11.67 - 29.84	.66 - 1.64
Deutan	20	2.82 - 29.31	.71 - 1.64
Tritan	3	1.35 - 1.68	1.79 - 1.98
RG & YB	9	1.76 - 30.77	1.83 - 15.06

Materials

Evaluation measurements were defined as clinical tests, computerized tests, or occupational tasks. The clinical tests and the computerized tests are available commercially; however, the occupational tasks were created in the laboratory (with the exception of the signal light gun) strictly to serve as work samples for validation purposes. Consequently, we had to determine an appropriate passing score for the *occupational* tasks. That process is described later in this paper. The *clinical tests* included: the Dvorine®, the Ishihara (-14, -24, and -38 plate versions), the Waggoner HRR®, the Waggoner PIPIC®, the Richmond Products HRR®, and the Stereo Optical 900® (OPTEC 900®). The *computerized tests* included: the Rabin Cone Contrast Test (RCCT®), ColorDx®, and the Colour Assessment and Diagnosis (CAD®) Test.

The *occupational tasks* included:

- Incandescent red (R) and white (W) precision approach path indicator lights (INC-PAPI), which included 26 pairs of lights, scored as 52 trials. Light pairs were presented with the following combinations: R-R, R-W, W-R, and W-W.
- Light-emitting diode red and white lights (LED-PAPI) that included 64 pairs of lights, scored as 128 trials. Light pairs were presented with the following combinations: R-R, R-W, W-R, and W-W.
- Signal light gun test (SLGT). The SLGT presents a single light of red, green, or white. A total of 6 lights were presented at 1,000 ft and 6 lights at 1,500 ft. To pass, FAA Order 8400 stipulates that no errors are allowed on the 12 trials to pass.

- Pilot cockpit display colors task was comprised of 10 targets (trials) for each of 8 colors. Trials were presented as colored text (red, white, green, blue, cyan, yellow, amber, or magenta). The total percent correct was scored using the number of correct selections minus false positives.

Strategies to achieve consistency

With our ultimate goal in mind—to find an appropriate cut-point on the screening test that ensures those passing are capable of performing critical, color-coded aviation occupational tasks—we tried several options. First, we explored setting pass/fail cut-points on the *occupational tasks* at 95% correct of all trials, separately for each task, which seemed like a reasonable standard generally accepted in academia to reflect good performance. The final decision was to use a performance equivalent to the 5th percentile of the normal color vision group because about 95% of all individuals have normal color vision. About 8-10% of men and less than ½ of 1% of women have a color vision deficiency. In other applications, using a pass/fail point of the 5th percentile of the normal color vision group is similar to setting a production goal for assembly-line workers based on a goal that 95% of the workforce meets or exceeds. Therefore, to set individual screening test cut-points, we took the following steps:

1. Determine the 5th percentile for the NCV group for each occupational task separately (NCV determined by CAD test)
2. Cross-tabulate the pass/fail of 5th percentile performance on the occupational tasks with CAD certification
3. Cross-tabulate 5th percentile occupational task pass/fail performance with each clinical test using the manufacturer's pass/fail criterion and evaluate their agreement
4. Graph the sum of correct trials (x-axis) by CAD RG thresholds (y-axis) and color-code points by pass/fail performance (at the 5th percentile of NCV group) on the composite of occupational tasks, separately, for each screening test
5. To guard against jeopardizing the integrity of screening tests and to prevent motivated examinees from memorizing a limited number of plates, the total trials administered must exceed 11 and the passing score must exceed a minimum of 7 trials correct
6. Using both the cross-tabulation tables and the graphs, we examined the FAA-defined, pilot cut-points (when available) for *clinical* tests to determine their effectiveness for passing those who performed the occupational tasks at the 5th percentile of the NCV group. When necessary, we altered the cut-point of the *clinical* tests to balance the false positive with the false negative cells to achieve optimal sensitivity and specificity scores (using composite *occupational* task performance as the criterion measure)
7. For those tests without FAA pilot cut-points, we followed the same procedure as previously described to set screening test cut-points
8. Once the pass/fail cut-points were selected, the burden on airmen could be evaluated—meaning the number of airmen that would be required to take medical flight tests (MFTs) as a result of failing their initial screening test

These strategies were used to find consistency between multiple screening tests, performance on several occupational tasks, and a linear scale of color vision ability in standardized normal units, which some might argue is a proven pilot certification test. This paper will focus on the techniques employed for setting cut-points rather than the cut-points assigned to each color vision screening test because we believe that when establishing cut-points, traditional methods may not allow the researcher, test developer, or test validator to fully see multiple comparisons that are essential to validating a test. Typically, when establishing a screening test cut-point, one seeks to find the point that maximizes the sensitivity and specificity of the test without sacrificing one for the other; but, our ultimate goal was to differentiate between those who can and cannot perform safely within a reasonable degree of accuracy and certainty. Several statistical tests can be used to measure the agreement between the screening test and the criterion measure (Milburn & Mertens, 2004). In medicine, the criterion measure may be whether

or not a person has a particular disease and screening tests are used to predict the disease when determining the presence or absence of the disease is expensive or invasive. Sometimes, screening tests are used to predict performance, such as standardized tests (e.g., ACT, SAT, or GRE, which are used to predict readiness for college or graduate school). Likewise, a screening test can be used to predict ability to perform a specific task. The FAA uses color vision screening tests to predict one's ability to decode and interpret vital color-coded information used in signal lights.

In the current research, the purpose of the occupational tasks was to simulate actual work tasks required by pilots that necessitate good color perception to perform safely. One possible solution would have been to hire several seasoned pilots to perform the occupational tasks and set a pass/fail point based on their performance; however, we honed the tasks to the bare essential requirement of identifying, naming, differentiating, or matching colors—tasks that did not require any piloting experience or knowledge.

It may be important to explain why, if we are using the CAD aviation certification test as a benchmark or a reference point for performance, we don't simply use the CAD certification exclusively for all pilot screening. The answer is that the test is not conveniently available at national test sites or at aviation medical examiners' offices; additionally, it is very expensive (about $9,000 US dollars). Consequently, it is unlikely to be widely purchased by aviation medical examiners when the return on investment ratio is minimal, and likely requires many years to break-even. Furthermore, a typical flight physical costs about $200 with color vision screening as a very small part of the examination.

Results

Using the strategy described above, setting the cut-point was easy in some cases and more difficult in others. For example, The Waggoner PIPIC is a relatively new pseudoisochromatic plate test, which we categorized as a clinical test. It was not in production when the FAA first set pilot-specific cut-points (Mertens & Milburn, 1993). Because all NCV participants passed all screening tests with a strict (NCV) criterion; and, because 5% of NCV participants were sacrificed to establish the 5^{th} percentile pass/fail point for the occupational tasks, we chose to isolate only the CVDs to examine the cost/benefit of setting a new cut-point. Using the manufacturer's criterion for determining NCV (12 of 14 correct), 13 CVD subjects failed the screening test but passed all of the occupational tasks with performance equivalent to, or exceeding the 5^{th} percentile of the NCV group. Table 2 is a cross-tabulation of performance on the Waggoner PIPIC using the new (9 of 14) cut-point by pass/fail performance on the composite of occupational tasks for only the CVD participants and shows a gain of 7 additional subjects passing the screening test that also passed all of the occupational tasks. There were also 6 subjects that failed the Waggoner PIPIC who were also able to perform the occupational tasks. There was a substantial improvement in the number passing the screening test, but as you can see from the scatterplot in Figure 1, some CVD participants scored very poorly on the screening test but well on the occupational tasks. This evidence supports the FAA's occupational color vision test (OCVT) or the FAA developing such a test that can be administered in the office, such as the air traffic color vision test (ATCOV), which serves that purpose for air traffic control applicants (Chidester et al., 2011).

Figure 1 best presents the outcome of using the strategy we described previously to set individual screening test cut-points. Essentially, if an examinee passes at least one of the 12 screening tests, he/she should be highly likely to accomplish the color-coded aviation occupational tasks that we examined. Some points are overlapping—47 subjects passed all 12 screening tests and 31 subjects failed all 12 and that information is not readily apparent on the scatterplot. Therefore, it is essential that cross-tabulation tables are used in conjunction with the scatterplots. Not all screening tests measure the same color perception abilities; hence, some subjects do well on some tests and not on others. We recognize that the manufacturer's cut-point, based on separating NCV from CVD, is an unfair requirement for some CVD

examinees. We found that regardless of which approved screening test the examinee takes and passes, it is highly likely that his/her performance on color-coded tasks will be essentially as good as the 5th percentile of the normal color vision group.

Table 2.

Composite Occupational Tasks for Pilot Cockpit, LED, and Incandescent Precision Approach Path Indicator (PAPI) Systems, and Signal Light Gun Test by the Waggoner PIPIC Pass/Fail Cross-tabulation[a]

		Waggoner PIPIC		Total
		Fail	Pass	
Composite Occupational Tasks for Pilot Cockpit, LED/Incandescent PAPI, and Signal Light Gun Test	Fail ANY	34	1	35
	Pass ALL	6	7	13
Total		40	8	48

a. Color Deficient subjects only

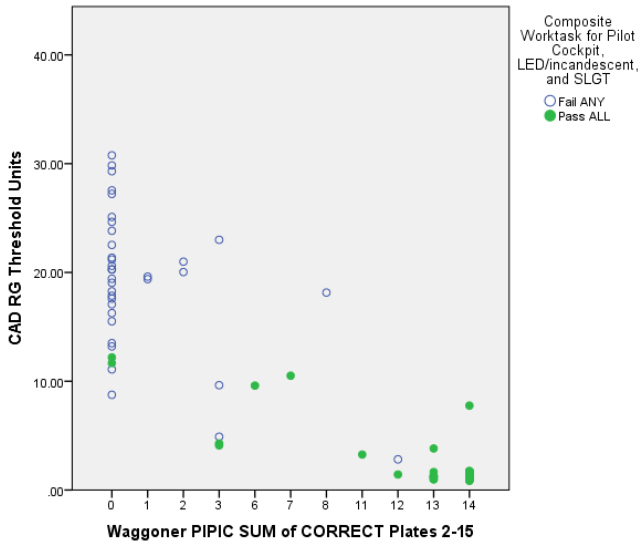


Figure 1. Scatterplot of the Waggoner PIPIC sum of correct responses on plates 2-15 by the Colour Assessment and Diagnosis Test red-green threshold, with the markers coded by passing or failing the occupational tasks.

One example of the improvement in Kappa agreement scores between screening test and occupational performance was the Waggoner HRR that changed from $K_{(92)}$= .62 to .72. Using this methodology to assign screening test cut-scores (rather than using the screening tests' designation of NCV to determine pass/fail), we found agreement scores, for the tests we examined, ranging between $K_{(92)}$=.70 to .78 and averaging .72. Agreement scores improved from those obtained using the previous

cut-points—but more importantly, based on our validation analyses, we are assured that those cleared based on accepted screening tests will be highly likely to perform well on aviation color-tasks.

Conclusions

By using coded markers on a scatterplot with a diagnostic quantitative measure in standard normal units on the Y-axis and screening test scores on the X-axis, the decision points can be easier to determine. In contrast, one must explore several different pass/fail points by coding variables for analysis, and then create multiple cross-tabulations with the criterion variable to examine the resulting kappa agreement scores—to check for improvements. We believe that the methods we used and presented here may be a helpful strategy that is applicable to other test developers, researchers, and test validators.

Acknowledgements

Research reported in this paper was conducted under the Flight Deck Program Directive / Level of Effort Agreement between the Federal Aviation Administration Headquarters and the Aerospace Human Factors Division of the Civil Aerospace Medical Institute sponsored by the Office of Aerospace Medicine and supported through the FAA NextGen Human Factors and Engineering Division.

References

Barbur, J., Evans, S., & Milburn, N. (2009). Minimum Color Vision Requirements for Professional Flight Crew, Part III: Recommendations for New Color Vision Standards. (Report No. DOT/FAA/AM-09/11). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Chidester, T., Milburn, N., Lomangino, N., Baxter, N., Hughes, S., & Peterson, L. (2011). *Development, Validation, and Deployment of an Occupational Test of Color Vision for Air Traffic Control Specialists.* (Report No. DOT/FAA/AM-11/8). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Cohen, J.A. (1960). Coefficient of Agreement for Nominal Scales. *Educational & Psychological Measurement*, 20:37-46.

Mertens, H., & Milburn, N. (1993). *Validity of FAA-Approved Color Vision Tests for Class II and Class III Aeromedical Screening.* (Report No. DOT/FAA/AM-93/17). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Milburn, N., & Mertens, H. (2004). *Predictive Validity of the Aviation Lights Test for Testing Pilots With Color Vision Deficiencies*. (Report No. DOT/FAA/AM-04/14). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

CONCEPTUAL AND PROCEDURAL TRAINING FOR SITUATION AWARENESS AND PERFORMANCE IN AN INSTRUMENT HOLDING TASK

Andrew R. Dattel
Jennifer E. Thropp
Embry-Riddle Aeronautical University
Daytona Beach, FL

An exploratory approach that investigated the differences between conceptually and procedurally trained participants in situation awareness (SA) and performance of instrument holds was conducted. The step-by-step actions required to fly instrument holds were emphasized in the procedural training group. The interrelationship of elements in a dynamic environment was emphasized in the conceptual group. Participants were tested in two simulated instrument holding pattern scenarios. The second holding pattern was designed to be more complex. A trend was found where the conceptual group showed less altitude deviation (M = 399.22) than the procedural group (M = 599.74). Participants were asked six SA questions in each task. In the first task, the conceptual group answered an average of 3.30 questions correctly, whereas the procedural group answered 2.75 questions correctly. In the more difficult task, the spread increased with the conceptual group answering an average of 3.20 questions correctly, whereas the procedural group answered only 2.25 questions correctly.

The minimum flight time to earn a private pilot's license in the United States is between 35 and 40 flight hours of training. However, the average flight hours for students to earn a private pilot's license is about twice the amount above the minimums (Federal Aviation Administration, 2006). Finding ways to maximize the effectiveness and efficiency of training for all phases of flight training can decrease time and cost needed to acquire flight certificates and ratings. This study examined the effect of conceptual and procedural training on the skill development and situation awareness (SA) for a critical instrument flight maneuver—instrument holds. Instrument holds are racetrack patterns that pilots fly to basically remain in a static position. Instrument holds are occasionally requested by air traffic control (ATC) when air traffic has become congested or backed up.

Conceptual and procedural training were used to train participants for this study. Conceptual training uses metaphors, analogies, diagrams, etc. to emphasize the interrelationships of elements in a dynamic environment and the *reasons* maneuvers are made. Explaining how instrument holds are similar to racetracks is an example of conceptual training. Procedural training emphasizes the required tasks, or step-by-step *actions* needed to successfully complete a maneuver. Reducing power to a set airspeed, turning to a set heading, setting a timer, etc. are examples of procedural training for instrument holds. Conceptual training teaches one "how a system works" and procedural training teaches one "how to work a system" (Bibby & Payne, 1993).

Both procedural and conceptual training are important for skill development in flight training. Procedural training is important for developing automatic processing for routine situations. For non-routine situations (maneuvers that are expected to occur less frequently), conceptual training is beneficial. Additionally, conceptual training is best for complex tasks and SA (Hockey, Sauer, & Watsell, 2007). SA is the comprehension of the relevant information in a rapidly changing environment (Durso, Rawson, & Girotto, 2007). Determining the best training approach to emphasize at various stages of learning a flight maneuver will maximize how well a student learns a maneuver.

Dattel, Durso, & Bédard (2009) found evidence that a review of the conceptual aspects of landings and traffic pattern maneuvers improved subsequent performance when compared to a control group and a group that received a procedural review. Dattel et al. (2013) found that conceptually trained participants showed the same level of SA

while executing holding patterns at varying degrees of difficulty, but procedurally trained participants showed poorer SA when encountering more difficult holding patterns.

The current study tested pilots who had no or minimal instrument training. Participants received introductory training for instrument holds with an emphasis in either conceptual training or procedural training. It was expected that the group that received an emphasis on conceptual training would show better SA than the group that received an emphasis on procedural training. Because instrument holds are a complex flight maneuver that require high level cognitive skills, such as retaining assigned heading and altitude information, as well as requiring quick algebraic and geometric calculations, it was expected that the participants who received an emphasis on conceptual training would perform better than the participants who received an emphasis on procedural training.

Method

Participants

Sixteen private pilots were recruited via a flyer sent to their listed mailing addresses. Participants received approximately 1 ½ hours of training followed by about 30 minutes of testing. They were paid $15 per hour. Participants were randomly assigned to either a conceptual training group or a procedural training group.

Materials

Training Material

All participants received text training and video training. The text used for procedural training included the actions required to conduct an instrument hold in a step-by-step fashion. The text used for conceptual training included reasons for conducting an instrument hold, as well as diagrams from a top-down perspective. The text training stimuli were presented in Adobe (.pdf) files and Microsoft PowerPoint (.ppt) files. Videos of two holding patterns were recorded using Microsoft Flight Simulator. A view of the cockpit from the pilot's perspective was used for the procedural training, and the same flight viewed from outside the aircraft, as from a bird's eye view, was used for the conceptual training.

Each participant received two sections of the text training and two sections of the video training. Three questions (with forced-choice answers and short answers) about the respective training material were developed for each section.

Testing Material

An Elite PI-135 Personal Computer Aviation Training Device (PCADT) was used to test participants. Two instrument holds were developed. The simulation began mid-flight about 7 miles from the initial holding fix (waypoint) for each test scenario. Holding instructions were provided to the participant before starting the flight. The first test scenario included a flight that included typical instrument hold maneuvers with standard rate turns. The second test scenario was more complex and required atypical maneuvers and non-standard turns. Six SA questions were created for each flight and presented in the SPAM format (see Durso & Dattel, 2004).

Procedure

Participants assigned to the procedural group read the procedural training text document followed by watching the procedural training video. After viewing each training section, participants had to answer three questions (multiple choice and short answer) about the respective training material. Participants had to answer each question correctly before continuing to the next question and the next section. Participants in the procedural group read the same procedural training text document and watched the same procedural training video for a second time, but had to answer a different set of questions. The participants were permitted and encouraged to review the training text document and the training video for assistance in answering a question for which they were unsure. Data was removed from analyses for participants who did not answer a question correctly after a third attempt.

Because it would be unrealistic to expect a participant to fly an instrument hold without having any exposure to the procedures necessary to fly an instrument hold, the conceptual participants received half of the amount of time of procedural training that the procedural group received. For the other amount of time, the conceptual group received the conceptual training stimuli. As with the procedural group, the conceptual group had to answer the same amount of training questions and were allowed to review the training material for assistance in answering the questions. Participants assigned to the conceptual group read one of the procedural training text documents, the conceptual training document, the procedural training video file, and the conceptual training video, in that respective order.

After completing the training, participants flew two simulated flights that required an instrument holding pattern in a flight simulator (PCATD). For each flight, participants were instructed to enter an instrument hold and remain in the hold until further notice. Participants were provided written holding instructions before each flight. Each test scenario lasted 10 minutes, at which time the simulation was stopped. Six SA questions were provided over a headset at approximately 1 ½ minute intervals. A warning bell was played before each SA question was played. Participants were instructed to answer the question aloud as quickly, but as accurately as possible. Participant responses were audio recorded. Each SA question was relevant to the test scenario.

Results

The mean age of the participants was 43.89 years (*SD* = 17.31). The mean number of flight hours of the participants in the conceptual group was 139.62 (*SD* = 87.78), while the mean number of flight hours of the participants in the procedural group was 275.00 (*SD* = 173.11); this difference was not significant ($p > .05$).

Date Preparation

Three performance measures were scored: heading deviation, altitude deviation, and airspeed deviation. For each leg of the instrument hold, participants should have flown a particular heading. As per training and test instructions, participants were to fly at a particular altitude and airspeed. The root mean square of the deviations from what the participants' heading, altitude, and airspeed were, to what they should have been, were the units of measurement in the analyses. Two measures of SA were recorded: accuracy and response time. Accuracy was measured by the total number of questions answered correctly in a scenario. Average response time of SA questions answered correctly was also calculated. Unfortunately, some of the flight performance and SA recordings were lost due to technical difficulties, leaving only about 70% of the original data useable.

Performance Data

For each participant, heading, altitude, and airspeed were recorded once per second in both an easy (Hold 1) and difficult (Hold 2) condition. The root mean square error (RMSE) was calculated to determine the participants' deviations from the prescribed flight plans. Raw RMSEs for heading, altitude, and airspeed for the conceptual and procedural groups (across difficulty levels) are presented in Table 1.

Table 1.
Average Deviation in RMSE in Performance Data for Both Holds Combined

	Altitude RMSE (*SD*)	Heading RMSE (*SD*)	Airspeed RMSE (*SD*)
Conceptual	399.22 (395.78)	56.67 (27.73)	16.20 (6.05)
Procedural	599.74 (205.49)	64.19 (26.80)	15.73 (7.18)

Figures 1, 2, and 3 show the raw score RMSEs for the easy (Hold 1) and difficult (Hold 2) hold patterns for heading, altitude, and airspeed, respectively.

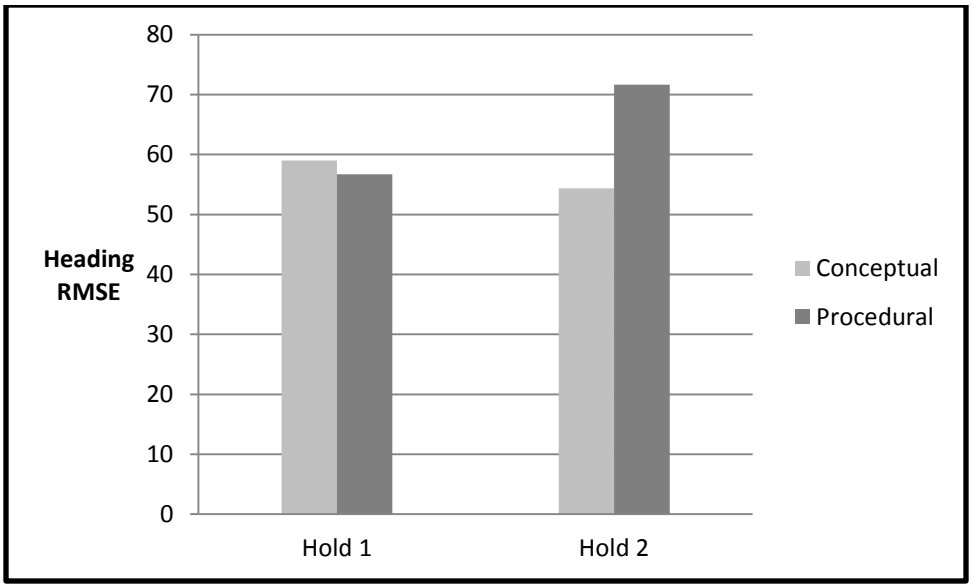


Figure 1. Heading RMSE for Holds 1 and 2.

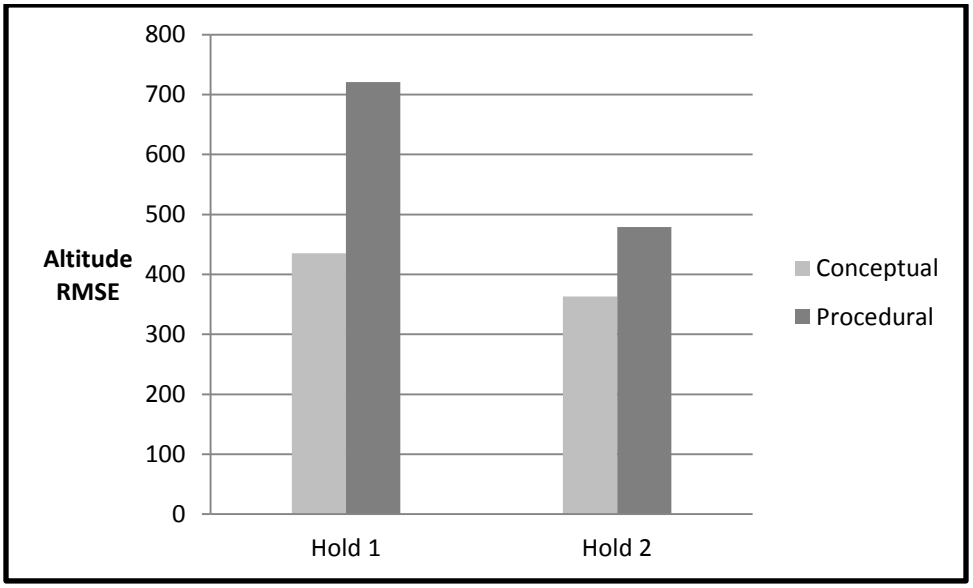


Figure 2. Altitude RMSE for Holds 1 and 2.

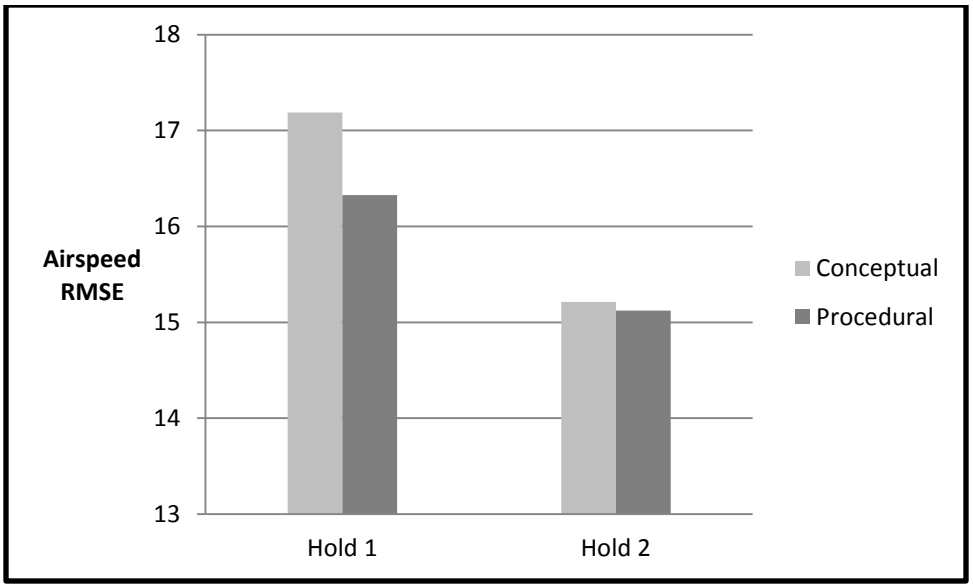


Figure 3. Altitude RMSE for Holds 1 and 2.

A log transformation was then performed on the RMSE data to correct for the positive skew in the data. A 2 (Difficulty Level) X 2 (Group: Conceptual vs. Procedural) mixed measures ANOVA was conducted on the log transformed altitude deviation RMSE. There was a trend for the conceptual group to have a lower altitude RMSE than the procedural group, $F(1, 11) = 3.564$, $p = .086$. There was no significant group effect for either heading or airspeed RMSEs (both $p > .05$).

Situation Awareness Data

A 2 (Difficulty Level) X 2 (Group: Conceptual vs. Procedural) mixed measures ANOVA was also conducted on SA accuracy and response time; no significant differences were found ($p > .05$). The means for SA and response time for both the conceptual and procedural groups are presented in Table 2.

Table 2.
Descriptive Statistics for SA Accuracy and Response Time

	SA Accuracy		SA Response Time	
	Hold 1 Mean (*SD*)	Hold 2 Mean (*SD*)	Hold 1 Mean (*SD*)	Hold 2 Mean (*SD*)
Conceptual	3.30 (1.70) *n* = 10	3.20 (1.87) *n* = 10	3.08 (2.44) *n* = 7	3.61 (1.73) *n* = 7
Procedural	2.75 (1.50) *n* = 4	2.25 (2.22) *n* = 4	2.31 (1.29) *n* = 3	2.87 (2.54) *n* = 3

Discussion

No significant differences were found between groups on any of the SA measurements. Although the conceptual group's accuracy for answering SA questions changed slightly across test scenario difficulty (3.30 questions answered correctly in Hold 1 compared to 3.20 questions answered correctly in Hold 2), this difference in more pronounced, albeit not significant, for the procedural group (2.75 questions answered correctly in Hold 1 compared to 2.25 questions answered correctly in Hold 2). Dattel et al. (2013) found poorer SA across task difficulty for a group that received an emphasis in procedural training compared to a group that received an emphasis in conceptual training. It is believed that if the data that was lost to technical difficulties followed the same trend, then a significant difference for SA accuracy would have been found between groups.

Differences between groups were not found for the airspeed and heading performance. However, the conceptual group had less altitude deviation across test scenarios when compared to the procedural group (p = .086). Heading and airspeed are two important performance parameters that must be maintained in instrument holds. Significant deviations from the proper tracking in an instrument hold can result in a pilot's aircraft drifting to the "unprotected" side which could result in a mid-air collision. Airspeed is also a critical parameter in a hold. Because aircraft are flown at slower than cruise airspeed in instrument holds, in addition to making turns that increase the stall speed, a pilot must be constantly aware of the aircraft's airspeed to prevent a stall.

The better performance of maintaining altitude by the conceptual group may indicate better implied SA. As noted above, the SA data had less power than the performance data. However, if pilots have a better understanding of SA, then one can expect better SA due to less individual workload. That is, the conceptual training pilots may have had more cognitive capacity to obtain to a performance measure that may not have seemed as critical at the time, consequently showing improved performance.

This study suggests that an emphasis on conceptual training for complex flight maneuvers is better for flight students than procedural training. Although this study had low power, partly due to technical difficulties, continuing to study how procedural and conceptual training impact different training of flight maneuvers is warranted. Future studies should compare the differences in procedural training and conceptual training for more cognitively complex flight maneuvers and more procedurally oriented flight maneuvers.

References

Bibby, P.A. & Payne, S. J. (1993). Internalization and the use specificity of device knowledge. *Human Computer Interaction, 8*, 25-56.

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury and S. Tremblay (Eds.). *A cognitive approach to situation awareness: Theory, measurement and application*, (pp. 137- 154). Burlington,VT: Ashgate.

Dattel, A. R., Durso, F. T., & Bédard, R. (2009). Procedural or conceptual training: Which is better for teaching novice pilots landings and traffic patterns? In *Proceedings of the 53rd Annual Human Factors and Ergonomics Society*, San Antonio, TX

Dattel, A. R. Kossuth, L., Sheehan, C. C., Green, H. J., Giannini, C., Decker, J., Mericle-Swingle, H. M., Crockett, S. A. (2013). Effects of conceptual training and procedural training for teaching aviation instrument holding patterns. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 1445-1449). San Diego, CA: Human Factors and Ergonomics Society.

Durso, F. T., Rawson, K. A., & Girotto, S. (2007). Comprehension and situation awareness. In F.T. Durso, R.S. Nickerson, S.T. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of applied cognition (2nd ed.),* (pp. 163-193). Hoboken, NJ: John Wiley & Sons.

Federal Aviation Administration (2006). *Student pilot guide* (Publication No. FAA-H-8083-27A). Retrieved from http://www.faa.gov/regulations_policies/handbooks_manuals/aviation/

Hockey, G. R. J., Sauer, J., & Wastell, D. G. (2007). Adaptability of training in simulated process control: Knowledge-versus rule-based guidance under task changes and environmental stress, *Human Factors, 49*, 158-174. doi: 10.1518/001872007779598000

PSYCHOLOGICAL ASPECTS OF THE ORGANIZATION OF INFORMATION AT THE INSTRUCTOR'S FLIGHT SIMULATOR WORKPLACE

Tetiana Bondareva
National Aviation University
Kyiv, Ukraine

This Paper presents a critical analysis of the classical procedure of training pilots on Full Flight Simulators (FFS), and developed new version of the procedure of instructor's actions during simulator training. The developed procedure will change an instructor's position as person, who sets parameters before and during performing exercise, to a person as involved participant, who is waiting for a forthcoming training flight circumstances together with the crew. This developed procedure implies a rearrangement of specific operations from the instructor on the software, and also the availability of an alternative monitor, on which the most significant events (selected for specific exercises and situations) of training flight will be displayed. The developed procedure of instructor's actions is aimed to reduce the quantity of processed information and actions, which will lead to increase of instructor's work efficiency, as a result of lowering of a level of fatigability.

To understand the classical procedure of an instructor's actions it is necessary to consider a brief review of the current approaches in the design of an instructor's workplace interface.

Analysis of the development of instructor's workplace interface from the 90's of 20th century reveals the relative immutability of approaches of instructor's workplace organization and a structure of the interface. The main type of instructor's interaction on station is touchscreen. Grouping of elements with wide and narrower categories is the main type of interface design.

The information model of instructor's workplace is considered on the example of a workplace inside the cockpit of a FFS for aircraft Antonov-148, as a typical representative of a current generation of flight simulators.

Man–machine interface IOS is based on the principle of grouping information with themed frames, which gives the opportunity to work in a touchscreen mode. Each frame contains a number of parameters, which ideologically belong to a particular group of management and control of a training process, operating mode of a simulator. Frame is the main screen of an instructor's workplace program, which is divided into two sections:

1) the field of an active frame;
2) quick access panel.

Quick access panel is presented all the time, regardless which frame is currently displayed in the active frame field. This panel allows to move to the desired frame, but the panel doesn't contain all possible frames, but only those that, according to the developer, are most

needed for quick access. The remaining frames, those which were not included in the quick access panel, are needed to be called in the active frame field using the main button "list of frames", which, when is clicked, opens a list of all frames, including those, which are presented in the quick access panel. Example: In order to display the frame of high-lift device failure, you can choose the button “input failure “(broad category) both on the quick access panel and through the full list of frames, and then choose a narrower category - “high-lift device”. To change the parameters of the equipment it is necessary to call a list of frames (broad category) and select a frame “equipment settings” (narrow category), because there is no such frame on the quick access panel.

If we generalize the concept of the formation of the instructor's station, the following options can be marked out:

1) graphical interface;
2) grouping of information in the form of frames;
3) interface is optimized for the touch screen work;
4) Grouping by category: wide categories include narrow categories**,** which are related to the certain systems of the simulator.

The advantages of this instructor’s station are in the fact, that with the help of considered categorization of parameters and touchscreen, it quickly gives access to the necessary information any time you need. It is also possible a scaling of parameters, i.e. inclusion of any number of monitors for parallel operation (control / input parameters) of instructor.

The disadvantage of this station is instructor’s overload as an operator. Because of the large flow of information, which he must have time to register, then he needs to change the parameters during performing a training flight, and at the same time he needs to monitor the actions of the crew, that performs this training flight. Increased productivity can be achieved by changing the classical procedure of instructor’s action. Instructor, being in the same involvement position as a manned crew, will be removed from the task to set parameters before and during performing training flight.

The developed procedure of the instructor’s action

The procedure aims to maximize the involvement of instructor in the situation of the training flight, by his minimal distracting to other parameters and circumstances, which are not directly related to his primary task - evaluation of psychophysiological potential of the piloting crew. According to the psychophysiological concept of simulator’s training (Gorbunov, 2001) FFS should not only keep track of skills working off of emergency situations (for which procedural simulators can be used) but should determine the level of psychophysiological load of manned crew. Psychophysiological load characterizes the efforts that are involved in the successful implementation of the task, which will adequately assess a degree of the pilot’s qualification, from the standpoint of human factors (Gorbunov,2000).

Deep instructor’s involvement in the situation of the training flight is needed for the evaluation of quality of work for each member of the crew, by identifying the procedural errors and psychophysiological assessment of pilot’s readiness. An important task of training flight on

FFS is to measure the psychophysiological potential, that pilot uses to solve a particular problem. Instructor's task is to be the finest instrument for the measure of the psychophysiological potential of each member of the crew, because in the modern practice there is no hardware's way of its measuring. For the decreasing of instructor's distracting level on secondary tasks it is necessary to get rid from large data flows and shift a part of the instructor's operations to the software.

The procedure will allow instructor to allocate resources in the most effective way. These resourses are: time, physical and psychological capability of the organism. Two significant improvements were added for the implementation and testing procedure into the design and software of the existing FFS:

- Possibility to switch to "Batch training" mode;
- Installation and using of alternative instructor's station (IOS-A);

"Batch training" mode

For the increase of automation's degree and exclusion of the majority of operations that instructor performs, each type of training has been formalized in the form of "training programs", which includes a series of "exercise". Each exercise is a part or a whole training flight for which specified:

- All initial conditions (position and condition of the aircraft, the state of the environment);
- All the failures that must occur in flight and the conditions of their occurrence;
- Exercise's completion criteria;

Thus, instructor, who distinctly follows the training program, just starts the next exercise, and after its completion will start the next exercise, or repeat current.

IOS-A

Another necessary step was to develop and connect instructor's station which provided the mode "batch training". Station connects to a host of simulator and the information is displayed on the relatively small monitor (which is installed in the area of instructor's workplace). The entire interface is composed of several frames, including the frame of the selection of Training Program. The main frame is the "Message Output Area", where in the course of exercises appeare text messages about the events, which are related to the training flight and the most useful for assessing by instructor the correctness of effectuation of a crew's flight task. This "instrumental" evaluation provides instructor with information, which is required to produce an instructor's main assessment - psychophysiological load of each crew member.

The Figure 1 shows an example of IOS-A during the exercise, which is aimed to practice takeoff, when an engine fails. Instructor's station screen is divided into the main frame (Message Output Area) , and the two panels (Actual Massage Region and Control Panel). On the main frame messages about situational events of training flight for this particular exercise are showed. Message№5 is the last message that appears at the bottom of the main frame, and after a while climbs up to the previous one (to the Message History Region). All messages are classified

into: errors, warnings, special situations and notifications. Color of the message's text matches to the category of messages.

Panel "Message mask" is intended to ensure, that at the end of exercise instructor has the possibility to filter events according to the categories on the main frame. Control panel has the following capabilities:

1) To suspend the exercise ("Pause" button), thus a time in pause mode is displayed;
2) To repeat the current exercise (button "repeat current");
3) Forcibly stop the exercise, (button "force complete");
4) To move to the next exercise (button "next exercise");
5) To obtain an information about the status of implementation of the program and the name of the exercise (information field on the Control Panel).

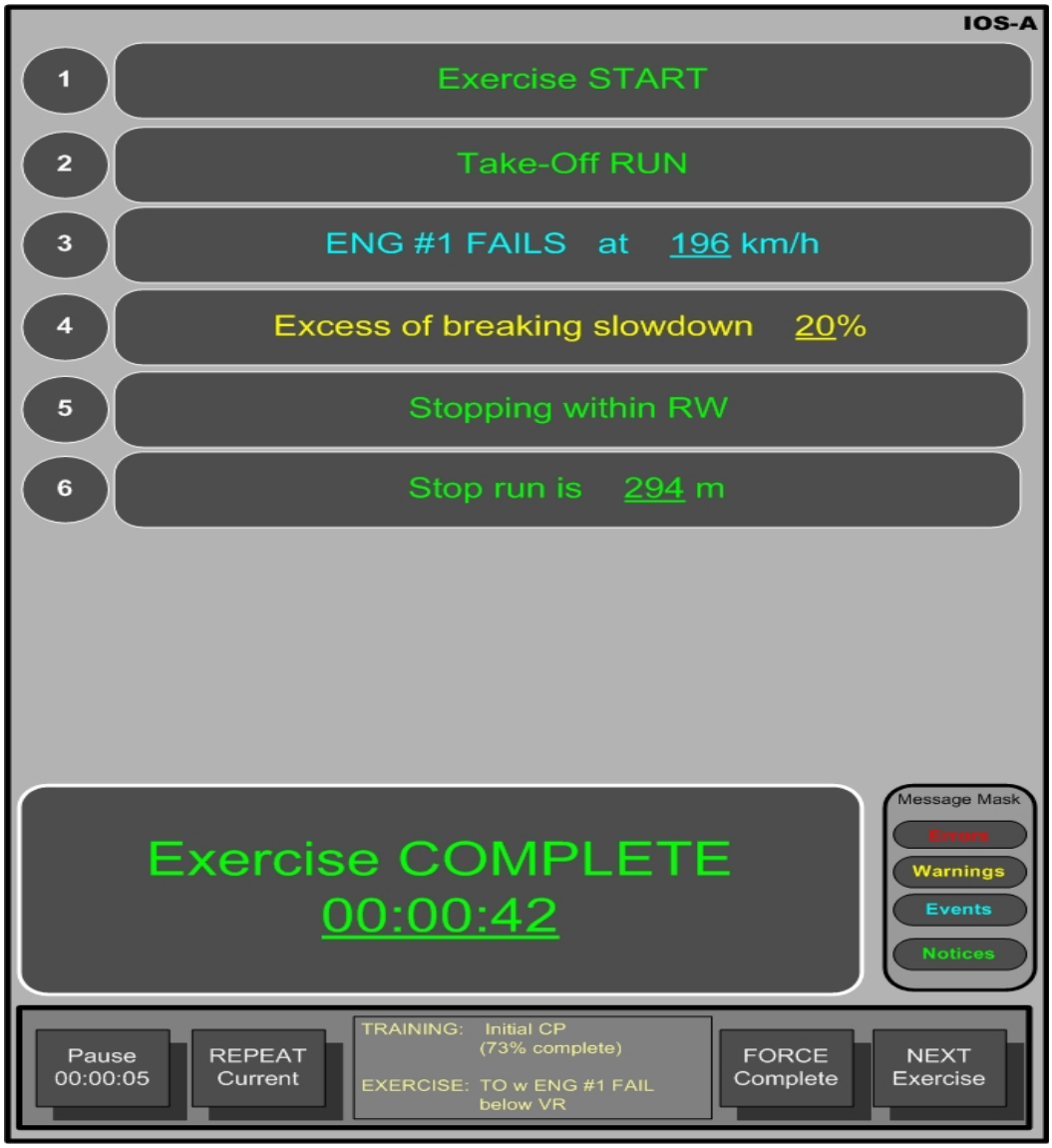


Figure 1. An example of IOS-A.

Conclusion

The base of the “batch training” is currently updated and testing of the developed technique to confirm the hypothesis about enhancing of productivity through reduction of a flow of information to instructor. Initial testing showed the perspective of this approach.

Acknowledgements

Author would like to thank ANTONOV's "Simulation & Training Technologies" (STT) Division and its lead programmer Mr. Oleksandr Sosnenko. The STT together with NAU and other scientific organizations support a lot of researches aimed at improving the pilots training efficiency.

References:

ANTONOV SE (2011). Antonov-148-100 FFS. Operational Manual., Kyiv,.

Gorbunov V.V.(2001). Hygiene of the labor. *The concept of psychophysiological reliability of the human operator and the "man-machine" system in general from the perspective of the human factor. Academy Of The Medical Science Of Ukraine* (pp.175-196).

Gorbunov V.V.(2000). Ergonomic in Ukraine. *Changes of pilot’s cardiac during training flights*. Kiev Military Institute of Management and Communication(pp. 87-91).

History of FFS. Retrieved from Royal Aeronautical Society library: http://www.raes-fsg.org.uk/20/image_gallery/?cat2=2.

Zinchenko V.P., Leonova A.B., Strelkov J.K. (1977). Psychometrics of fatigue. Moscow University (pp.87-91).

CONCEPT OF FLIGHT INSTRUCTOR ASSISTANCE IN HELICOPTER EMERGENCY MEDICAL SERVICE USING PILOT TRAINEE'S WORKLOAD DETERMINATION

Felix Maiwald and Axel Schulte
Universität der Bundeswehr München (UBM),
Institute of Flight Systems (LRT-13), 85577 Neubiberg, Germany
{felix.maiwald, axel.schulte}@unibw.de

This article focuses on the development of a tool chain to support the training of helicopter rescue pilots. The aim is to support the training instructor for comprehensible, objective and reliable assessment of the mental state of pilot trainees. Hence this article investigates a method for on-line estimating the mental workload of the pilot and his free/needed cognitive and sensorimotor resources during flight. We further provide a description of the methodological approach and details on the implemented prototype of a flight instructor station as part of our research simulator. In a first simulator study with four subjects the system has been rated as helpful and effective. A possible application can be found in the more objective evaluation of the pilot students' learning progress. For this purpose the recorded missions are analyzed during debriefing in order to identify workload peaks. Furthermore, the continuous analysis of workload can be used for an on-line adaptation of the training lessons. However both application fields require further development and validation of the methods used in this specific task environment.

Introduction

The training for commercial helicopter pilots CPL (H) is divided into a theoretical part and a practical flight training. In accordance with the rules of the EASA Part-FCL the practical training (at least 30 hours with a flight instructor) comprises the type rating, IFR-training and skill tests. Depending on the intended application purpose, civil helicopter pilots are also trained in rescue missions (Helicopter Emergency Medical Service, HEMS), off-shore or mountain operations. Proficiency checks (e.g. "OPC" and "TRPC") have to be passed by the pilots semi-annually.

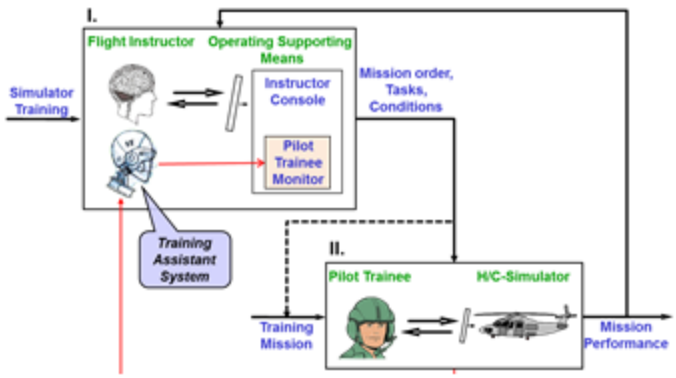


Figure 1. Work system configuration for flight instructor and pilot trainee

In recent years, practical training has been shifted to helicopter simulators more and more. Highly accurate helicopter cockpit replica in conjunction with realistic dynamic simulation of the aircraft can provide great cost savings. Furthermore independence of flight time and outside weather conditions is achieved. Such an integrated air ambulance training center for helicopter pilots (HEMS-Academy) is operated by the German ADAC for instance.

While analyzing simulator training for helicopter pilots the following work processes are of importance: (I) work process of the flight instructor and (II) the work process of the pilot trainee. Subsequently, the mutual dependencies of these processes are revealed with Figure 1 (cf. Onken & Schulte 2010).

The work process of the flight instructor is hierarchically superordinate to the work process of the pilot trainee. Basis for every work process is the specific mission order. Figure 1 also outlines the functional components, mandatory for the execution of the mission order in the two work processes. In the considered use case, the mission order of the flight instructor is made up of simulator training of pilot trainees. For this purpose the flight instructor enters mission orders into his instructor console or manipulates environmental conditions (e.g. weather). The results of his work process are inputs to the pilot trainees work process. This comprises the work objective ("training mission") and other information relating the environmental conditions for the work process of the pilot trainee.

The work process of pilot trainee is then transferred to the more technical view of a work system (c.f. Figure 2). Here, the work system of the pilot trainee is considered under the specific terms of observability by the flight instructor. As a result, the pilot trainee follows the mission order provided by the flight instructor under the given environmental conditions. The overall performance of the pilot trainee work system is measured through the

fulfillment of the work objective and the achieved work results. The performance is heavily dependent to the behavior parameters, i.e. the interactions of pilot trainee with the helicopter simulator. Behavior parameters, e.g. the learning success are available to the training staff by cameras, built-in the training simulator. The observable behavior of the pilot trainee is an expression of his inner states. In this context Pina, Donmez & Cummings (2008) coin the term "behavior precursor". This includes the topics of mental workload (WL), trust in automation and emotional situation of humans.

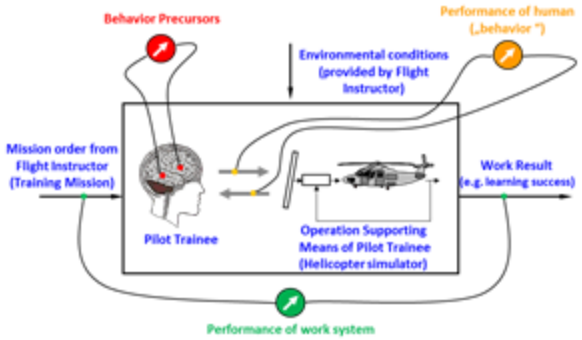


Figure 2. Observability of the pilot trainee work system by the flight instructor (according to Schulte, 2014, Lecture Notes "Flight Guidance & Automation")

One challenge of the flight instructor is the understanding of the hidden internal states of the pilot trainee (behavior precursor). This will be based on the behavior and learning progress of the pilot trainee. Information regarding the workload can be used during training to push the pilot trainee into his mental limitations temporarily. In addition, training schedules are adapted on the basis of the acquired skills and knowledge of the pilot trainee. However, it can be expected difficulties in objective and always comparable assessment of pilot trainee's mental workload from the flight instructors' individual experiences. Therefore a continuous and objective documentation will not be available. Furthermore, behavioral changes (i.e., "self-adaptive strategies"; Sperandio, 1978; Donath, 2012) of the pilot trainee due to high demands have to be interpreted correctly by the flight instructor. Due to these factors, an objective comparison between different pilot trainees and instructors is not possible.

To address this problem, this article focuses on a technical approach to support the flight instructor in pilot state determination (see pilot trainee monitor in Figure 1). The aim is to on-line identify an objective estimation of the current mental workload of the pilot trainee and his remaining resources. The results should be made available to the flight instructor by a "pilot trainee monitor". This "pilot trainee monitor" is a first step towards the development of an instructor-assistant system. Suitable implemented and tested concepts are derived in later sections of this article.

Detailed concept for determining Workload

In the domain of ergonomics, however, there is no standard definition for workload. Following Gopher & Donchin (1986) workload is understood as psychological construct subjectively perceived by the human and therefore not directly measurable. However, there exist different approaches and methods to operationalize workload.

	Question-naires	Analytical approaches	Performance measures	Physiological measures	Behavior analyzes
Proactivity	-	(X)	-	(X)	X
Real-time capability	-	X	X	X	X
Broadbent sensitivity	X	X	-	X	-
No intrusion	(X)	X	(X)	(X)	X

Table 1. *Evaluation of common methods for workload determination*

The workload determination for our pilot trainee monitor should meet the following requirements: (I) Proactivity, able to predict future states of workload, (II) Real-time capability of the measurement, (III) Broadband diagnosis in a wide workload area and (IV) Non-intrusive. An overview of common methods (c.f. summarized in Table 1) is available in Young et al. (2015), Gopher & Donchin (1986), Donath (2012).

Maiwald & Schulte (2014) applied an analytical concept for determination of workload and the mental state of a pilot in military missions. Maiwald (2013) proposes an analytical approach to estimate workload to direct dialogues generated by a pilots' assistant system to the perceptual modality and code, which can be assumed to provide spare resources. The implemented methods provided a broadband sensitivity, high user acceptance and real-time capability. Therefore, we will apply and enhance the method to the domain of civilian helicopter rescue missions (HEMS). The implemented concept is summarized in Maiwald & Schulte (2014). For realization of the pilot trainee monitor we incorporate two models:

1.) *Model of pilot tasks* for the purpose of determining the current tasks of the pilot
2.) *Model of pilot resource consumption* to estimate the resource consumption and WL for current tasks

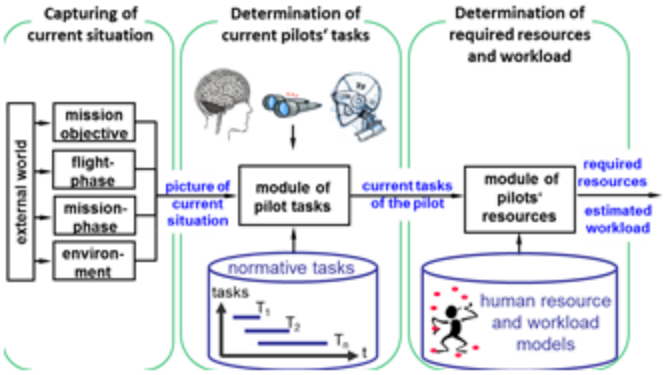


Figure 3: Concept for determination of workload and available/required resources of the pilot trainee

Model of pilot tasks

In the first step, we capture all external influences on the pilot during the HEMS mission (i.e. the state of the helicopter, the mission objective as well as environmental conditions). The flight status is derived from the simulated flight systems and sensors (e.g. navigation). A planning function generates the initial H/C-task agenda, which serves as a basis for the evaluation of the mission progress. This agenda represents a rough mission framework and combines mission relevant tasks with each other. After aggregating all available data into a full situational picture the current tasks the pilot should be executing will be determined. For this purpose, we implemented normative models of mission-typical task situations using state transition networks representing the knowledge acquired in experiments with professional pilots. In a next step we synchronize the tasks described by the static model with the tasks the pilot is actually executing. Therefore, human-machine-interactions such as visual information acquisition (i.e., measures eye fixations) as well as manual interactions are analyzed (cf. Maiwald & Schulte, 2014). In this context, simple models are used to draw conclusions on the tasks actually processed by the human operator from measurements of the eye movements and observations of the manual interactions. Manual interactions taken into consideration are the currently displayed page on the various screens, pushed buttons, current system settings (e.g. landing gear), as well as manual control stick inputs. Visual interactions taken into account are provided by a commercial camera based eye-tracking system (Smarteye®) and its integrated object-related gaze tracking.

Model of pilot resource consumption

Task	information acquisition				processing		reaction	
	VS	*VV*	*AS*	*AV*	*CS*	*CV*	*M*	*V*
Approach to Pickup-Zone	3	2	2	0	2	2	3	0
change zoom on map	1	0	0	0	2	0	1	0

Table 2. *Demand vectors for two sample tasks*

In the next step the actual task(s) are associated with task-specific values of mental resource consumption. Our model of resource demands is based on Wickens' (Wickens & Hollands, 2000) so called multiple-resource theory and describes the required resources by use of eight-dimensional demand vectors (Wickens, 2002). Every demand vector represents the demand a single task poses on the human operator expressed in the terms of information acquisition, information processing and response. Hence, data were gathered through knowledge acquisition experiments, in which helicopter pilots had to rate individual resource demands that arise during the various mission tasks. To eliminate subjective influences from these models as far as possible, laboratory experiments have been conducted to better match the predicted resource conflicts within distinct task situations with the objectively measured pilots' performance (c.f. Maiwald & Schulte, 2014). Table 2 shows an example of demand vectors in detail for the sample tasks "Approach H/C to Pickup-zone" and "Change zoom on map". To estimate the current individual resource utilization, a modified Visual-Auditory-Cognitive-Psychomotor model (VACP; Aldrich & McCracken, 1984) is used. Based on the assumption of a maximum capacity provided by the VACP model a measure of the remaining individual resources of the pilot trainee can be calculated. In addition we look at the resource conflicts which stem from simultaneous task performance to compute the current overall pilots' workload. For this purpose, the demand vectors of the current tasks are fed into a modified workload index model (W/INDEX; Wickens, 2002). The modification we applied to the W/INDEX computation eliminates any limitation on the number of tasks to be examined in parallel (for details c.f. Maiwald & Schulte, 2014).

Preliminary Experimental Testing

A first engineering test has been conducted in our flight simulator to investigate our functional chain predicting the workload und resource utilization of the pilot trainee. The purpose is to gain knowledge how to support the flight instructor by suchlike information. Here we focus on the appropriateness of the implemented methods and possible enhancements of functions in the context of instructor assistance.

Apparatus

Figure 4. Workplace of the flight instructor with three display elements. Left: operator console, center: resource monitor, right: cockpit displays

To test the functional chain, the method has been implemented as a prototype in our generic two-person side-by-side helicopter simulator, used for research projects at the Institute of Flight Systems. Our simulator consists of four multi-function displays (MFDs), each equipped with a multi-touch screen. Depending on the configuration, display formats such as a Primary Flight Display (PFD), a digital map, BOS-transponder status (non-public mobile VHF land mobile service) and pages for radio communication as well as transponder settings can be shown. The pilot is provided a digital map where he enters mission-relevant constraints (e.g. "pickup injured at position X") via touchscreen. Based on this information, the automatic mission planner generates the task agenda by using simple hierarchical task networks. The terrain-conformal route is generated by sample based route planning algorithms (A*-search). Mission specific information such as radio communication and transponder settings may be entered into a Control and Display Unit (CDU). For the simulation of the external environment, a three-channel projection system with a lateral field-of-view of approx. 180° was used. Gaze tracking is realized via four cameras.

The configuration of the prototype workstation for the flight instructor includes the following three displays (cf. Figure 4). The screen on the right position depicts the display formats of the pilot mirrored for evaluation purposes. The resource monitor (center position) represents the utilization of the eight considered resources and the overall workload for a period of 200 seconds. The operator console (left position) shows the telemetry data of the helicopter. It is additionally equipped with the scene camera representing the current visual focus of pilot trainees' information acquisition. Additionally, the instructor is allowed to manipulate mission parameters and individual system parameters of the helicopter. As part of the evaluation the instructor may initiate an engine failure. The telemetry and workload data are recorded and can be replayed by the instructor during debriefing.

Mission

In our scenario, we consider a typical civilian HEMS mission recovering an injured person in the northern Alps in a single pilot configuration. A second trained pilot acts as a flight instructor. At first, the rescue helicopter is located on his base and receives the mission order via voice communication. To keep the mission plan up to date, the pilot has to coordinate with several agencies (e.g. flight information services, land-based emergency services on ground) throughout the mission. Additionally the pilot has to re-plan the mission one or more times (e.g. concerning selection among several suitable hospitals). The overall mission takes about 25 minutes.

Subjects

Three helicopter pilots of the Germany Navy and one rescue helicopter pilot of the ADAC participated in the experimental campaign. The age of the subjects ranged from 25 to 48 with an average of 32 years. The flight experience ranged from 300 to 3500 hours at an average of 1139h with different helicopters (EC135, EC145, BO105).

Hypotheses

The examined scientific questions relate to the following hypotheses:
(1) The implemented functional chain reflects the workload of the pilot.
(2) The implemented functional chain correlates with individual and observable behaviors of the pilot.
(3) The realized instructor station is a valuable tool for pilot training.

As dependent variable we used the predicted workload value and the subjective observations of the experimenter. In addition the manual pilot control inputs ("steering entropy") were used as dependent measure. The assumption to correlate workload with manual steering activity is supported by the research of Nakayama et al. (1999) in the field of vehicle guidance.

Test procedure

Due to the small number of subjects we choose a within subject design for the experiments. Hence, test subjects alternate in the role as flight instructor as well as pilot trainee. To get into routine each subject first conducted a training mission in the Alps. It consists of all elements of the following measurement mission. Landing

in the mountains and emergency procedures for engine failure were rehearsed several times. After completing the training mission, the experimental mission was executed. During flight from the pickup injured to the hospital the flight instructor initiated an unexpected engine failure (independent variable).

Findings

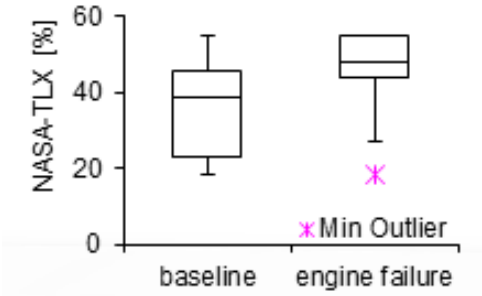


Figure 5. NASA-TLX in baseline and during engine failure

NASA-TLX questionnaires were presented to the pilots for a baseline measure (i.e., during enroute, takeoff, landing) and then for the engine failure condition. Due to inter-individual differences of the workload scales, all NASA-TLX-ratings were normalized. As depicted in Figure 5, pilots rated the baseline with 35.5% workload at the average. In contrast the engine failure condition was rated with an averaged workload level of 48.2%. The increase of workload was proved weak significant by a two side t-test ($t(33)=1.75$, $p=0.0897$, $SD=13.2$, $n_1=26$, $n_2=9$).

In a second step we observed the manual control stick activity of the pilots. During the baseline condition enroute-flight all pilots showed only little control activity (cf. Figure 6). Although pilot 1, 2 and 4 were almost equally experienced, they showed much different control activity during landing and take-off. Huge differences in control inputs are observed under the engine failure condition. In particular, pilot 2 showed a very high control activity. In contrast, pilots 3 and 4 exhibited much less control activity in this situation. To sum up, the experiments revealed individual pilot behavior in comparable situations.

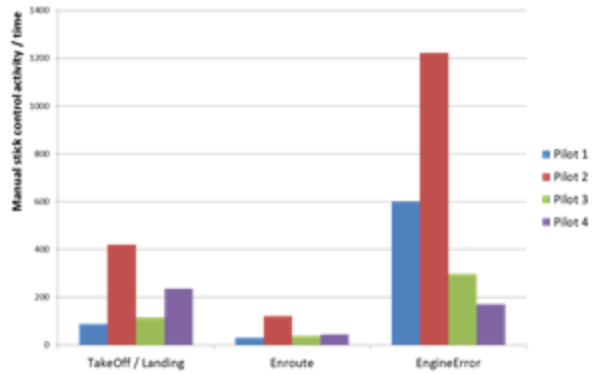


Figure 6. Manual control input of test subjects in examined mission phases

Figure 7 depicts the predicted workload of the pilot during enroute-flight and during engine failure condition. Therefore, one engine had been shut down by the flight instructor at approximately - $t = -150s$. For the examination of hypothesis (1) we compare the relative values of the predicted workload with the experimenter's subjective observations and additionally with the manual stick control activity and the NASA-TLX. As depicted in figure 5, the NASA-TLX revealed an increased workload in engine failure condition. The increase of workload (predicted by our functional chain for the test subjects) between enroute flight and the emergency maneuver "engine failure" (c.f. Figure 7) show a similar characteristic. However, for each test subject a different workload was estimated. Despite the emergency maneuver the estimated workload for pilot 4 is in moderate range. This is consistent with the experimenter's subjective observations of the test person's behavior, because he acted in a very structured way with only a few control inputs. A higher workload was estimated for subject 1 and 3. Nevertheless both subjects performed well the mission tasks. In contrast to this, a very high workload was estimated for subject 2 in the emergency situation. This finding is consistent to the observed behavior of the pilot because he did not respond to auditory communication with the flight instructor in this situation.

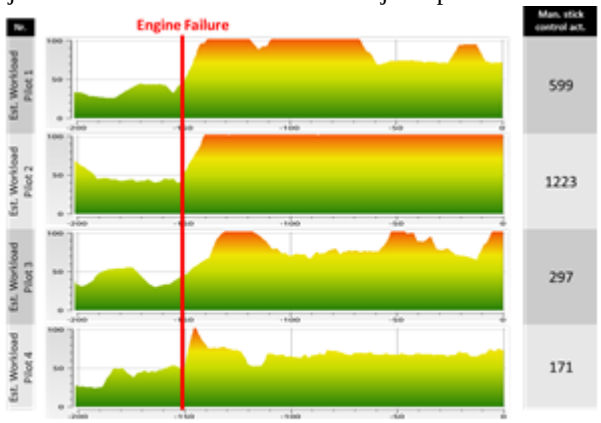


Figure 7. Estimated workload for 4 subjects during enroute with transition to engine failure. Manual stick control activity is excluded in workload determination

The results encourage using individual behavior parameters (manual, visual and auditory interactions) as part of workload prediction. Figure 7 shows a correlation between the manual control activity and the predicted workload. So, in future models the control activity shall be included in addition to the manual, auditory and visual interactions of the pilots. Consequently hypothesis (2) is worthwhile to be further examined.

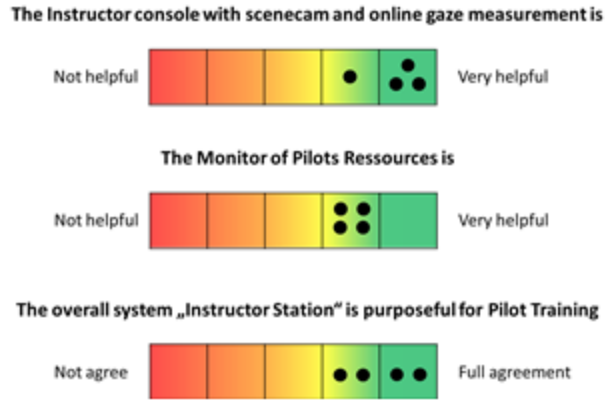


Figure 8. Questionnaires (subjective ratings) of pilots for instructor assistance

Using further questionnaires (cf. Figure 8) the pilots rated the instructor station and integrated tools (e.g. scene camera, workload monitor) as helpful and purposeful for pilot training. They felt well supported in assessing the workload of the pilot. The hypothesis (3) can thus be confirmed.

Conclusions

The implemented system represents an initial approach to assist the flight instructor in the objective and continuous assessment of the pilot trainees' mental state. A possible application is the support of debriefings by use of offline analyses of recorded missions. Thereby, workload peaks could be identified in correlation with specific behavior. Such an assessment could form the basis for future adaptations of workload intensive procedures and may result in improved aviation safety.

Benefit is also expected through the online analysis of workload to optimize the training of pilot trainees. This would enable the training staff to continuously monitor the pilot trainees' mental state and allow the flight instructor to purposeful stimulate the workload (e.g. maxing out trainees).

However, the presented approach requires a further development and validation in the domain of helicopter emergency missions. Our future work will incorporate trials for a profound validation of our resource model prototype, in particular the demand vectors. Also further effort has to be placed in the secure determination of pilots' activity. Here we are investigating on the application of uncertainty theories.

References

ALDRICH, T.B. & MCCRACKEN, J.H. (1984). *A computer analysis to predict crew workload during LHX ScoutAttack Missions*. Vol.1, Fort Rucker, Alabama, US Army Research Institute Field Unit.

BILLINGS, C.E. (1991). *Human-Centered Aircraft Automation: A Concept and Guidelines*. NASA Technical Memorandum 103885. Moffet Field, NASA-Ames Research Center.

DONATH, D. (2012). *Verhaltensanalyse der Beanspruchung des Operateurs in der Multi-UAV-Führung („Behavior Analysis of Workload of multi-UAV Operators").* Dissertation. Universität der Bundeswehr München.

GOPHER, D. & DONCHIN, E. (1986). Workload – An Examination of the Concept. In: BOFF, K.R., KAUFMAN, L. & THOMAS, J.P. (Eds.) Handbook of Perception and Human Performance. Vol.2 (41), pp .1–49. New York: John Wiley.

MAIWALD, F. (2013). *Maschinelle Beanspruchungsprädiktion zur ressourcengerechten Adaption eines Pilotenassistenzsystems („Automatic Workload Prediction for Ressource Adaptive Pilot Assistence").* Dissertation. Universität der Bundeswehr München. (in print)

MAIWALD, F. & SCHULTE, A. (2014). Enhancing Military Helicopter Pilot Assistant System Through Resource Adaptive Dialogue Management. In: VIDULICH, M., TSANG, P. & FLACH, J. (Eds.) *Advances in Aviation Psychology*, Ashgate Publishing Ltd, ISBN: 978-1-4724-3842-3.

MCCRACKEN, J.H. & ALDRICH, T.B. (1984). Analysis of *selected LHX mission functions: Implications for operator workload and system automation goals.* Fort Rucker, Alabama, US Army Research Institute Aircrew performance and Training.

NAKAYAMA, O., FUTAMI, T., NAKAMURA, T., BOER, E. (1999) Development of a Steering Entropy Method for Evaluating Driver Workload. In: *Society of Automotive Engineers Technical Paper Series*: 1999-01-0892

ONKEN, R., & SCHULTE, A (2010). *System-ergonomic Design of Cognitive Automation-Dual Mode Cognitive Design of Vehicle Guidance and Control Work Systems.* Heidelberg: Springer.

PINA, P.E., DONMEZ, B. & CUMMINGS, M.L. (2008). *Selecting metrics to evaluate human supervisory control applications* (No. HAL2008-04). Cambridge, MA: MIT Humans and Automation Laboratory.

SPERANDIO, A. (1978). The regulation of working methods as a function of workload among air traffic controllers. In: *Ergonomics*. Vol.21, No.3, pp. 195–202.

WICKENS, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), pp. 159-177.

YOUNG, M.S., BROOKHUIS, K.A., WICKENS, C.D. & HANCOCK, P.A. (2015). State of science: mental workload in ergonomics. In: *Ergonomics*, Vol.58, No.1, pp.1-17.

CONTRIBUTION OF MULTIMETHODOLOGY TO HUMAN FACTORS IN AIR NAVIGATION SYSTEMS

CABRAL, Lisia Maria Espinola da Silva Pacheco
Brazilian Airport Administration Organization (INFRAERO)
Rio de Janeiro, RJ, Brazil
ESTELLITA LINS, Marcos Pereira
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ, Brazil

This article presents a general view of a post-graduation study developed from 2011 to 2014 into some civil Air Navigation contexts of a brazilian public organization, to promote System and Rational Thinking, and Metagovernance, aiming at structuring, understanding and monitoring problems prone to variability, dynamics and unpredictability, to contribute to minimum risks management and opportunities of changes in real work. The study was developed to support TRM (Team Resources Management) behavior abilities, introduced as a Program since 2009, and adopted: a qualitative and collective method; Multimethodology as a predictive methodology, in which Conceptual Map was the central instrument, based on Soft Operational Research (OR) principles; and complexity paradoxes, in which Metacognition and Selfdeception was the central one. Multimethodology assumed interdisciplinary iterations, interactions and integrations among professionals of different operational activities and organizational levels, applied in four yearly cycles.

In Brazil, Air Navigation and Aviation activities are prescribed by the military Aeronautic Command (COMAER), both for aeronautical accidents and incidents´ safety and investigation areas, civil and military, as follows: Air Navigation ones by the Air Space Control Department (DECEA); and Aviation ones by the Aeronautical Accidents and Incidents´ Investigation and Safety Center (CENIPA). Besides, Civil Aviation activities for civil aeronautical accidents and incidents´ safety and investigation areas are prescribed by the civil National Civil Aviation Authority (ANAC), which is the extinct military Civil Aviation Department (DAC). COMAER (DECEA and CENIPA) and ANAC are brazilian authorities, respectively: aeronautical and civil aviation ones. Both have specific complementary and, sometimes, conflicting standards, in fulfillment to the International Civil Aviation Organization (ICAO) standards, of which Brazil is a member.

Civil Air Navigation activities are developed either by military segments or civil services providers, duly homologated by DECEA, as follows: four military Integrated Air Space Defense and Control Centers (CINDACTA), located at Brazil´s Capitol (Brasília), Manaus (Amazonas), Recife (Pernambuco) and Curitiba (Paraná); one military Regional Protection Flight Service (SRPV), located at São Paulo (São Paulo); some military units (Destacamentos), located all over Brazil; some civil Air Traffic Control and Aeronautical Telecommunications Permitted Stations (EPTA), spread out all over Brazil. These military segments and civil services providers are components of: the Brazilian Air Space Control System (SISCEAB), which DECEA is its central institution; and the Accidents´ Investigation and Safety System (SIPAER), which CENIPA is its central institution.

The brazilian authorities standards take as reference: the Brazilian Aeronautical Code (CBA) (BRASIL, 1986), which has been up-dated for the last years; and the ICAO standards. ICAO´s Safety Management Manual (CANADA, 2013) standard of 2006, up-dated in 2011 and 2013, is fulfilled by all countries´ members to increase, continually, safety in the world, with recommendations for the implementation of: a National Safety Program (PNSO) by the Aviation and Air Navigation authorities; and Safety Systems (SGSO) by each of the services providers, based on PNSO. The Brazilian Operational Safety Program for Civil Aviation (PSO-BR) (BRASIL, 2009) is our PNSO, which is divided in two Specific Operational Safety Programs (PSOE), for the brazilian authorities with the main guidelines to the Civil Aviation and Air Navigation services providers to develop their own SGSO: PSOE-COMAER (DECEA and CENIPA) (BRASIL, 2010); and PSOE-ANAC (BRASIL, 2009a).

Human Factors (CANADA, 1989; BRASIL, 2012) have been responsible to a high contribution in aeronautical accidents and incidents´ occurrences, resulting in the following trainings´ standards: Crew Resource Management - CRM (CANADA, 1989a) for Civil Aviation activities, since 1972; and Team Resource Management - TRM (CANADA, 2008 and 2008a) for Air Navigation activities, since 2002. In Brazil: CRM standards were created in 2003 and up-dated in 2005 by DAC, adopted by ANAC (BRAZIL, 2005); and in 2005, by DECEA (BRAZIL, 2005a), adopted for Civil Air Navigation activities.

CRM and TRM have the common purpose to improve behavior team abilities (CABRAL, 2006): communication assertiveness; situational awareness; stress and health management; team dynamics and leadership; decision process. CRM (CANADA, 1989a) and TRM present two main focus: Error Management - EM; Threat and Error Management - TEM (CANADA, 1989 and 1989a; BRAZIL, 2005 and 2005a). CRM and TRM embodies, mainly, two modalities - conceptual and

practice, which the last one is applied in simulator environments, based on the following standards in Brazil: Line Oriented Flight Training (LOFT) for Civil Aviation activities (BRAZIL, 2005; CANADA, 1989 and 1989a); and practice in simulator environments for Civil Air Navigation activities, still being standardized.

This study is about the implementation of a methodology to deal with complexity (ESTELLITA LINS, 2010 and 2011) in complex systems (ESTELLITA LINS, 2010), to support going parallel, ahead and beyond: TRM (BRAZIL, 2005a and 2012b; CANADA, 2008); and NOSS (BRAZIL, 2012a; CANADA, 2008a).

THEORETICAL REFERENCES

The study was based on the following fundamental concepts, which will be commented in this article:

Complexity Paradoxes (ESTELLITA LINS, 2011) – Group of collective paradigms to increase the understanding of complex systems´ (ESTELLITA LINS, 2010) variability (CANADA, 2002; HOLLNAGEL, 2007), as follows:

Internal (personal subject) X External (common object to all) – In simple systems, the object of analysis gets either into a natural and phenomenal environment, or an artificial and laboratorial isolation submitted to control and manipulation by independent methods, without the need of observers (internal); in complex systems, the object of analysis gets into a permeable and systemic isolation submitted to multi-phenomenal factors, as membranes of each organizational whole, representing individuals´ location in different sets (external).

Processes preservation X Transcendence opened to changes – Simple systems enable to manage behaviors from the functional decomposition of their components in situated actions to preserve their processes; complex systems have self-organizational, emergent and non-analytical properties, accessible and identified, characterizing continuous variations and evolutions to enable transcendence opened to changes.

Isolated parts X Interdependent whole – Simple systems have a functional consistence where functions are restricted to one or a few areas of knowledge, and causal relations among their parts are controlled; complex systems have a multifunctional ambiguity, in which different systems´ components can be grouped as another system and focused by other systems´ components, in different areas of knowledge.

Located information X Distributed systemic information – In simple systems, information nature, power representation and decision making are centralized into production and science, commonly relegated to politic points of view; in complex systems, information is distributed and associated to "hologram" metaphor, in which each part represents an integrated view of the whole, with different degrees of precision, incorporated in constructivism.

Subject indivisibility X Subject multiplicity – Simple systems are characterized by neutrality, professed by an absolute and arbitrary disjunction between the observer and the observed object, requiring a dissociation (dynamics´ suppression) of other inadequate perceptions of the object (indivisibility) ; complex systems (social and productive) require engagement and constant observation of reality, under different points of view, in different, similar, conflicting and complementary activities (multiplicity).

Metacognition (FLAVELL, 1976) X Selfdeception (RUMSFELD, 2012 APUD WAGNER, MURPHY & KORNE, 2012) – According to the Mind Theory (PREMACK & WOODRUFF, 1978), followed by the Theory of Theory (GOPNIK & WELLMAN, 1994 APUD LANGDON, 2005), human mind represents others and ourselves in terms of mental states, therefore real world representation results from individual to social transformations and are associated to human behavior. In simple systems, people, apparently, get to preserve more Metacognition, because of the following processes: propositional, experimental, performing and epistemological dimensions of knowledge, interfering in perception; alternatives to deal with consciousness levels and metacognition; elements analysis to build references and thoroughfares from subjective to inter-subjective attributes; characterization and measurement of subjective attributes as relevant to decision processes (MINGERS, 2006). In complex systems, people tend to give preference to selfdeception and look for the acceptance of other groups not to go against the *status-quo*, and, without perceiving, reinforce certain wrongdoings instead of promoting continuous changes on processes to increase integrity and balance in the whole system. In this case, metacognition needs to be increased with cooperative learning about changes, to allow transcendence from individual to systemic paradigm in all system´s parts.

Unification X Diversification and integration – Simple systems present few differenced characteristics of uniform pattern and routine processes, not requiring much efforts for human cognition to deal with permanent stimulus. Complex systems

present diversification characteristics, requiring creativity, resilience, conflicts management, anticipation and agreement capacity to achieve balance and integration of the differences among components in complex environments.

Systemic Thinking (ACKOFF, 2005; GHARAJEDAGHI, 2011) and Rational Thinking (SENGE, 2008) – This concerns a collective way of thinking about qualitative problems, solutions and possibilities of changes, promoting "iterative loopings" or "iterations", characterized by multiple cycles, and continuous interactions (GHARAJEDAGHI, 2011).

Multimethodology (MINGERS, 2006) and Conceptual Map (ESTELLITA LINS, 2010) – Based on Soft Operational Research (OR) principles (ESTELLITA LINS, 2010; ARÊAS, 2009), it involves a group of instruments used to increase Metacognition (FLAVELL, 1976) about various aspects (material, social, organizational and individual) of real world, considering that only one methodology is very limited to embody complexity (ESTELLITA LINS, 2010 and 2011).

Metagovernance (JESSOP, 2002) – Governance assumes dependent relations in which a centralized and localized power of an upper isolated part of a system regulates what other lower parts have in common. Metagovernance (JESSOP, 2002) assumes interdependent relations with non-hierarchical and spanned coordination in all organization levels, intensifying positive criticism by different perspectives ("requisite variety") for effectiveness of: economic controls; collective purposes; and associated values.

CONTEXT, PARTICIPANTS AND PURPOSE

This article complements another presented at ISAP / 2013 (CABRAL, 2013) of a post-doctoral study (CANADA, 1989; BRAZIL, 2012) developed in five civil EPTA of a public organization of indirect administration, ruled by the Labors Laws´ Consolidation (CLT), homologated by DECEA, as civil Air Navigation service providers. These EPTA are, as well, SISCEAB, SIPAER and SGSO components, in fulfillment to PSOE-COMAER (DECEA and CENIPA) (BRASIL, 2010) and in complement to TRM (BRASIL, 2005a and 2012b). Taking advantage of the need to implement the Psychological Monitoring Program, limited to Air Traffic Controllers - ATC (PTA), the study expanded this focus to the participation of Meteorology Professionals (PMET), Meteorologists (MEG), Aeronautical Information Service´s Professionals (PNA-TIA), Aeronautical Telecommunication Service´s Professionals (PNA-OEA) (BRASIL, 2010b), Airport Safety Professionals (PSA), and leaderships (managers, coordinators and supervisors), embodying the following activities: Air Traffic Control and Monitoring, Meteorology; Aeronautical Information Service and Aeronautical Telecommunication Service (BRASIL, 2010a).

The main purpose of the study was to promote Systemic Thinking (ACKOFF, 2005; GHARAJEDAGHI, 2011), Rational Thinking (SENGE, 2008) and Metagovernance (JESSOP, 2002) towards a collective reflection of operational reality, under different actors´ point of view, in order to structure, analyze and monitor, appropriately, problems focused to common goals in safety operation of the civil Air Navigation EPTA studied. Debate was emphasized to allow, continuously, balance in the complexity paradoxes (ESTELLITA LINS, 2011) mentioned, in complement to: TRM concepts (BRAZIL, 2005a and 2012b; CANADA, 2008); TRM practice in simulator; and NOSS (BRAZIL, 2012a; CANADA, 2008a). This proposition was not limited to quantify and solve organizational problems´ impacts on civil Air Navigation operations, but mainly to enhance a collective and qualitative understanding of emerging threats for adequate risk management, prioritizing Human Factors.

METHOD, METHODOLOGY PHASES AND INSTRUMENTS

This study adopted a qualitative and collective method, and Multimethodology (MINGERS, 2006) as a predictive safety methodology (CANADA, 2013), which did not follow a prescribed approach, but developed its own instruments, applied gradually and cyclical in complex civil Air Navigation contexts (EPTA), to: structure problems; enable perspectives of changes and improvements; complement to TRM (BRAZIL, 2005a and 2012b; CANADA, 2008) and NOSS (BRAZIL, 2012a; CANADA, 2008a); and look for balance in the complexity paradoxes (ESTELLITA LINS, 2011), of which Metacognition (FLAVELL, 1976) and Selfdeception (RUMSFELD, 2012 APUD WAGNER, MURPHY & KORNE, 2012) is the central one.

Multimethodology (MINGERS, 2006) used some instruments, distributed in four yearly cycles, to allow up-dating and structuring problems in real work (VIDAL & MÁSCULO, 2011), by iterative "loopings" and interactions (GHARAJEDAGHI, 2011), with active participation, positive criticism and explicit communication, looking for balance in the complexity paradoxes (ESTELLITA LINS, 2011). The instruments used (Table 1) were: Brainstorm; Simbolization and Simulation; Speeches; Feedback to Leaders; Debates; Group Dynamics; Conceptual Maps (ESTELLITA LINS, 2010) as the central one; and Reports.

Table 1:
Study Phases and Related Instruments of Multimethodology (MINGERS, 2006)

QUALITATIVE AND COLLECTIVE METHOD IN SOFT OPERATIONAL RESEARCH (OR)			
PHASES	**INSTRUMENTS OF PREDICTIVE MULTIMETHODOLOGY IN SAFETY**		
1st. PHASE - Problems´ Awareness, Representation and Consolidation	**2011**		
Awareness	1st. Group Dynamics: "Alphabet"		
	1st. Speech: "Problems´ Definition, Monitoring and Resolution"		
Representation	*Group Exercises*		
	Brainstorm Registration	Simbolization and Simulation Registration	1st. Oral Presentation
Consolidation	*Post-visit*		
	1st. Conceptual Map	1st. Report	Debriefing to Leaders
2nd. PHASE - 1st. Up-date of Problems´ Awareness, Representation and Consolidation	**2012**		
Awareness	2nd. Speech: "Psychology under Human Factors and Ergonomics Focus in Air Navigation Context"		
Representation	*Group Exercises*		
	1st. Debate	2nd. Oral Presentation	
Consolidation	*Post-visit*		
	2nd. Conceptual Map	2nd. Report	

QUALITATIVE AND COLLECTIVE METHOD IN SOFT OPERATIONAL RESEARCH (OR)		
PHASES	**INSTRUMENTS OF PREDICTIVE MULTIMETHODOLOGY IN SAFETY**	
3rd. PHASE - 2nd. Up-date of Problems´ Awareness, Representation and Consolidation	**2013**	
Awareness	2nd. Group Dynamics: "Without Thinking too Much"	
	3rd. Speech: "Problems Monitoring to Find Collective Solutions"	
Representation	*Group Exercises*	
	2nd. Debate	3rd. Oral Presentation
Consolidation	*Post-visit*	
	3rd. Conceptual Map	3rd. Report
4th. PHASE - 3rd. Up-date of Problems´ Awareness, Representation and Consolidation	**2014**	
Awareness	3rd. Group Dynamics: "Your Activity"	
	4th. Speech: "Cognitive Restructuring"	
Representation	*Group Exercises*	
	3rd. Debate	4th. Oral Presentation
Consolidation	*Post-visit*	
	4th. Conceptual Map	4th. Report

ANALYSIS AND CONCLUSION SUMMARY

Based on the complexity paradoxes (ESTELLITA LINS, 2011), here will be presented a general analysis of the study:

Internal (Personal Subject) X External (Common Object to All) – There is a trend to individual and personal interests and benefits, reinforcing the isolation of systems´ parts, as an obstacle to achieve common goals of the whole system´s planning and operation, for instance: medical orders optimizing work absence as a compensation of continuous hard shift working ; training, workload distribution and vacations planning based on personal criteria by some managers. This points out to internal and subjetc criteria in place of external and systemic ones, contributing to a negative organizational climate and poor communication.

Processes of Preservation X Transcendence Opened to Changes – There is a Quality Program to certificate EPTA after periodic audits, which non-compliances are corrected by pressure, due to the need of attempting prescribed procedures and quantitative references, distant to operational difficulties. Multimethodology (MINGERS, 2006) came up with a qualitative and collective method to evidence emergent problems in real work (VIDAL & MÁSCULO, 2011), although there is a disproportional slowness of upper organizational levels (Governance) to respond to their demands. Some examples involve inappropriate physical and cognitive operational conditions in terms of equipment, material and personnel in all kinds of air traffic plan, control and management. This emphasizes a trend to processes´ preservation, proper of public organizations, in place of structures, functions and processes´ transcendence to provide changes, solutions and improvements.

Isolated Parts X Interdependent Whole – Initially, in the first phases of the study, participants of different activities or of the same activities but different working groups, showed unfamiliarity among themselves, tending to add value to its own context. As well, there were some operators physically located nearby others, but working in different activities, who had never talked to each other. So, different positions´ relationships and diverse activities´ routines used to be addressed separately, as isolated parts of the same systems (SIPAER and SISCEAB), in spite of considering common safety goals. Besides, the rapport among EPTA´s operators and leaderships with actors of Aeronautical Authority (CENIPA and DECEA) presented hierarchical and non-systemic characteristics, based on poor communication and interaction. This kind of attitude was less issued in TRM (BRASIL, 2005a and 2012b) than in Multimethodology (MINGERS, 2006) because Conceptual Map (ESTELLITA LINS, 2010) showed contribution to promote Systemic and Rational Thinking (ACKOFF, 2005; GHARAJEDAGHI, 2011; SENGE, 2008), although not enough to advance from military values of dependence to Metagovernance (JESSOP, 2002), based on interdependence.

Located Information X Distributed Systemic Information – Multimethodology (MINGERS, 2006) with emphasis on Conceptual Map (ESTELLITA LINS, 2010), in complement to TRM (BRASIL, 2005a and 2012b), helped to distribute systemic information, although there is still a centralized bureaucratic culture involving located information, linear and dependent interactions, and emphasis on parts more than on the whole system. This contributes to emphasize loss of self-confidence, motivation, confidence on the organization, development of a fragile commitment from top to bottom organizational levels, in place of focusing to effective integration on relationships involving common structures, functions and processes. Some examples are: privatization of airports; Human Factors´ abilities devaluation; deficit of operators and hard shifts working; linear and confuse communication; organizational pressures to attend prescribed standards trading-off deep problems on operational reality.

Subject Indivisibility X Subject Multiplicity – There is a trend to passive and complacent behavior to accept organizational pressures and bureaucratic prescriptions, not taking chances to develop subject multiplicity in order to face possible changes focused to systemic and common goals. To change this paradigm, in complement to TRM (BRAZIL, 2005a and 2012b; CANADA, 2008) prerogatives, the study´s iterations and interactions reinforced cooperation, assertiveness in communication, team dynamics and flexible parameters related to system´s latent failures and individual´s active errors, in place of culpable and competitive parameters. Promoting subject multiplicity still remains a challenge to be reached continuously.

Metacognition (FLAVELL, 1976) **X Selfdeception (RUMSFELD, 2012 APUD WAGNER, MURPHY & KORNE, 2012)** – This complexity paradox (ESTELLITA LINS, 2011) involves the others and was permanently debated in the study in order to: optimize individual and group perception of different human needs and points´ of view; understand operational reality from a systemic dimension of structures, functions and processes; and increase metacognition in place of selfdeception.

Unification X Diversification and Integration – Gradually, the study led to interactions among different parts in terms of different operations, activities and positions, which enabled to exchange knowledge and experience, ideas and propositions, aiming at diversification on collective learning. Integration was observed in lower intensity, once it requires a continuous maturity up-grade in all organizational levels, based on Metagovernance (JESSOP, 2002), which wasn´t implemented, as desired.

Multimethodology (MINGERS, 2006) as a predictive methodology, with emphasis on Conceptual Map (ESTELLITA LINS, 2010), based on a collective and qualitative method, seemed to succeed on achieving the main purpose of the study: to promote Systemic Thinking (ACKOFF, 2005; GHARAJEDAGHI, 2011) and Rational Thinking (SENGE, 2008) towards a collective reflection of operational reality focused to common safety goals, enabling to structure and monitor problems. Balance in complexity paradoxes (ESTELLITA LINS, 2011) is a continuous process needed according to each context to stimulate which Metacognition (FLAVELL, 1976). Metagovernance (JESSOP, 2002) still needs future studies. Detailed description of this study may be found in its doctoral thesis and related papers.

References

ACKOFF, Russell Lincoln (2005). Thinking about the Future. Transcript of the talk given at the Tällberg (Sweden), Forum. In: http://ackoffcenter.blogs.com/ackoff_center_weblog/2014/01/thinking-about-the-future.html; http://ackoffcenter.blogs.com/files/ackoffs-tallberg-talk-doc-copy-1.pdf.

BRASIL (1986). Brazilian Aeronautical Code [Código Brasileiro de Aeronáutica - CBA, Lei nº 7.565, Presidência da República].

BRASIL, National Civil Aviation Agency - ANAC (2005). Civil Aviation Instruction Instruction [Instrução de Aviação Civil - IAC] 060-1002A: Corporate Resource Managing - *CRM* [Gerenciamento de Recursos de Equipes].

BRASIL, Air Space Control Departament - DECEA (2005a). Aeronautical Command Instruction [Instrução do Comando da Aeronáutica - ICA] 37-228: Team Resource Management - TRM.

BRASIL, Aeronautical Command - COMAER and National Civil Aviation Agency - ANAC (2009). Brazilian Operational Safety Program [Programa de Segurança Operacional - PSO-BR]. Joint Act [Portaria Conjunta] nº 764 / GC5, COMAER and ANAC.

BRASIL, National Civil Aviation Agency - ANAC (2009a). Specific Safety Operational Program of National Civil Aviation Agency [Programa de Segurança Operacional Específico - PSOE-ANAC], ANAC.

BRASIL. Air Space Control Departament - DECEA (2009b). Aeronautical Command Instruction [Instrução do Comando da Aeronáutica - ICA] 63-22: Vigilance of Operational Safety Program for Air Navigation Services [Vigilância do Programa de Seguamça Operacional para os Serviços de Navegação Aérea].

BRASIL. Aeronautical Command - COMAER (2010). Specific Safety Operational Program of COMAER [Programa de Segurança Operacional Específico - PSOE-COMAER]. Act of COMAER [Portaria do COMAER] nº. 368 and COMAER Bulletin [Boletim do COMAER] nº. 108.

BRASIL, Air Space Control Departament - DECEA (2010a). In: http://www.redemet.aer.mil.br.

BRASIL, Brazilian Airport Infra-structure Organization - INFRAERO (2010b). In: http://www.infraero.gov.br.

BRASIL. Air Space Control Departament - DECEA (2012). Aeronautical Command Instruction [Instrução do Comando da Aeronáutica - ICA] 63-15: Human Factors in Operational Safety Management of Brazilian Air Space Control Sysetm [Fatores Humanos no Gerenciamento da Segurança Operacional do Sistema de Controle do Espaço Aéreo Brasileiro - SISCEAB].

BRASIL. *Air Space Control Departament* - DECEA (2012a). Aeronautical Command Instruction [Instrução do Comando da Aeronáutica - ICA] 63-14: Risk Management to Operational Safety Manual [Manual do Gerenciamento de Risco em Segurança Operacional].

BRASIL. Brazilian Airport Infra-structure Organization - INFRAERO (2012b). Proceedures Manual [Manual de Procedimentos - MP] 16.11 (NAE): Team Resource Management - TRM [Gerenciamento de Recursos de Equipe].

CABRAL, Lisia Maria Espinola da Silva Pacheco (2006). Dissertation of Master Post-graduation Course: A Methodological Proposition for Training of Crew´s Behavior Based on Computational Games [Proposta de Metodologia para Terinamento Comportamental de Equipe Baseada em Jogos Computacionais]. Federal University of Rio de Janeiro - UFRJ, Alberto Luis Coimbra Post-graduation and Research Institute in Engineering - COPPE, Civil Engineering Program.

CABRAL, Lisia Maria Espinola da Silva Pacheco (2013). Article: The Main Characteristics of the Use of Multimethodology and Concept Maps for Structuring, Analyzing and Monitoring Problems and Decision Making Processes at Public Organizations - A Practical Example. ISAP.

CANADA. International Civil Aviation Organization - ICAO (1989). Doc 9683 / AN 950: Human Factors Training Manual. In: http://www.icao.int/publications.

CANADA. International Civil Aviation Organization - ICAO (1989a). Circular 217 / AN 132: Human Factors Digest nº. 2 - Flight Crew Training: Cockpit Resource Management (CRM) and Line-Oriented Flight Training (LOFT).

CANADA. International Civil Aviation Organization - ICAO (2002). Doc 9806 / AN 763: Human Factors Guidelines for Safety Audits.

CANADA. International Civil Aviation Organization - ICAO (2008). Doc 314 / AN 178: Threat and Error Management - TEM in ATC.

CANADA. International Civil Aviation Organization - ICAO (2013). Doc 9859 / AN 474: Safety Management Manual. In: http://www.icao.int/safety/SafetyManagement/Documents/Doc.9859.3rd%20Edition.alltext.en.pdf.

CANADA. International Civil Aviation Organization - ICAO (2008a). Doc 9910 / AN 463: *Normal Operation Safety Survey* - NOSS.

GHARAJEDAGHI, J. (2011). System Thinking, Managing Chaos and Complexity - A Platform for Designing Business Architecture: Chapter 7 - Design Thinking. Ed. Elsevier, USA.

HOLLNAGEL, Erik (2007). Safer Complex Industrial Environments: Chapter 3 - Extending the Focus of Human Factors. CRC Press, NY and London.

ESTELLITA LINS, Marcos Pereira; ANTOUN NETTO, Sérgio Orlando; BISSO, Cláudio R. S. (2010). Written *Annotation* [Apostila]: Social Complex Problems: Structuring through Conceptual Maps [Estruturação de Problemas Sociais Complexos com Mapas Conceituais]. Federal University of Rio de Janeiro - UFRJ, Alberto Luis. Coimbra Post-graduation and Research Institute in Engineering - COPPE, Doctoral Course of Operational Research in Production Engineering Program - PEP.

ESTELLITA LINS, Marcos Pereira (2011). Lecture [Palestra]: Complexity Paradoxes in Structuring Problems Methods [Paradoxos da Complexidade em Métodos de Estruturação de Problemas, UFRJ, COPPE, Doctoral Course of Operational Research in PEP.

FLAVELL, J.H.(1976). Metacognitive Aspects of Problem Solving. In: L. B. Resnick (Ed.), The Nature of Intelligence (pp. 231- 235). New Jersey, NJ: Lawrence Erlbaum.

SENGE, Peter (2008). The Fifth Discipline. Ed. Best Seller.

VIDAL, Mario Cesar; MÁSCULO, Francisco Soares (2011). Ergonomics: Adequate and Efficient Work [Ergonomia: Trabalho Adequado e Eficiente]. Ed. Elsevier.

WAGNER, Daniel A.; MURPHY, Katie M.; KORNE, Haley De (2012). Learning First: A Research Agenda for Improving Learning in Low-Income Countries - Working Paper 7. Graduate School of Education, University of Pennsylvania, Center for Universal Education at Brookings.

A VALID AND RELIABLE SAFETY SCALE FOR PASSENGER'S PERCEPTIONS OF AIRPORT SAFETY

Stephen Rice, Florida Institute of Technology
Rian Mehta, Florida Institute of Technology
Scott Winter, Florida Institute of Technology
Korhan Oyman. Florida Institute of Technology

Previous research has developed various customer satisfaction scales in many applied areas; however, to our knowledge, there is not a validated scale for measuring commercial airline passengers' ratings of personal safety based on airport security. The current study seeks to address this missing gap by developing a valid and reliable safety scale for commercial airline passengers (SS-CAP). We first solicited words and phrases that are related to a passenger's feeling of safety from potential consumers and experts in the field. We then narrowed down the list to 7 remaining items. Lastly, we tested the scale using participants from Amazon's ® Mechanical Turk ®, which is reliable source for participants in online surveys. A principle components factor analysis with varimax rotation revealed that all items loaded strongly on one factor, accounting for 78% of the variance in the model. A Cronbach's alpha test revealed high internal consistency, $r = 0.95$. A Guttman split half test showed high reliability, $r = 0.95$. These results provide strong evidence for a valid and reliable scale of passenger ratings of personal safety. The scale statements include: I feel safe, I feel secure, I feel protected, I feel guarded from danger, I feel shielded from harm, I feel at ease, I feel sheltered from threats. Participants should respond on a 5-point Likert type scale scored from strongly disagree (-2) to strongly agree (+2).

Introduction

Safety is of the utmost importance, especially in high consequence industries such as aviation (Janic, 2000; Maurino, 2000; Sarter & Alexander, 2000). As aviation continues to grow, countries are investing in developing, expanding, and building new airports. The Atlanta Hartsfield-Jackson airport recently opened a fifth runway and Abu Dhabi and Istanbul are investing in the development of new airports. As this growth continues so must the safety record within this industry. Numerous authors (Patankar & Sabin, 2010; Sarter & Alexander, 2000) express concern that this growth could result in an increase in accidents.

Safety is defined as being "free of harm" and "the state of being safe" (Merriam-Webster, n.d.). Safety is also framed within the social construct (Maurino, 2000) and is related to risk. What is deemed safe or risky in one society may be outside the tolerances of another. Therefore, safety is somewhat of a continuum, but there is no question that in high consequence industries, such as aviation, the margin of error is extremely low.

Airports play a key role in the aviation system. The majority of all flights originate and arrive at airports. These facilities serve as the backbone of a large network of flights were passengers and cargo are loaded and aircraft are maintained. The complex maze of taxiways and runways must provide for safe movement of aircraft, and if a mistake is made, the results could be catastrophic. Of interest in the current study was to develop a valid and reliable scale that could be used to measure consumer's perceptions on airport safety.

Human error remains the leading cause of safety infractions in aviation. Experts estimate that as many as 70% to 80% of accidents are attributed to human error (Sarter & Alexander, 2000). However, it is possible that accidents are the results of compounding issues. Poor airport design or layout could compile the human related factors. Many airports are now being designed to reduce the number of possible locations where a runway incursion (possible collision or loss of separation) could occur. Recent studies have focused on the measurement of safety management systems and safety culture (Gill & Shergill, 2004; von Thaden & Gibbons, 2008), locus of control (Hunter, 2002), and commercial aviation safety culture (von Thaden, Wiegmann, Mitchell, Sharma, & Zhang, 2003).

Studies by Gill and Shergill (2004), von Thaden and Gibbons (2008), and von Thaden, Wiegmann, Mitchell, Sharma, and Zhang (2003) reviewed safety management and views toward safety culture. However, these studies were focused on the perceptions of operators within the system and not the consumers of these processes. Operators have more extensive levels of training within the system and therefore different perspectives. Hunter (2002) studies the measure of locus of control. Locus of control is defined as the level a person believes that the outcomes of certain situations are under their personal control. Hunter's measure was pre-existing and adjusted to be a valid and reliable instrument that could be used in the aviation field.

While these studies all provide for accurate measures of safety by operators, there appears to be a gap in the literature related to 1) consumer perceptions of safety and 2) their views toward the safety of airports. Consumer perceptions can have powerful influence within aviation. Many consumers did not like flying on turboprop aircraft, and now few of them are used for commercial transportation. In fact, some consumer ticketing websites have checkboxes to remove those options that include turboprops as part of the flights (e.g. Kayak). Therefore, having a valid and reliable instrument to measure consumer perceptions towards airports may provide valuable insights into those items that are most important to consumer's views toward airport safety.

Methodology

Stage 1: Word Generation

In Stage 1, we began generating items for the scale. We first solicited words and phrases from experts in the field and similar scales in the literature. We then selected participants for an online survey in order to generate items from people who might actually use the scale in the future. Since consumers are the ones who will be responding to the scale, we felt that this increased the validity of the process.

Participants. Seventy-two (28 females) participants from the United States were recruited via a convenience sample using Amazon's ® Mechanical Turk ® (MTurk). MTurk provides participants who complete human intelligence tasks in exchange for monetary compensation. Prior research shows that data from MTurk is as reliable as normal laboratory data (Buhrmester, Kwang, & Gosling, 2011; Germine, et al., 2012). The mean age was 32.98 (*SD* = 9.35). Three additional participants with expertise in safety and/or security provided further items for review. Lastly, the trust literature was reviewed, and items were added accordingly.

Materials and Stimuli. Participants were presented with the following scenario: "*In the context of a commercial airport, please enter 5 characteristics of safety in the spaces provided below. Each answer should include only one word or short phrase.*" After providing the list of 5 words or phrases, participants were debriefed and dismissed. This exercise generated 232 unique words or phrases. These items were then reviewed for correct spelling and de-capitalized when necessary to ensure uniformity.

Stage 2: Nominal Paring

In Stage 2, we began eliminating words or phrases that were not perceived by participants as being related to the construct of safety as it relates to a commercial airport.

Participants. Forty-nine (23 females) participants from the United States were recruited via a convenience sample using Amazon's ® Mechanical Turk ® (MTurk). The mean age was 32.65 (*SD* = 11.26).

Materials and Stimuli. Each of the 232 items generated in the first stage were presented to participants, along with the following statement, "*In the context of a commercial airport, please rate whether each word below is related to (similar to) safety, not related to (not similar to) safety, or you don't know.*" Forty-two items were chosen to be related to trustworthiness by at least 85% of participants.

Stage 3: Likert-scale Paring

In Stage 3, we continued narrowing down the list of items that would be retained for the final scale. Looking for a more sensitive measure of the relationship between the items and safety, we used a Likert-type scale instead of a nominal scale.

Participants. Forty-eight (22 females) participants from the United States were recruited via a convenience sample using Amazon's ® Mechanical Turk ® (MTurk). The mean age was 32.23 (*SD* = 10.59).

Materials and Stimuli. The 42 items retained from Stage 2 were presented to participants with the following statement, "*In the context of a commercial airport, please rate how strongly each word below is related to safety.*" Participants responded based on a Likert-type scale from "Not at all related to safety" (0) to "Extremely related to safety" (+3). Seven items averaged 2.5 or higher and were retained for the final scale.

Stage 4: Scenario-based Testing

In Stage 4, we began collecting evidence of validity and reliability for the new scale. The seven items were converted into statements that could be rated on a 5-point Likert-type scale from strongly disagree (-2) to strongly agree (+2).

Participants. Two hundred and twenty-two (83 females) participants from the United States were recruited via a convenience sample using Amazon's ® Mechanical Turk ® (MTurk). The mean age was 31.09 (*SD* = 10.55).

Materials and Stimuli. In this stage, participants were presented with the following scenario: "*Please try to remember the last commercial airplane flight that you flew on. Think about the airport security that you interacted with. In the context of the airport security described above, please respond to the following statements to the best of your ability.*" Participants were presented with the questionnaire (see Appendix A) and asked to provide statements of agreement or disagreement on a 5-point Likert-type scale (coded from -2 to +2).

Scale Development. A factor analysis using the principle components and varimax rotation resulted in all items strongly loading on one factor. A Cronbach's Alpha test was conducted to measure internal consistency within the scale. The resulting coefficient of 0.95 indicated high internal consistency. A Guttman split-half test was conducted as well. The resulting coefficient of 0.95 indicated high reliability.

Discussion

The purpose of this study was to create a valid and reliable instrument for obtaining airline consumer ratings of personal safety based on airport security. Human beings value their personal safety and security, and is oftentimes of the highest priority when making decisions. This remains true when passengers consider which airports they use for their travels. If a passenger believes an airport has a lower level of security as compared to a neighboring airport, it can be assumed that they would choose the safer of the two options. The consumers' perception of the safety factor of an airport can have a significant influence on airport traffic. An additional purpose of this research was to fill a void in the aviation related literature regarding a valid and reliable measure that can be used to capture consumer perceptions on airport safety.

Only positively scored items related to aircraft safety were enlisted to develop the scale. In doing so, it prevents the need for the participant to cognitively switch between negative and

positive words. This in turn eliminates the possibility of cognitive confusion. Research has indicated that a negative psychometric effect could be observed with the use of reverse scored items (Harrison & McLaughlin, 1991).

The creation of a valid and reliable scale is of practical value to the airport operations field as well as the research community. The study fills a gap by developing a metric that measures a vital airport consideration. The additional benefit lies in the fact that the developed scale is has been statistically proven for validity, reliability, and discriminability. While knowledge of the inner working of the industry are important, it is relevant to note that this scale was developed using words generated by consumers and not industry experts. A consumer perception scale specifically for airport security could be of valuable assistance to most airport management staffs across the country. Additionally, the creation of such a scale lays the foundation for future scales to be developed, within the realm of security, as well as for other airport facets.

While it is beneficial to create a consumer scale using actual consumers, the means of data collection have certain limitations. Each stage of the generation and statistical analysis phase use responses generated by participants for compensation from Amazon's ® Mechanical Turk ® (MTurk). Research by Buhrmester, Kwang, and Gosling (2011) state that this form of data is a reliable as laboratory data. Since the recruited participants were only recruited from the United States, the generalizability of the findings is limited to U.S. airports and consumers. Future steps along this line of research could seek to field test the instrument in person at airports around the United States. Lastly, further studies could seek to better understand the relationship, if any, between the frequency of a consumer's travels and their perception of airport security.

Conclusion

The aim of this research endeavor was to fill a gap in scientific community by creating a valid and reliable scale to measure consumer perceptions of airport security. In doing so, it allows airport operators to utilize such an instrument in order to better understand, and perhaps address, certain concerns or misconceptions of the passengers. Consumer responses through a multi-stage process using word generation, word paring, and scenario-based examples were used to create the instrument. It is the hope that this scale could be used as a tool for data collection on perceptions of airport security, and enable improvements in that sector.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(3), 3-5.

Gill, G. K., & Shergill, G. S. (2004). Perceptions of safety management and safety culture in the aviation industry in New Zealand. *Journal of Air Transport Management, 10*, 233-239.

Harrison, D. A. & McLaughlin, M. E. (1991). Exploring the cognitive processes underlying responses to self-report instruments: Effects of item content on work attitude measures. *Proceedings of the 1991 Academy of Management annual meetings*, 310-314.

Hunter, D. R. (2002). Development of an aviation safety locus of control scale. *Aviation, Space, and Environmental Medicine, 00*(0), 000-00.

Janic, M. (2000). An assessment of risk and safety in civil aviation. *Journal of Air Transport Management, 6*, 43-50.

Maurino, D. E. (2000). Human factors and aviation safety: What the industry has, what the industry needs. *Ergonomics, 43*(7), 952-959.

Merriam-Webster (n.d.) In Merriam-Webster online dictionary. Retrieved from: http://www.merriam-webster.com/dictionary/safety

Patankar, M., S., & Sabin, E. J. (2010). The safety culture perspective. In Salas, E. & Maurino, D. (Eds.), *Human Factors in Aviation*, (95-122). Burlington, Elsevier.

Sarter, N. B., & Alexander, H. M. (2000). Error types and related error detection mechanisms in the aviation domain: An analysis of aviation safety reporting system incident reports. *The International Journal of Aviation Psychology, 10*(2), 189-206.

von Thaden, T. L., & Gibbons, A. M. (2008). The safety culture indicator scale measurement system (SCISMS). Human Factors Division Institute of Aviation. Savoy, Illinois.

von Thaden, T. L., Wiegmann, D. A., Mitchell, A. A., Sharma, G., & Zhang, H. (2003). Safety culture in a regional airline: Results from a commercial aviation safety survey. Proceedings of the *International Symposium on Aviation Psychology*, Dayton, Ohio.

Appendix A

Please respond to each of the following statements:

I feel safe.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel secure.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel protected.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel guarded from danger.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel shielded from harm.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel at ease.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

I feel sheltered from threats.
Strongly Disagree DisagreeNeutral Agree Strongly Agree

AGE AND TRUST IN FLIGHT ATTENDANTS: A COMPARISON BETWEEN TWO COUNTRIES

Rian Mehta, Florida Institute of Technology, Melbourne, Florida
Natasha Rao, Florida Institute of Technology, Melbourne, Florida
Ethan Labonte, Florida Institute of Technology, Melbourne, Florida
Stephen Rice, Florida Institute of Technology, Melbourne, Florida

It is important for passengers to trust their flight attendants, especially in case of an emergency. There is ongoing debate in India regarding trust and the lowering of retirement age, which is currently mandatory at 58 years in flight attendants. Some believe this is in order to acquire younger, more attractive flight attendants. The current study asked 384 Indians and Americans to rate their trust in flight attendants based on an emergency situation. Results showed that Indians trusted the younger flight attendants (25 years old) more than their older counterparts (55 years old). These findings have theoretical and practical implications in the ongoing debate. A future intention of the researchers is to conduct a mediation analysis to determine if affect is a possible mediator.

This study looks to analyze the consumer's trust in flight attendants. It analyzes cross-cultural data from India and the U.S. Increases in scheduled flights have caused a shortage in the number of flight attendants. Passengers have the closest interaction with their flight attendants and therefore, understanding consumer perceptions is of utmost importance.

Today, the retirement age in India is 58 years, while the United States has no age restriction. Whereas, In 1954, women who were hired as flight attendants were required to sign a cont ract stating that they would be required to retire at 32 years of age (Lessor, 1984). Culture is one of the factors that helps us explain these differences. (Chen and Staroata, 1998) defined culture as "pattern of shared basic assumptions of society according to national, organizational, regional, ethical, religious, linguistic, and social characteristics". Hofstede (1980, 2001) found that Thai, Chinese, and Indian cultures displayed high levels of collectivism. The difference between India as collectivist and the United States as individualistic has been acknowledged and documented by previous research (Markus & Kitayama, 1991). India is in general a collectivistic culture but may also exhibit individualistic features (Rice et al. 2014). Collectivists are taught to trust. Persons born into individualistic cultures however, are taught to be independent of one another and to not trust without questioning (Han & Shavitt, 1994).

Other factors that help us understand tendencies towards aging flight attendants, is age and gender. Even though gender is a factor in this study, the effect of age is more pronounced. Prejudice on the basis of age is known as ageism. It has not been found that an individual's job performance can be predicted based on their age (Cleveland & Landy, 1983). Meta-analytical reviews have shown that older adults are devalued and perceived to have lower competence (Kite & Johnson, 1988; Kite, Stockdale, Whitley, & Johnson, 2005). When reviewing previous research on gender differences, it is evident that cultural differences influence gender perceptions. Jackman and Senter (1981) found that 78 percent of men believed that women are not emotionally equal to men. Some gender stratification analyses have cited son preference and the

low position of women, as primary contributing factors to the discrimination against females (Arokiasamy, 2004), this is due to India's patriarchal nature.

Trust has an important role to play in this study. For the purpose of this study trust is defined as "expectation of technically competent role performance" (Barber, 1983, p. 14). Studies suggest that stigmas could have an effect on trust. As per the social identity theory, people are biased towards their 'in-group', or the group that they identify themselves within.

Current Study

The study aimed to understand the trust dynamic between passengers and flight attendants. This relationship is an under-analyzed sector of the scientific community. Participants were asked to respond to a hypothetical situation. Age and gender of the flight attendants were manipulated as a part of the hypothetical situation and ratings of trustworthiness were collected from Indian and American participants. The research predictions were as follows:

1) There would be differences in trust ratings based on the country of origin of the participant.
2) There would be differences in trust ratings based on the age of the target flight attendant in the scenarios.
3) There would be differences in trust ratings based on the gender of the target flight attendant in the scenarios.

Methods

Participants: Three hundred and eighty-four (135 females) from India and the United States participated in this study and the number of participants per country was equal. The mean age was 31.06 (SD = 7.19). The mean ages did not differ as a function of country ($p > .10$).

Procedure, Materials and Stimuli: First, the participants were asked to fill out a consent form and given instructions. A hypothetical question about a commercial airline was asked. The participants were told that the flight attendants were 25 and 55-year-old male and 25 and 55-year-old female followed by questions such as "How do you feel", "How much do you trust the flight attendant in an emergency situation?" and "How trustworthy do you think this flight attendant would be in an emergency situation?" The responses were collected along three different Likert-type scales from Extremely negative/unfavorable/bad to Extremely positive/favorable/good and Extremely negative/unfavorable/bad to Extremely positive/favorable/good. There was a zero neutral option for each scale. The participants were then asked for demographic information, debriefed and dismissed.

Design. A three-way between-participants factorial design was employed, whereby the three independent variables were: 1) Age of the flight attendant; 2) Gender of the flight attendant; and 3) Country of origin of the participants. The dependent variable was the participants' trust scores.

Results

A Cronbach's Alpha test was conducted on the data which was combined for further analysis due to high scores of internal consistency for Trust (*r* range from = .84 to .97). Trust data was subjected to a three-way ANOVA with Age and Gender of the flight attendant, and Country of origin of the participants as the factors. A main effect of Age, $F(1, 376) = 5.05$, $p = .025$, *partial-eta squared* = .01, and Country, $F(1, 376) = 36.25$, $p < .001$, *partial-eta squared* = .09 was found which were qualified by a significant interaction between Age and Country, $F(1, 376) = 29.28$, $p < .001$, *partial-eta squared* = .07. No other significant effects were found. As in Figure 2, both nationalities trusted younger flight attendants equally while Americans trusted older flight attendants more. Post hoc tests showed that trust dropped as a function of Age for Indians, but went up for Americans (all *p*s < .05).

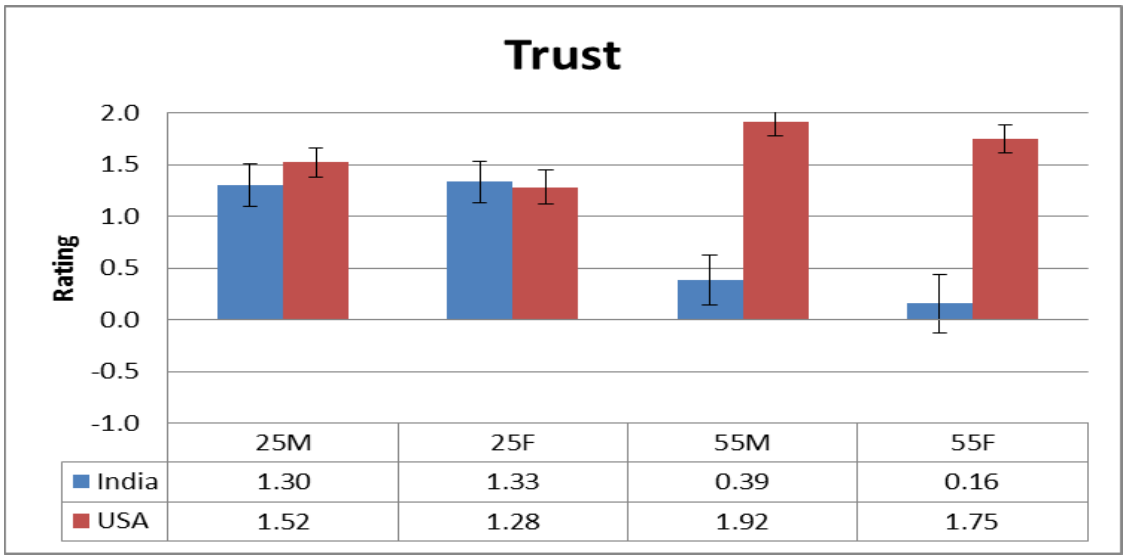


Figure 1. Trust data from the experiment. SE bars are included.

Discussion

The study was conducted in order to understand cultural differences and perceptions towards ageing flight attendants. The researchers predicted that there would be statistically significant differences in trust ratings when analyzing the country of origin of the participant, and the age of the flight attendants. The results showed a main effect of culture and country of origin, which supported the first and second hypotheses. The study found that Americans trusted older flight attendants more than their counterparts from India. One explanation for this is that Americans believed that 55 year olds would be better at handling most situations due to their experience. Indians on the other hand trusted younger (25 year old) flight attendants more than their American counterparts. It is possible that Americans view flight attendants as responsible for onboard safety whereas Indians view them more in terms of onboard service. What strikes as interesting is that collectivist cultures such as Indians view their elders in high regard. This would have implied that Indian participants would have more faith in the older flight attendants, and so these findings are insightful for commercial air service operators. The third hypothesis

predicted by the researchers stated that there would be differences in trust ratings as a function of gender of the target flight attendant in the scenarios. The results of the study however failed to support the predictions and suggested that there was no statistically significant differences in trust ratings as a function of the flight attendants' gender. Even though it was initially hypothesized that people would stereotype flight attendants as young females, it was interesting to note that participants felt that both genders were equally capable of fulfilling the role of a flight attendant.

Practical Implications and Limitations

Since aviation is hugely dependent on its consumers, their perceptions hold value. This study shows the growing sentiment against ageing flight attendants that is rampant in India and how it is fuelled by emotion. The study also looks to identify gender and age preferences. This could help airlines in personnel placement and help educating passengers about air safety. Moreover, future research could help track consumer perceptions over time.

This study is subject to certain limitations. First being that Amazon's ® Mechanical Turk ® was used collect data. The experimenter therefore has no control over the environment. Conversely, previous research has suggested that Mturk data is as reliable as data collected in a laboratory setting (Buhrmester, Kwang, & Gosling, 2011; Germine, et al., 2012). Moreover, the participants were compensated for the survey. This could affect their mindset and motivation for completing the survey.

Aviation is a global industry and involves almost every country on the planet. This research only analyzes perceptions of participants from two countries, India and the United States. For this reason, the study cannot generalize its inferences to the entire industry. Future research should attempt to collect data from several other countries, cultures and regions to enhance the understanding of these consumer perceptions. In addition, another future avenue for research lies in attempting to understand the reason behind such decisions of the participants. One possible method doing so would be to replicate this study with the addition of a mediation analysis. A possible mediator could be that of affect, which would imply that the decisions of the participants were based on emotions.

Conclusion

The study is aimed at further understanding the mindset of the travelling public, and the complex relationships that affect their trust in different aspects of the aviation industry operations, in this case the trust in the flight attendants. The purpose of this study was also to examine cross-cultural differences in trust of flight attendants using Indian and American participants. The study showed that Indians trusted older flight attendants more than the older counterparts. Conversely, it was seen that Americans trusted older flight attendants more. These findings have practical implications to the aviation industry. While these findings are important, they will serve as the basis for the researchers to analyze whether emotions (affect) mediates the relationship between the condition and trust.

References

Arokiasamy, P. (2004). Regional patterns of sex bias and excess female child mortality in india. *Population, 59*(6), 833-863.

Barber, B. (1983). The logic and limits of trust. New Brunswick, NJ: Rutgers University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6(3), 3-5.

Chen, G. M., & Starosta, W. J. (1998). Foundation of intercultural communication, MA: Allyn & Bacon. Chu, S. K. (1997). *China since 1911, Pacific Affairs, 70(1), 122-124*.

Cleveland, J. N., & Landy, F. J. (1983). The effects of person and job stereotypes on two personnel decisions. *Journal of Applied Psychology, 68, 609-619.*

Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., & Wilmer, J.B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5), 847-857.

Han, S. & Shavitt, S. (1994). Persusion and Culture: Advertising Appeals in Individualistic and Collectivistic Societies. *Journal of Experimental Social Psychology, 30,* 326-350.

Hofstede, G (1980). Culture's Consequences: National Differences in Thinking and Organizing. California: Sage Press.

Hofstede, G. (2001). Culture 's Consequences: Comparing Values, *Behaviors Institutions, and Organizations across Nations - 2nd Edition*, Thousand Oaks: Sage Publications.

Kite, M. E. & Johnson, B. T. (1988). Attitudes toward older and younger adults: A meta-analysis. *Psychology and Aging, 3, 233–244.*

Kite, M. E., Stockdale, G. D., Whitley, B. E. & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61, 242–262.*

Lessor, R. (1984). Social movements, the occupational arena and changes in career consciousness: The case of women flight attendants. *Journal of Occupational Behavior (Pre-1986), 5*(1), 37.

Markus, Hazel R.; Kitayama, Shinobu (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, Vol 98(2), 224-253*.

Rice, S., Kraemer, K., Winter, S. R., Mehta, R., Dunbar, V., Rosser, T. G., & Moore, J. C. (2014). Passengers from India and the United States Have Differential Opinions about

Autonomous Auto-Pilots for Commercial Flights.*International Journal of Aviation, Aeronautics, and Aerospace*, *1*(1), 3.

Rice, S., Trafimow, D., Hughes, J., & Hunt, G. (2011). Extending a courtesy stigma to a computer programmer, his work product, and his associates. *International Journal of Technology, Knowledge, and Society, 7*, 79-91

A REGRESSION OF CONSUMER ATTTITUDES TOWARD AIRPORT WATER REUSE

Ismael Cremer
Florida Institute of Technology
Melbourne, Florida
Stephen Rice
Florida Institute of Technology
Melbourne, Florida
Scott R. Winter
Florida Institute of Technology
Melbourne, Florida

Recent studies have focused on characterizing and understanding the public's perceptions of risk with respect to general reuse projects (Baggett, Jefferson & Jefferson 2005; Hurlimann 2011; Toze, 2005). These studies have shown varying attitudes toward water reuse and are necessary to assess the public's risk perception and acceptance of water reuse before implementing it. To date, no studies have examined whether certain variables affect people's attitudes toward the water reuse concepts at airports. Four hundred and four participants from India and the United States participated in a study wherein various socio-economic were collected along with their attitude scores toward water reuse. For "Flushing Toilet", the resulting model included two of the original predictors: Political Preference and Ethnicity. Liberals and Americans generated higher scores compared to their counterparts. This model accounted for 5% of the variance in the criterion. For "Washing Hands", the resulting model included four of the original predictors: knowledge of environmental science, water reuse knowledge, political preference, and the amount of water use. Participants with greater knowledge of environmental science, less water reuse knowledge, liberal, and used less water in general generated higher scores compared to their counterparts. This model accounted for 10% of the variance in the criterion. For "Drinking Water", the resulting model included two of the original predictors: Knowledge of Environmental Science and Ethnicity. In this model, participants with higher environmental science knowledge and Indians generated higher scores compared to their counterparts. This model accounted for 14% of the variance in the criterion.

Recycled water is an engineering process that can be implemented in areas that have shortages in water. These shortages can occur due to drought, urbanization, and increasing industrialization. Many European countries are also experiencing situations that are related to water stress (Hochstrat & Wintgens, 2003). Although this engineering process is safe and clean, public doubt of using such water still exists. It is imperative for the public to be willing and accepting to use recycled water, particularly in areas that have reduced natural water supply. There have been many recycled water projects around the world have failed due to the lack of support by the public community.

It is important to assess the public's risk perception and acceptance of water reuse before implementing it. Recent studies have focused on characterizing and understanding the public's

perceptions of risk with respect to general reuse projects (Baggett, Jefferson & Jefferson 2005; Hurlimann, 2011; Toze, 2005).

The case for using recycled water for various aspects at airports can be made from the projected 5% annual growth of aviation. Airports are large users of water. Atlanta alone uses over 252,600,000 gallons per year (Atlanta Sustainability Report, 2011). There are many airports around the world. They are vital to the economy; yet, can be seen as a large eco-footprint that impacts the natural state of the environment. Recycled water can be used in different ways at an airport, ranging from flushing toilets, to drinking.

Review of Literature

There are different factors that can affect an individual's acceptance of using recycled water. Their ethnicity, gender, knowledge, and past experiences can play a role. A study investigated the risk perception of two different groups in Australia, where one had a shortage of water and the other did not. The acceptance of using recycled water for various uses was assessed (Hurlimann, 2007). The study's results indicated that an individual's perception of risk did not vary between locations regarding water scarcity, however, background experience played a significant and positive impact on risk perception for drinking, showering, washing hands, and clothes washing. This study only focused on Australians, therefore, the current study will be looking at two cultures, as well as their background information to see what factors are associated with predicting their level of acceptance to using recycled water.

There are 18 countries that are currently considered to be at extreme risk with respect to their water security according to the Water Security Risk Index. Even the United States has water scarcity, particularly in areas that have high population growth, large water consumption, and a low level of natural water sources. European countries also have shortage in areas that use intensive irrigation practices. New water supplies to provide or supplement the current sources will be necessary. Water recycling is a method that allows for this to happen. In certain countries and states the implementation of indirect and direct water reuse is implemented. These countries include Israel, Spain, Italy, Australia, and Greece. In Namibia, direct potable reuse is implemented to supply drinking water to the public, and this is highly accepted by the population.

Current Study

Previous research has looked at the public's perception of water reuse risk on a county scale (Hurlimann, 2007), comparing two counties with varying levels of water scarcity. The current study expands on previous research and contributes a unique aspect of culture differences, and examines various predictor variables that contribute to predicting acceptance levels of using recycled water at various levels. This is important as success of specific projects, particularly at airports, depend on the public's attitudes and acceptance.

Methods

Participants

Four hundred and four (159 females) participants took part in the study. There were 204 (90 females) participants from India, and 199 (69 females) from the United States. The overall mean age was 30.86 (SD = 9.21). The average age for the Indian participants was 30.81 (SD = 8.96), and the average age for the United States participants was 30.90 (SD = 9.47). The differences in age were not significant between countries, $t(401) = .09$, $p = .93$.

Instrument

The study was presented online using FluidSurveys ®. Participants were recruited via Amazon's ® Mechanical Turk ® (MTurk). MTurk is a global online service that enables participants (Turkers) to participate in Human Intelligence Tasks (HITs) in exchange for monetary compensation. Participation in any HIT is voluntary and anonymous. MTurk provides data that is shown to be as reliable as laboratory data (Buhrmester, Kwang, & Gosling, 2011; Germine, et al., 2012).

Procedure

Participants first signed an electronic consent form. Following this, participants were asked to imagine that they were at an international airport terminal. They were told, "In an effort to conserve freshwater, the water used to flush the toilet waste is clean recycled water from a wastewater (sewage) treatment plant". In the other two conditions, participants were told the recycled water was for washing hands, or for drinking. Thus, there were a total of three different uses for the recycled water.

In order to measure affective responses to the questions, we used the same methodology as previous studies measuring affect (e.g. Rice, Richardson & Kraemer, 2014). Participants were asked in three different ways how they felt about these uses of recycled water, and they responded by choosing the appropriate Likert-type scale response from (-3) extremely negative/uncomfortable/unfavorable to (+3) extremely positive/comfortable/favorable. There was a neutral option a zero for each question. After the scenarios were completed, demographic data which included Gender, Ethnicity, Knowledge of Environmental Science, Knowledge of Water Reuse, Age, Political Preference, Income, Times flown per year, Water usage in gallons per day, and Education level as predictor variables was collected. The participants were then paid and dismissed.

Results

We began by conducting a regression analysis of the initial dataset using acceptance of using recycled water for Flushing Toilets as the criterion variable, and Gender, Ethnicity, Knowledge of Environmental Science, Knowledge of Water Reuse, Age, Political Preference, Income, Times flown per year, Water usage in gallons per day, and Education level as predictors. We used backward stepwise regression to eliminate ineffective predictors. For this stage the resulting model included two of the original ten predictors: Ethnicity and Political Preference. The regression model from this dataset was:

$$Y = 1.276 + 0.387X_1 - 0.145X_2$$

Where Y is the predicted acceptance score to using recycled water for Flushing Toilets, and X_1 and X_2 are Ethnicity and Political preference respectively. This model accounted for 5% of the variance in the criterion, $F(2,380) = 19.58$, $p < 0.001$.

We conducted the same analysis to examine the predictors with respect to using recycled water for Washing Hands. The criterion variable was acceptance of using recycled water for Washing Hands, and Gender, Ethnicity, Knowledge of Environmental Science, Knowledge of Water Reuse, Age, Political Preference, Income, Times flown per year, Water usage in gallons per day, and Education level were predictors the predictor variables. Here, the resulting model included four of the original ten predictors: Knowledge of Environmental Science, Knowledge of Water Reuse, Political Preference, and typical individual Water Usage in gallons per day. The regression model from this dataset was:

$$Y = 0.404 + 0.370X_1 – 0.263X_2 – 0.112X_3 – 2.989E\text{-}6X_4$$

where Y is the predicted acceptance score for Washing Hands with recycled water, and X_1 through X_4 are Knowledge of Environmental Science, Knowledge of Water Reuse, Political Preference, and typical individual Water Usage in gallons per day respectively. This model accounted for 10% of the variance in consistency, $F(4,380) = 10.94$, $p < 0.001$.

We lastly conducted the analysis again to examine the predictors with respect to using recycled water for Drinking. The criterion variable was acceptance of using recycled water for Washing Drinking, and Gender, Ethnicity, Knowledge of Environmental Science, Knowledge of Water Reuse, Age, Political Preference, Income, Times flown per year, Water usage in gallons per day, and Education level were predictors the predictor variables. Here, the resulting model included two of the original ten predictors: Ethnicity, and Knowledge of Environmental Science. The regression model from this dataset was:

$$Y = -0.916 – 06.58X_1 + 0.572X_2$$

where Y is the predicted acceptance score for Drinking recycled water and X_1 and X_2 are Ethnicity, and Knowledge of Environmental Science respectively. This model accounted for 14% of the variance in consistency, $F(2,380) = 33.09$, $p < 0.001$.

Discussion

There are different means of implementing water reuse projects. Depending on the level of need and water scarcity could determine the type of water reuse projects to initiate at an airport. While other studies have investigated differences between cultures and groups with respect to the level of contact with recycled water (Cremer, Rice, & Winter, in press), the attributes that may explain the acceptance levels have not been investigated. The purpose of this study was to identify those traits that have the greatest impact on the acceptance of type of water reuse. The results of this study have given us an outline of the main factors that contribute to acceptance level depending on the level of contact.

Flushing Toilet with Recycled Water

The two main factors that explained the level of acceptance of using recycled water for flushing the toilet purposes were ethnicity and political preference. The results indicate that U.S. individuals who were liberal were more willing to use recycled water for flushing toilets. A possible explanation to this is that in some states the practice of using recycled water for flushing toilets has started to grow in many other aspects such as in theme parks and in certain large public locations. Although they may not be as popular in airports, the exposure to this in the United States might explain the regression model. Furthermore, it can be argued that liberals will have a more open stance with respect to the interaction with a project that is a means of contributing positively to the welfare socially and to the environment. Using recycled water to flush a toilet where there is no direct physical contact is easier to accept than a process that may involve physically touching the water.

Washing Hands with Recycled Water

There were four factors that explained the level of acceptance of using recycled water for washing their hands: Knowledge of Environmental Science, Knowledge of Water Reuse, Political Preference, and typical individual Water Usage in gallons per day respectively. In this case it did not matter whether the individual was from the U.S. or from India. The results indicate that participants with greater knowledge of environmental science, less water reuse knowledge, liberal, and used less water in general generated higher acceptance scores compared to their counterparts. A possible explanation could be that individuals who have a higher knowledge overall of environmental science would be more willing to use it because they may understand the salient concepts of sustainability and the need to implement certain projects to reduce the impact to the environment from certain industries such as aviation. With a slightly less understanding of the processes to water reuse, acceptance is higher perhaps due to not fully understanding the technical aspects but trusting the science overall. Again, liberal participants are more willing to use this type of water, and those that use less water overall may also be more eco-conscious and thus willing to use innovative measures to reduce their environmental impact even further.

Drinking Recycled Water

For “Drinking Water”, the resulting model included two of the original predictors: Knowledge of Environmental Science and Ethnicity. In this model, participants with higher environmental science knowledge and Indians generated higher scores compared to their counterparts. Indians may be more willing to drink recycled water due to culture differences than Americans. Indians are considered to be a collectivist culture and thus more accepting of regulations imposed by authority. Thus, if the airport authority deems the water to be safe to drink, and the participant has a higher level of environmental knowledge, they may have a higher propensity to accept that the water is safe, and the concept of using recycled water for drinking is having a positive impact on the environment and society as a whole.

Conclusions, Limitations, and Recommendations

The purpose of this study was to examine the possible demographic predictors in acceptance toward water reuse between two different cultures. While prior research has

examined attitudes toward water reuse, no prior study has examined attitudes toward water reuse between different cultures, gender, and types of water reuse at a large-scale facility such as airports. This study found that different types of water reuse aspects had different predictor variables associated with it. Moreover, the scenario regarding drinking recycled water had the highest explained variance through the resulting predictor variables.

The main limitation of this study includes the use of an MTurk convenience sample. Another limitation is the use of the specific set of demographic data collected to be included in the regression analyses. Future research should replicate the study with more participants, preferably airport passengers, and enlarge the scope of predictor variables that will be examined.

References

Atlanta Sustainability Report (2011). *Hartsfield-Jackson Atlanta International Airport Sustinability report*. Retrieved from http://www.atlanta-airport.com/docs/Airport/Sustainability/2011%20Annual%20Sustainability%20Report%2011-15-12.pdf on Jan 3, 2015.

Bagget, S., Jeffrey, P., & Jefferson, B. (2006). Risk perception in participatory planning for water reuse. *Desalination,* 187, 149-158.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(3), 3-5.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(3), 3-5.

Cremer, I., Rice, S. & Winter, S. (in press). Attitudes toward sustainability between Indians and Americans on water reuse for different purposes at airports. *International Journal of Sustainable Aviation*.

Crook, J., MacDonald, A. J., & Trussell, R. Rhodes. (1999). Potable use of reclaimed water. *American Water Works Association, 91*(8), 40-49.

Hochstrat, R. and Wintgens, T. (2003). Report on Milestone M3.I, Draft of wastewater reuse potential estimation, Interim report, AQUAREC.

Hurlimann, A. and McKay, J. (2004). Attitudes to Reclaimed Water for Domestic Use: Part 2. Trust. Water, Journal of the Australian Water Association 31(5), 40-45.

Hurliman, A. C. (2007). Recycled water risk perception – a comparison of two case studies. Water Practice & Technology 2(4), 1-11.

Rice, S., Richardson, J. & Kraemer, K. (in press). The emotional mediation of distrust of persons with a mental illness. International Journal of Mental Health.

Toze, S. (2006). Water reuse and health risks – real vs. perceived. Desalination, 187, 41-51.

DEVELOPMENT OF THE AIR TRAFFIC CONTROL TOWER ALERTS STANDARD

Edmundo A. Sierra, Jr.
Federal Aviation Administration
Washington, D.C., United States of America
Michael Buckley
HumanProof
Arlington, VA, United States of America

FAA HF-STD-008 Air Traffic Control Tower Alerts Standard specifies functional requirements, alarm and alert human interaction characteristics, and threshold levels in systems that use an alert mechanism to capture human attention in air traffic control tower environments. FAA HF-STD-008 was developed to address a shortfall in the general criteria for alerts found in FAA HF-STD-001 Human Factors Design Standard. FAA-HF-STD-008 was developed in three phases: literature review and draft development, subject matter expert working group review and development, and stakeholder comment and adjudication. The results of the work include specific requirements for alerts and additional evidence of a repeatable human factors standard development procedure. There were shortfalls in general human factors requirements that were addressed by a standard with more specific requirements. Although human factors standards and the standardization process of human factors areas are relatively new in the FAA, there has been increased recognition of the significance of human factors requirements for system design.

At the end of a solo cross country flight, a private pilot on final approach hears the following.

LOW ALTITUDE ALERT CESSNA THREE FOUR JULIET,
CHECK YOUR ALTITUDE IMMEDIATELY.

The pilot recognizes that an air traffic controller has issued the alert to her. Was she too close to terrain or was there some other obstruction? How much time does she have to climb or should she do nothing? Whatever her level of situation awareness, her safety depends on the course of action she takes.

In the scenario above, both the pilot and the controller have to recognize and assess the situation in complex conditions. The pilot in this fictitious scenario soon recognizes that she should climb because a controller has issued a safety alert. For the controller; workload, traffic volume, the quality and limitations of the radar system, and the available lead time to react are factors (FAA Order 7110.65V, Air Traffic Control) impacting her ability to quickly assess the situation. The controller is able to recognize this situation with the help of a Minimum Safe Altitude Warning (MSAW). The topic of this paper is how requirements analysis contribute to the design, development, and implementation of effective alarms and alerts.

The Federal Aviation Administration (FAA) Systems Engineering Manual (SEM) Version 1.0, states "The Next Generation Air Transportation System (NextGen) is a comprehensive overhaul of the National Airspace System (NAS) to make air travel more efficient and dependable, while ensuring each flight is as safe and secure as possible." The FAA SEM describes the NAS as a System of Systems. The system engineering processes for completing the transformation to NextGen are described therein. The system engineering processes include Operational Concept Development, Functional Analysis, Requirements Analysis, Architectural Design Synthesis, and Cross-Cutting Technical Methods. The specific challenge reviewed in this paper is Requirements Analysis.

The FAA SEM continues, "Requirements Analysis is an iterative process that defines the essential system characteristics for all system components required for the product's successful development, production, deployment, operation, and disposal." Requirements Analysis is composed of two distinct activities: Requirements Development and Requirements Management. The approach described in this paper has a direct impact on both, but especially on Requirements Development. The activity develops functional requirements from the functions developed through the Functional Analysis Process. The authors' approach, described herein, was to develop a standard, which is a primary input to the Requirements Development process.

According to Rodrick, Karwowski, and Sherehiy, "Unlike other fields, standards and the standardization process in human factors and ergonomics are relatively new" (Rodrick et al., 2012, p. 1512). This statement pertains to human factors standards for FAA applications. Until recently (see FAA HF-STD-002 Baseline Requirements for Color Use in Air Traffic Control Displays [3/26/2007]), FAA HF-STD-001 Human Factors Design Standard (2003) was the only FAA reference providing formal input to Requirements Development. Although FAA HF-STD-001 includes requirements for air traffic control and maintenance; human factors specialists had to further analyze, decompose, and derive specific requirements for each system implementation. For example, requirements for air traffic control would be further analyzed and perhaps extrapolated for Terminal versus En Route. For Terminal, requirements would be further analyzed for Terminal Radar Approach Control needs versus Airport Traffic Control needs - and so on.

This paper describes the analysis of requirements for FAA HF-STD-008 Air Traffic Control Tower Alerts Standard (8/8/2014). Standards developers developed FAA-HF-STD-008 in three phases: literature review and draft development, subject matter expert (SME) working group review and development, and stakeholder comment and adjudication. The developers identified and compiled a preliminary, foundation set of requirements for FAA HF-STD-008 from FAA HF-STD-001, and also from additional sources from the literature. The developers compiled new candidate requirements from the literature because of the age of FAA HF-STD-001, and also because more detailed requirements were needed to support specific implementations for the tower environment. The developers also gathered additional requirement inputs from a team of FAA SMEs consulted at several diffferent FAA Towers.

After the standard developers identified and captured requirements from the literature, they matched the requirement to the appropriate level in the standard. Next, FAA SMEs analyzed the requirements. The standards developers and FAA SMEs repeated the process until the

requirements were stable. Finally, the standards developers sent out a draft FAA HF-STD-008 for stakeholder comment. The standards developers adjudicated the stakeholders' comments before publishing the standard.

The standard developers achieved two things. First, they developed a body of requirements for the design and implementation of alarms and alerts for systems supporting tower operations in the form of a published FAA HF-STD-008 Air Traffic Control Tower Alerts Standard (8/8/2014). Second, this work provided evidence of a repeatable human factors standard development procedure. This paper describes method and results in detail in the sections that follow.

Method

The standard developers began requirements development with a review of the literature. They reviewed government documents that included military, and non-military federal agency standards, handbooks, and specifications. They reviewed non-government publications from organizations such as the American National Standards Institute (ANSI) and the International Electrotechnical Commission (IEC). Standards developers also reviewed research on alarms and alerts. FAA HF-STD-008 lists applicable documents in Section 2 and references in Appendix B.

A spreadsheet was used to support Requirements Management during the processes of compiling, filtering, sorting, organizing for access, and evolution through SME review. Compared to other FAA programs, the size and complexity of this project was small. A spreadsheet sufficed to manage the number of requirements. The spreadsheet application was used to capture, compile, and track the evolution of requirements. The spreadsheet was also used to maintain source documents traceability to the evolving requirements.

The standards developers identified and captured requirements from the literature in the spreadsheet. Next, the standards developers analyzed, filtered and sorted requirements for the tower environment. FAA SMEs, a group of FAA Senior Scientific & Technical Advisors for Human Factors and Senior Engineers, reviewed the requirements and further analyzed and evaluated them for applicability, accuracy, and conciseness. The standards developers and FAA SMEs repeated the analysis and evaluation in three successive reviews of succeeding drafts. Finally, the standards developers sent out the Draft FAA HF-STD-008 for stakeholder comment. The standards developers adjudicated the stakeholders' comments and submitted resolved comments for FAA review before drafting the final version of the standard.

Standard development took about a year. The standard developers identified and captured requirements from the literature within three months. They worked with FAA SMEs for six months. Public comment, adjudication, and FAA review took three months. This time frame did not include project planning and project management activities that were important, but occurred before and after standards development.

In addition to the literature review, FAA SME analysis and evaluation, and public comment; the standard developers used additional methods to enhance the viability of the product and ensure the validity of the requirements. Standard developers conducted tower

facility visits with the Senior Scientific and Technical Advisor for Terminal. Standard developers also consulted with Engineering Research Psychologists and other specialists. These activities ensured that the standard developers more fully embraced and accommodated stakeholder needs, known constraints, current interface limitations, operating environments, and modes of operation.

There was one known limitation of the process as implemented. To allow for the best solutions for NextGen, requirements must be solution agnostic. For this effort , requirements were molded to facilitate unbiased and measurable evaluation of various solution alternatives. Standard developers analyzed and evaluated requirements for applicability to air traffic control, then for Terminal, and then once more for Airport Traffic Control. To be truly solution agnostic, Human factors specialists will need to also analyze and evaluate requirements for specifications of alerting systems such as MSAW, Conflict Alert (CA), or Far Field Monitor (FFM).

Results & Discussion

The method used by the standard developers performed well. During the literature review, a large number of candidate requirements were collected and compiled. During the FAA SME Review Phase, there were hundreds of suggested additions, simplifications, deletions and edits on the body of candidate requirments. By the time the draft was ready for stakeholder comment, most comments were administrative and very few were more than editorial in impact. That is, they addressed items such as typographical, format, and grammatical errors. There was also a fairly short cycle time from requirement change initiation to approved resolution. Finally, the number of validated requirements to total proposed requirements was not highly variable.

Enhancements for tower are summarized in Table 1.

Table 1.
FAA HF-STD-008 Alert Enhancements for Tower.

Property or Atttribute	FAA HF-STD-001: Chapter 7: Alarms, Audio & Voice Communications	FAA HF-STD-008
Coverage	High level, general coverage of alarms and alerts	Detailed coverage of all types of alarms and alerts for Tower operations
Focus	Audio and voice	Audio, visual and tactile requirements addressed
High-level organization	High-level coverage that addresses general functions and attributes, implementation concerns, and the intrinsic characteristics of audio and voice signals	Organized consistent with FAA-STD-068; includes a treatment for general and detailed requirements
Signal treatment	Audio-relevant requirements only	Includes non-modal-specific requirements
Signal characterization	Few characterization specifics	Detailed characterization and construction specifics

Property or Atttribute	FAA HF-STD-001: Chapter 7: Alarms, Audio & Voice Communications	FAA HF-STD-008
Implementation	Includes a few implementation-specific requirements	Includes a wide assortment of both non-modal-specific and modal-specific implementation requirments
Coding	Very few coding-specific requiremets	Many coding-specific requirements, including coding for each mode
System-specific treatment	Some equipment-specific treatment: controls, handsets, headsets, telephone systems	Generic system requirements relevant to alarms and alert systems in general

The method did address the problem, but not entirely. FAA HF-STD-008 Air Traffic Control Tower Alerts Standard (8/8/2014) will likely perform as intended. However, truly successful Requirements Development is measured by the acceptable transformation of stakeholder needs into discrete, verifiable, specific and applicable requirements. Many of these requirements will meet this criteria. However, the scope of FAA HF-STD-008 requires that human factors specialists further analyze and evaluate requirements for alert systems in light of changing needs and evolving technologies.

Consider requirement *4.1.1.3 Minimize response time*. The requirement reads, "An alarm and alert system must minimize the time required for the operator to detect and assess the situation and to initiate corrective action(s)." It is critical that human factors specialists analyze this requirement to enable requirements verification and compliance. The analysis may one day lead to a timely alert that the pilot in the introduction, and many others, will appreciate.

TRAFFIC ALERT CESSNA THREE FOUR JULIET,
ADVISE YOU TURN LEFT, AND CLIMB IMMEDIATELY.

As a final observation, the authors think it is important to note that the socialization of the content in the evolving standard ranked only slightly below the significance of the product's final content. The mechanics of the three working group reviews of the draft in its successive forms started a conversation on component and philosophical issues that continued throughout the review process, culminating in the resolution of final comments following the wider stakeholder review. Each working group member received a copy of the latest draft for his examination prior to the meeting. Whether the working group members came prepared with a marked up copy with specific embedded comments, or just showed up with personal notes, questions and issues, each member had an opportunity to make sure the draft remained headed in the right direction. During the meeting, and with capture completed a short time later, all comments were captured in the Comments Resolution Matrix, giving each member direct feedback on the resolution of their concerns. Along the way, working group members became invested in the process and the product, sometimes becoming minor champions of specific decisions made during the evolution of the document.

Further, what was started with the working group reviews, continued in the wider stakeholder review. Though the working group numbered relatively few individuals, they were widely dispersed both organizationally and geographically. Widely networked, the fact that the

working group played such a crucial role in the mechanics of the evolution of the standard surely benefitted participation during the final stakeholder review.

References

Ahlstrom, V., & Longo, K. (2003). Human Factors Design Standard (HF-STD-001). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

FAA Systems Engineering Manual (Version 1.01). (2014). Washington, D.C.: Federal Aviation Administration, 800 Independence Avenue SW Washington, DC 20591.

Federal Aviation Administration. (2007). *Baseline requirements for color use in air traffic control displays* (DOT/FAA/HF-STD-002). Washington, DC: U.S. Department of Transportation, FAA Human Factors Research and Engineering Group.

Federal Aviation Administration (2014). *Air traffic control tower alert standard* (DOT/FAA/FAA HF-STD-008). Washington, DC: U.S. Department of Transportation, FAA Human Factors Division.

Federal Aviation Administration. (2014). *Air Traffic Control* (FAA Order 7110.65V). Retrieved from http://www.faa.gov/regulations_policies/orders_notices/index.cfm/go/document.information/documentID/1023549

Rodrick, D., Karwowski, W. and Sherehiy, B. (2012) Human Factors and Ergonomics Standards, in *Handbook of Human Factors and Ergonomics*, Fourth Edition (ed G. Salvendy), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781118131350.ch55

Acknowledgements

FAA HF-STD-008 Air Traffic Control Tower Alerts Standard was developed with funding from FAA R&D Program, *A11.i Air Traffic Control/Technical Operations Human Factors*. This project was sponsored by Chuck Perala, Scientific & Technical Advisor at the Federal Aviation Administration. The authors gratefully acknowledge colleagues who assisted in developing the standard by actively participating in the working group reviews: Bert Howells (HumanProof), Charles Jones (AJW-131 Maintenance Automation Team), Chuck Perala (AJM-352 Specialty Engineering Team), Fred Brooks (National Institute of Aerospace), Jerry Crutchfield (AAM-500 Office of Aerospace Medicine - Aerospace Human Factors Division), Sehchang Hah (ANG-E25 Human Factors Branch), and Steve Cooley (AJM-352 Specialty Engineering Team). The authors are also grateful to FAA facility managers and tower controllers who hosted facility visits and to the aviation human factors community for their time reviewing and commenting on FAA HF-STD-008. The views of the authors do not necessarily reflect the views of the Federal Aviation Administration or HumanProof.

A^{3}IR-CORE AND FLIGHTPROFILER: AN ACADEMIC-INDUSTRY PARTNERSHIP FOR SMS DEVELOPMENT

John H. Mott
Department of Aviation Technology, Purdue University
West Lafayette, IN

Mark C. Ball
Department of Aviation Technology, Purdue University
West Lafayette, IN

FlightProfiler, a safety management system software for general aviation that has been under development since 2000, quantifies and illustrates how 79 different factors collectively affect planning for general aviation flights. The software uses advanced collaborative decision-making technology and NextGen analytics to prescreen an entire flight cycle, with objectives of improving flight safety and reducing costs. The Advanced Aviation Analytics Institute for Research (A^{3}IR-CORE) at Purdue University has entered a partnership with the software developer to improve the usability of the product in a collegiate aviation environment. This includes creating process flow diagrams of the software and of Purdue flight operations, gathering flight data, and improving the presentation of the output. Purdue will in turn be given access to the FlightProfiler tool for use in its flight program. The software development team consists of mathematicians, meteorologists, graphic technologists, and computer scientists who are in communication with A^{3}IR-CORE faculty and students to effectively collaborate and organize the task. Participating students will acquire business development skills working on a student-designed project plan in addition to the skills needed in industry to analyze, formulate and apply logical techniques for work task improvement.

The Purdue Masters of Science degree in Aviation and Aerospace Management is designed to prepare graduate students with the background necessary to become future leaders in the aviation industry. Some of the primary areas of concentration in this program include operational analysis, project management, human factors, safety and security, environmental sustainability, and resource analysis. AT 52000, a course taught by faculty associated with the Advanced Aviation Analytics Institute for Research (A^{3}IR-CORE), focuses on critical points in process workflow and their effects on customer service, employee relationships, cycle time, and profit. The course utilizes a combination of lecture and group projects and discussion to link subject material to practical applications thereof that are relevant to students' professional interests and career goals. The students in the course in the fall semester of 2014 were divided into multiple project teams; the research described herein resulted from the work of the FlightProfiler team.

The benefits to FlightProfiler included feedback regarding the utility of the software and recommendations for improvements thereto from users in an academic setting, as well as the potential for integration of the decision-making tool into the Purdue flight program. The benefits to the students included development of students' skills such as communication, project design and management, and professionalism, as well as growth of students' professional networks.

Literature Review

According to the U.S. Department of Transportation, there were 440 fatalities and 1,471 general aviation accidents in 2012. Although these numbers have generally decreased over time, they are still large compared to the zero fatalities and 27 commercial aviation accidents recorded in the same year (DOT, 2014). In an attempt to mitigate the relatively high proportion of general aviation accidents, the FAA identified a need for a risk assessment tool for general aviation pilots. There are several products on the market that are being tested and refined in addition to the FlightProfiler product that is the focus of the current research. These additional products are described below.

FltPlan.com, a web-based flight planning service, is one of the largest such services in North America, with a customer base of over 150,000 active pilots. FltPlan.com recently "enhanced its Safety Management System program by adding multiple Flight Risk Assessment Tools that are customizable for both a flight department's operation and the department's different aircraft" (General Aviation News). In addition to flight planning and safety management capabilities, the site also offers flight tracking, runway analysis, weight and balance calculators, an eLogbook program, and FBO and airport information. All of this information is helpful to pilots, but a new opportunity exists in effectively communicating potential flight risks to pilots. FltPlan's enhanced safety management system aims to capitalize on that opportunity.

Leveraging new technologies to produce new risk assessment tools can result in a major potential benefit to general aviation pilots. The new iPad Flight Risk Assessment Tool (iFRAT) offers pilots an easy way to examine risks associated with their flights (Sniderman, 2015). As the commercial airline industry moves further toward adoption of iPad technology, the general aviation industry would appear to be remiss in not adopting similar technology. The process of assessing a particular flight's risk can be accomplished in a few minutes either before or after a flight. iFRAT offers interactive risk displays that provide critical information in an easy-to-understand manner.

An additional product with enhanced risk assessment features is the Lockheed Martin Flight Services website. Lockheed-Martin provides flight planning services to over 80,000 general aviation flights per week, covering areas including pre-flight, in-flight, operational and special services, enroute communications, search and rescue services, and meteorological and aeronautical briefings (Lockheed Martin, 2015). When a general aviation user enters his or her flight information at the site, there are specific risks that are highlighted through the reports with which the user is presented. An easy-to-follow risk assessment is provided to the pilot through the appropriate smartphone, tablet, or computer application. Any concerns or warnings are highlighted in yellow or red, depending on the severity of the concern, and are also ranked as text in terms of importance to the pilot's flight and depicted on a map to help the pilot understand where the risks are positioned in relation to the flight. This information is presented in an easy-to-read and easy-to-remember format with a visually appealing layout and information flow.

Methods Used to Form Recommendations

A^3IR-CORE faculty and students met with FlightProfiler to establish business development objectives, which included testing the program in an academic setting, collecting feedback, drawing information from other current resources, and developing recommendations. To achieve these objectives, the team developed a project plan spanning the fall semester. The plan included the creation of process flow diagrams of the Flight Profiler website and Purdue flight operations using Microsoft Visio. Development of the flow diagrams included process map construction, critical sequence streamlining, and value-added analysis. As a result, the team could easily determine process overlaps, weaknesses, and other highlighted items.

The team first gathered information from Purdue flight operations to develop an understanding of all of the steps necessary to fly a visual flight rules (VFR) training flight. This included documenting all necessary steps, from arriving with the instructor to accepting the aircraft. Next, the team decomposed all of the steps necessary to complete a flight risk assessment on the FlightProfiler website. Once the processes were understood, the team facilitated focus groups of student pilots in which those pilots accessed Flight Profiler and tested it. 20 Purdue students enrolled in the professional flight program each created a personal flight risk assessment on the website, discussed observations with the researchers, and completed an online response form. Face-to-face communication was considered important for high-quality feedback, so a team member was present when the testing was conducted. The recommendations herein were derived from focus group response data, direct verbal feedback in focus groups, and the research team's own suggestions and findings.

FlightProfiler Recommendations

Purdue flight students currently use a flight risk assessment tool (FRAT) (see appendix) before every flight. The research team found that the FRAT tool covers many areas in a manner similar to that of the Flight Profiler questionnaire. The FRAT tool is grouped into preflight info, flight operations, weather, and training flight criteria, which each question weighted appropriately. At the end of the questionnaire, the pilot is given a score. If that score

is above the minimum, then the pilot can fly the particular flight in question. If the score is in a "warning" range, the pilot must get approval from the chief pilot. If it is below a minimum, then the pilot cannot fly.

Figure 1. FlightProfiler output.

The team looked at a typical FlightProfiler output (Figure 1). The solid blue line shows the amount of risk a particular flight is expected to encounter during the course of the flight. The higher the line, the less risk the flight is predicted to have. The software will overlay other flights that have been assessed to show how a particular flight compares with those flights and to provide a general threshold of safety. The example in Figure 1 shows only a single flight between the listed departure and destination airports. The upper dashed line shows a realistically optimum flight and the lower dashed line depicts a minimum threshold, based on Federal Aviation Regulations. In this example, the flight begins below FAA minimums and terminates close to those minimums. Potential causes for this include a departure airport that may be below weather minimums and pilot fatigue as the flight arrives at the destination airport.

With regard to the entering of information on the website, the team suggested a better categorization of airports. Currently, the website has an alphabetical dropdown list of all commercial airports in the United States. Inclusion of a search bar or a means to sort airports by location allow more convenient entry of origin, destination, and alternate airports. In addition, the home airport for the Purdue flight program (Lafayette-Purdue University Airport [KLAF]) was not listed.

The team also suggested the addition of a mechanism to include personal preferences and safety minima. Because different kinds of training flights are flown by pilots possessing different levels of experience, it would be helpful for pilots to choose their own safety minimums, including safe minimum runway lengths, winds, ceilings, visibilities, and so forth.

Some of the specific questions that were asked during the data entry process were not applicable to Purdue's training flights. For example, it would be helpful to have the software calculate proper fuel loads. The current software does not do this; rather, it asks for entry of the fuel load in units of thousands of pounds, which is not appropriate for most general aviation training aircraft. Questions relative to aircraft age were not relevant for

Purdue's new fleet and were asked twice. A question relative to arrival time may also be irrelevant, as most training flights may not have scheduled arrival times; such questions should be optional.

The software incorporates a color-coded radar image (Figure 2) with weather data that shows the area over which the pilot will fly during flight, and depicts the weather conditions relative to precipitation that may be anticipated at the origin and destination airports. This image also includes a color-coded line showing the change in risk over time. The line in the example below changes from red to green, indicating a general decrease in risk as the flight progresses.

Figure 2. Color-coded radar map.

Another output display the team suggested is a hazards display similar to that produced by Lockheed Martin's web-based software, which shows current NOTAMS, PIREPS, and other restrictions and warnings (Figure 3). All of these warnings are highlighted in yellow, orange, and red based on the level of risk. The information is summarized in boxes that are also color-coded in terms of risk level. For example, information presented in an orange box is high-risk, while information shown in a yellow box is intermediate-risk in nature. The blue line represents a sample training flight path and the user can see the risk of encountering the precipitation to the east of the path would be relatively low.

Figure 3. Hazards display.

The team recommended that an option be included to allow for determining alternate departure times associated with lower overall risk levels. This option could also suggest alternate airports that could result in reduced overall risk.

Relevance to the Next Generation of Pilots

Students are generally uninterested in large quantities of information. Younger general aviation pilots are accustomed to sending quick text messages and communicating in an abbreviated style. In order for information to be conveyed effectively, that information must be concise. The team agreed with the focus groups in their assessment that the FlightProfiler weather pages were not overly useful because of their lack of specificity and readability, and their failure to precisely depict which risks to a particular flight might exist. Other information presented was found to be excessive, which too much reliance on the pilot to sort through information to make an adequate determination of risk.

In order to provide relevant information to the next generation of pilots, the industry must emphasize the use of images, animation, and color-coding to help convey important information easily and quickly, an issue that was the primary basis for the team's recommendations. There is a need to create a risk display design that meshes satisfactorily with the decision-making process, to organize risks for quicker reference, to omit unneeded information, and to form better visuals of weather tailored to specific routes and time periods. Understanding and being able to evaluate risk is a very important skill for general aviation pilots. It takes continuous study and experience to develop these skills and to apply them to specific flights. Given information about levels of risk, there will likely always be pilots who choose to fly and others who do not. But the tools discussed here can assist the pilot in developing a better understanding of hazardous conditions that may impact that pilot's ability to safely complete a flight.

References

General Aviation News. (December, 2014). FltPlan adds flight risk assessment tools. *General Aviation News.* Retrieved from http://generalaviationnews.com/2014/12/05/fltplan-adds-flight-risk-assessment-tools/

Lockheed Martin. (2015). Lockheed Martin Flight Services. *Lockheed Martin.* Retrieved from http://www.lockheedmartin.com/us/products/afss.html

Sniderman, D. (2015). New iPad Application Helps Pilots Assess Flying Risks. *AVNET.* Retrieved from http://www.em.avnet.com/en-us/design/technical-articles/Pages/Articles/New-iPad-Application-Helps-Pilots-Assess-Flying-Risks.aspx

United States Department of Transportation. (2014). Table 2-14: U.S. General Aviation (a) Safety Data. *Office of the Assistant Secretary for Research and Technology*. Retrieved from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_02_14.html

Appendix

Purdue FRAT Tool

Pre Flight Info	YES	NO
solo flight (pre-private)	1	
Student less than 50 flight hours	2	
Student 50-150 flight hours	1	
Instructor's first semester teaching	1	
Instructor has CFII or MEL	-1	
Stress factor	2	
Flight Operations		
Runway less than 4000'	2	
Night landing	2	
No precision approaches available at destination (IFR only)	2	
Non-towered airport	2	
Unfamiliar airport	2	
Class C operations	2	
Student has not flown in the last 2 weeks		
Last Sleep Period (Less than 4hrs)		
Last Sleep Period (4 to 6 hrs)		
Last Sleep Period (6 to 8 hrs)		
Show time (between 7-8am)	2	
Show time (after 6pm)	3	
Maintenance test flight	3	
First flight after a phase or 50hr inspection	3	
Weather		
Departure- MVFR	2	
Departure- IFR	3	
Enroute- Turb. Forecasted along route	1	
Enroute- Thunderstorms forecasted	2	
Arrival - MVFR	2	
Arrival- IFR	3	
Arrival – winds > 15kts		
Training Flights		
Behind flight schedule	4	
Flying multiple approaches (253)	2	
Pattern work		
Instructor or student back to back training flights	1	
Class immediately before/after flight	1	

NIGERIA'S AVIATION AT A GLANCE: THE ASSESSMENT OF NIGERIANS' PERCEIVED TRUST LEVEL IN NIGERIA'S AVIATION INDUSTRY

Ibrahim G. Miya
Stephen C. Rice
Florida Institute of Technology
Melbourne, Florida

The recent occurrences of fatal aviation crashes in Nigeria have significantly affected Nigerians' trust in the overall performance efficiency of Nigeria's aviation. In the context of Africa's aviation, Nigeria in particular, it appeared that very little is being done on trust. This study assessed Nigerians trust level in Nigeria's aviation industry with respect to "Familiarity-Based Trust Model," (Zhang, Ghorbani & Cohen, 2007). The study used a 7-point Likert-type survey questionnaires as the primary data collection tools. Ten predictor variables (income, age, gender, political view, aircraft ownership, purpose of flying, class ticket, relationship status, distance flown, and flight frequency) were regressed on four dependent variables (pilots, airline, government, and aircraft). The result indicated that relationship status, annual flight rate, age, average annual income, and class ticket were significant predictors of Nigerians' perceived trust in the industry. The result was an eye opener for further scholarly research on trust in nation's aviation.

In the last few years, Nigeria's aviation industry has been in serious turmoil that degraded the level of Nigerians confidence in the industry. "The problems to confront are legion—and, of course, not just confined to the aviation sector: lack of transference, lack of management skills, errant government interventions, funding and infrastructure shortages" (Asiegbu, Igwe & Akeku, 2012, p.138). Unless critical steps and essential measures are being taken to restore Nigerians confidence in the market, the possibility remains that Nigerians' trust level in the industry will continue to suffer. Nigeria's Government, airlines, pilots, and aircraft have so far being identified by Nigerians as the four major players in the nation's industry. The purpose of this study is to assess the factors that influence Nigerians trust in the nation's aviation industry.

Literature Review

Familiarity-Based Trust Model

The theoretical framework upon which this study was grounded was familiarity-based trust model developed by Zhang, Ghorbani, and Cohen (2007). This theoretical model described trust as a function of factors such as experiences, repeated exposures, level of processing and forgetting rates of an agent. Several studies have explored the significant relationship between trust and familiarity. It was found that individuals often preferred familiar investments, and fear change and unfamiliarity (Zhang, Ghorbani & Cohen, 2007). According to Dani et al., (2006), trust is influenced by familiarity between agents over a significant period, shared experiences, reciprocal symbiosis between the agents, and "demonstration of non-exploitation expressed over time" (p.592). Familiarity is more than just underlying value-systems between two agents (Cater and Ghorbani, 2004). Familiarity, on the other hand, is defined as a complex understanding, which is often based on past interactions, experiences, and learning of others (Luhman, 1979; Zhang, Ghorbani & Cohen, 2007). It was stated that "trust is determined by the interplay of individuals' values, attitudes, moods, and emotions" (Jones & George, 1998, p.531). Familiarity is a paramount concept that mediates trust and reliance between agents.

Trust

Over the years, scholars have studied trust from different disciplinary perspectives, for example, psychological, sociological, and political science perspectives. It appeared obvious in the trust literature that there was no explicit integration about the definition of trust (Bhattacharya, Devinney & Pillutla, 1998). "Trust makes interactions easy. Supervisors and subordinates can coordinate their work efforts more effectively in the context of mutual trust. Likewise, international relations can progress rather than stall or regress when parties trust each other" (Lount, 2010, p.420). Trust as simple as it sounds, requires an exchange between trustor and trustee (usually

called agents). Agents in the context of the trust literature included but not limited to institutions, individuals, machines or organizations (Giustiniano & Bolici, 2012).

Definition of Trust

Review on the current researches showed that there hardly was a single definition of trust (Bhattacharya, Devinney & Pillutla, 1998; Kramer, 1999; Brewer, 1996; Calton, 1998). Lee and See (2004); Hoffmann and Sollner (2014) defined trust as a belief that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. It is a willingness to be vulnerable to the actions of another individual based on the expectations that the other party will perform a particular action important to the trustor, irrespective of the trustor's ability to monitor or control the trustee (Mayer, Davis & Schoorman, 1995; Zang, 1997; Rousseau, Sitkin, Burt & Camerer, 1998). Trust is a form of social capital; it is organizational resources; it is a mental state, and it is the foundation of social interaction in a group among agents on the basis of positive expectation (Bugdol, 2013). The economic definition of trust, however, is more centered on how institutions are created to minimize anxiety and uncertainty resulting from interactions between agents (Bhattacharya, Devinney & Pillutla, 1998). Researchers from a broad range of disciplines have examined the role of trust in the mediating relationship between individuals, groups, and organizations and emphasized the significance of trust in organizational settings (Lee & See, 2004; Rotter, 1967, Rampel et al., 1985; Johns, 1996; Moorman et al., 1993; Davis & Schoorman, 1995).

Organizational Trust

Study on organizational trust is gaining momentum and attention in the trust literature. Many researchers have stated that not much was done in the past on organizational trust, (Connell, Ferres & Travaglione, 2003; Tan & Tan, 2000; Mayer & Davis, 1999; Clark & Payne, 1997; Kramer & Tyler, 1996; Hosmer, 1995; Mayer, Davis & Schoorman, 1995)."This interest has been fueled, at least in part, by accumulating evidence that trust has a number of significant benefits for organizations and their members" (Kramer, 1999, p.569). In the context of employee-organization trust, it was found that increased distributive justice and procedural justice in the workplace; increased organizational commitment and decreased organizational turnover, are very critical to developing and strengthening trust (Rousseau, Sitkin, Burt & Camerer, 1998; Tan & Tan, 2000). Developing trust in the organizational context is imperative most especially in the event of crisis management because the business culture alone can influence trust (Rousseau, 1998; Calton, 1996; Tan & Tan, 2000; Calton, 1998; Connell, Ferres & Travaglione, 2003; Reychav & Sharkie, 2010).

Interpersonal Trust

Interpersonal trust is an integral part of organizational trust. Interpersonal trust could be politically-based, economically-based, or socially-based forms of trust. Studies on interpersonal trust on social and organizational sciences are among the important emerging areas in the trust literature. Teams are composed of different people where social relationships is crucial; hence, trust and distrust are very relevant factors to defining the kind of relationships that exist among such individuals (Baba, 1999). Trust is a cultural and sociological phenomenon, and it operates in several ways (Mechanic, 1996). Trust between various stakeholders in an organization, for example, employees and managers; managers of different subunits; firms and their customers; buyers and suppliers are necessary to enable effective and profitable business transactions (Fukuyama, 1995). Propensity to trust and trustee characteristics are paramount determinant of interpersonal trust. Propensity to trust depends on personality traits, cultural backgrounds and personal experiences (Hassan & Semerciöz, 2010).

Trust in Automation

Automation is the execution by a machine agent of a function that was previously carried out by human (Parasuraman & Riley, 1997).In the aviation domain, for instance, designers of mechanical aids do try to automate everything that could have an economic benefit without adequately addressing all the associated human factor problems such as insufficient feedback from the mechanical aid to the human operator, mode awareness, mode errors, mode management, and defect with the system reliability which impacts operator's trust on the system (Parasuraman & Riley, 1997; Norman, 1990; Sarter & Woods, 1994; Parasuraman, Molloy & Sign, 1993). Trust in automation is believed to be among the most important factors that affect operator's reliance, use, misuse, disuse,

and abuse of an electronic system as it does to humans (Parasuraman & Riley, 1997; Lee & See, 2004; Muir, 1988; Zhang, Ghorbani & Cohen, 2007). There are several studies which argued that the same factors affecting human-human trust, affect human-machine trust (Sheridan, 1975; Sheridan & Hennessy, 1984). People are more likely to trust machines that are believed to be reliable and trustworthy (Muir, 1988; Lee & See, 2004; Parasuraman & Riley, 1997; Zhang, Ghorbani & Cohen, 2007).

Design, Setting, and Methodology

The study constructed four regression models and determined coefficients of multiple determinations for each model that measured the variance in criterion variables (pilots, airline, government, and aircraft) being explained by the predictor variables. The predictor variables were gender, relationship status, private aircraft ownership, average annual distance travel, annual flight frequency, and most common purpose of traveling, class ticket, age, average annual income, and political view.

Population and Sample

The targeted population for this study was "Nigerians." The sample for this study was a randomly and a conveniently selected Nigerians who met the eligibility requirements for the study. The total sample size of (N = 110) Nigerians was collected. The sample had a total of (n = 30) females and (n = 80) males. There was a total of (n = 43) participants who had related knowledge and experiences in aviation. The a priori "Power Analysis" was performed using G*Power 3.0.10 (Faul, Erdfelder, Lang & Bucher, 2007). This analysis yielded a minimum sample size at (α = .05) of 82 participants. The post hoc power analysis was .99. The primary data collection tools were survey questionnaires. The survey instrument was a semantic differential scale that asked the participants to rate their perceived trust level in each of the four levels of the criterion variables (pilots, airline, government, and aircraft) using a 7-point Likert scale that ranged from negative 3 (-3), extremely distrust to positive 3 (+3), extremely trust. The predictor variables for this study were: gender, relationship status, private aircraft ownership, average annual distance travel, annual flight frequency, and most common purpose of flying, class ticket, age, average annual income, and political view.

Hypotheses

The study methodology was quantitative. The raw data collected were analyzed using descriptive multiple regression. The multiple regression analyses evaluated the relationship and predictability between the dependent and the independent variables. The study tested the following hypotheses:

> H_0: $\beta = 0$ (stated that there are no significant predictors that explain Nigerians' perceived trust level in Nigeria's aviation industry and Nigeria's government policy toward aviation, Nigeria's airline service quality, Nigeria's pilots flying skills, and Nigeria's aircraft service condition).
>
> H_A: $\beta \neq 0$ (stated that there are significant predictors that explain Nigerians' perceived trust level in Nigeria's aviation industry and Nigeria's government policy toward aviation, Nigeria's airline service quality, Nigeria's pilots flying skills, and Nigeria's aircraft service condition).

The data analysis produced the coefficients of multiple determinations, *t*-ratios, and *f*-ratios. The a priori alpha-level of significance was set at (α .05). The decision to reject the null hypothesis was made based on the *f*-ratios and *t*-ratios, respectively. The Statistical software used was SAS JMP® 11 software. As a parametric study, the data collected were tested and satisfied the statistical assumptions of linearity, normality, and homoscedasticity assumptions.

Results

Descriptive Statistics

The study used a sample size (N = 110) participants: (n = 30) females and (n = 80) respectively. The age distribution of the study group showed a mean and standard deviation (M = 38.25, SD = 9.34). The sample data showed that participants have mean annual income (M = 1,630,606, SD = 1,302,265) Nigeria's Naira. The income distribution showed a widespread income disparity among Nigeria's population. Annually, each participant travels, on average distance of (M = 2,271, SD = 2,805) kilometers. The mean and standard deviation of absolute trust in Nigeria's aviation were (M = .70; SD = 1.23).

Inferential Statistics

ANOVA used trust as a dependent variable and the four levels (group) of Nigeria's aviation (pilots, aircraft, airline, and government) as independent variables. The null hypothesis for this test stated that there is no significant difference between the group means of pilots, aircraft, airline, and government ($\mu_{pilots} = \mu_{aircraft} = \mu_{airline} = \mu_{government}$). The alternative hypothesis stated that at least one group mean is different. The result indicated statistically insignificant difference between the group means for pilots, aircraft, airline, and government. The differences in means between pilots (M = .81, SD = 1.64), aircraft (M = .66, SD = 1.44), airline (M = .71, SD = 1.43) and government (M = .66, SD = 1.60) were not statistically significant, $F(3, 439) = .34$, $p > .794$, $\eta p2 = .001$.

Regression Output

The first regression analysis used trust in Nigeria's pilots flying skills as the dependent variable. Backward stepwise regression was employed to delete the ineffective predictors that were statistically insignificant. The resulting model included two of the original predictors. Relationship status and annual flight frequency significantly predicted Nigerians' perceived trust in pilots' flying skill. Relationship status: $\beta = .28$, $t(110) = 3.11$, $p = .003$; annual flight frequency: $\beta = .28$, $t(110) = 3.09$, $p = .004$.This model accounted for 17% of the variance in the dependent variable, $F(2, 109) = 10.98$, $p < .001$. The result showed that the more Nigerians fly in aircraft piloted by Nigeria's pilots, the more they trust the pilots' flying skills. Also, the result showed that married Nigerians flyers tend to trust Nigeria's pilots flying ability more than the unmarried Nigerians.

The second regression analysis was conducted using perceived trust in Nigeria's airline service quality as dependent variable, with the same response variables. Age significantly predicted perceived trust in Nigeria's airlines services, $\beta = .28$, $t(110) = 3.07$, $p = .003$. This model accounted for 8% of the variability in the criterion variable, $F(1, 109) = 9.40$, $p = .003$. The results indicated that the older generations of Nigerians believed that the Nigeria's airlines are doing better.

The third regression analysis was performed using perceived trust in aircraft maintenance status as the criterion variable; the same predictor variables were used. Annual income and class ticket were significant predictors for trust in Nigeria's aircraft maintenance condition. Average annual income (Naira): $\beta = .39$, $t(110) = 3.86$, $p = .002$; class ticket: $\beta = .32$, $t(110) = 3.14$, $p = .002$. The model accounted for 14% of the variance in the dependent variable, $F(2, 109) = 8.62$, $p = .003$. Nigerians with higher income and who can afford more expensive flight tickets believe that the maintenance status of the nation's aircraft is acceptable

In the last regression model, the criterion variable was perceived trust in Nigeria's government's aviation policies, with the same predictor variables. Average annual income and class ticket significantly predicted perceived trust in Nigeria's government aviation policies. Annual income: $\beta = .53$, $t(110) = 5.60$, $p < .001$; class ticket: $\beta = .34$, $t(110) = 3.57$, $p = .005$. The model accounted for 23% of the variance in the criterion variable, $F(2, 109) = 16.32$, $p < .001$. The result showed that the first class passengers in Nigeria perceived that the Nigeria's government policies toward aviation are within acceptable tolerance.

Decision on Hypotheses

The null hypothesis was rejected in favor of the alternative hypothesis using both classical and p-value decision rules. This affirmed that that there are significant predictors that explained Nigerians' perceived trust level

in Nigeria's aviation industry and Nigeria's government policy toward aviation, Nigeria's airlines service quality, Nigeria's pilots flying ability, and Nigeria's aircraft service condition. Perceived trust in pilots flying skills: $F(2, 109) = 10.98$, $p < .001$; perceived trust in airline service quality: $F(1, 109) = 9.40$, $p = .003$; perceived trust in aircraft maintenance status: $F(2, 109) = 8.62$, $p = .003$; and perceived trust in government aviation policies: $F(2, 109) = 16.32$, $p < .001$. The results showed statistical significance compare to baseline ($p = .05$). The following variables, therefore, were found to be statistically significant predictors of Nigerians' perceived trust level in the performance efficiency of the Nigeria's aviation: annual flight frequency, relationship status, age, average annual income, and class ticket.

General Discussion

It was found in this study that relationship status, flight frequency, age, average annual income, and class ticket were significant predictors of trust in Nigeria's aviation industry. Readers should remember that trust is a social phenomenon; it is a psychological state, and the foundation of social interaction in various aspects of human life (Bugdol, 2013). Trust is also a social capital; is hard to build but easy to dissipate (Kramer, 1999). The measurement of trust is very subjective. Individuals' perception to trust or distrust another agent could be affected by their mood, emotion, personality traits, and expectation. In the trust literature, it is accepted that trust is a history-dependent process based on fair dealings between agents. In the context of this study, the better the service quality Nigerians receive from Nigeria's aviation; the more positive their expectation about swhat the industry can offer, and the stronger their trust in the industry.

According to familiarity-based trust model, trust is a combination of self-esteem, reputation, and familiarity (Zhang, Ghorbani & Cohen, 2007). A Nigeria's airline, for example, that builds a positive reputation through quality service delivery, hiring highly qualified pilots, and ensuring a high standard of aircraft maintenance, can enhance stronger public trust compare to an airline that did not do as well. In another words, Dani at el., 2006 defined trust as an agent's belief that another agent makes a reasonable effort to behave in accordance with the expected level of commitment. Nigerians who fly more frequent have positive experience about the aircraft maintenance status and the pilots' skills; hence, they tend to have a stronger faith in pilots and aircraft. The age of flying public, for example, was a significant predictor of trust in airline service quality. Annual income and class ticket revealed the differences in experiences between first class and economic class passengers in term of inflight services. Passengers with higher annual income can afford first class tickets; tend to reflect positively about the airline service and the government policies.

Conclusion

This study evaluated the factors that influenced Nigerians' trust in Nigeria's aviation. The study did not include all the predictor variables that affect people trust. Trust is believed to be influenced by individuals' attitudes, emotion, expectations, and predisposition to trust other agents (Kramer, 1996). It is recommended that future research should consider additional variables such as: emotions, educational level, personality, and predisposition to trust, among others.

References

Asiegbu, I. F., Igwe, P., Akeku A. N., (2012). Physical evidence and marketing performance in airlines in Nigeria. Retrieved from http://www.aijcrnet.com/journals/Vol_2_No_12 December_2012/15.pdf.

Baba, M. L. (1999). Dangerous liaisons: Trust, distrust, and information technology in American work organizations. Human Organization, 58(3), 331-346. Retrieved from http://search.proquest.com/docview/201158163?accountid=27313.

Connell, J., Ferres, N., & Travaglione, T. (2003). Trust in the workplace: The importance of interpersonal and organizational support. *Journal of Management Research*, 3(3), 113-118. Retrieved from http://search.proquest.com/docview/237226557?accountid=27313.

Calton, J. M. (1998). Trust in organizations: Frontiers of theory and research. Business and Society, 37(3), 342-346. Retrieved from http://search.proquest.com/docview/199480299?accountid=27313.

Dani, S. S., Burns, N. D., Backhouse, C. J., & Kochhar, A. K. (2006). The implications of organizational culture and trust in the working of virtual teams. Proceedings of the Institution of Mechanical Engineers, 220, 951-959. Retrieved from http://search.proquest.com/docview/195139996?accountid=27313.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. Academy of Management. The Academy of Management Review, 23(3), 531-546. Retrieved from http://search.proquest.com/docview/210977294?accountid=27313.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. Annual Review of Psychology, 50, 569-98. Retrieved from http://search.proquest.com/docview/205830786?accountid=27313.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46 (1), 50-80. Retrieved from http://search.proquest.com/doc view/216440925? accountid=27313 1253 (1993).

Murray, J. (2013). Likert data: What to use, parametric or non-parametric*? International Journal of Business and Social Science*, 4(11) Retrieved from http://search.proquest.com/docview/1446602622?accountid=27313.

Moreland, R.L., Zojanc, R.B., (1982). Exposure effects in person perception: Familiarity, and attraction. J. Exp. Social Psychology. 18(5), 395-415 (1982). Retrieved from http://www.sciencedirect.com/science/article/pii/0022103182900622.

Norman, G. (2010). Likert scales, levels of measurement and the laws. Adv. in Health SciEduc, 5:625-632, DOI 10, 1007/s10459-010-9222-y. Retrieved from Springer: http://www.fammed.ouhsc.edu/research/FMSRE%20Orientation%20&%20Hand out%20 Material/Handouts%205%20Science/Likert%20Scales.pdf.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230. Retrieved from http://search.proquest.com/docview/216444294?accountid=27313.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision-warning systems. Ergonomics, 40, 390-399.

Rousseau, D. M. (1998). Trust in organizations: Frontiers of theory and research. Administrative Science Quarterly , 43(1), 186-188. Retrieved from http://search.proquest.com/docview/203987577?accountid=27313.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). *Journal of Personality and Social Psychology Review*, 23(3), 531-546. Retrieved from http://search.proquest.com/docview/210977294?accountid=27313.

Thettacharya, R., Devinney, T. M., & Pillutla, M. M. (1998). A formal model of trust bsased on outcomes. Academy of Management. The Academy of Management Review, 23(3), 459-472. Retrieved from http://search.proquest.com/docview/210961965?accountid=27313.

Tan, H. H., & Tan, C. S. F. (2000). Toward the differentiation of trust in supervisor and trust in the organization .Genetic, Social, and General Psychology Monographs, 126(2), 241-60. Retrieved from http://search.proquest.com/docview/231482545?accountid=27313.

Zhang, J., Ghorbani, A. A., & Cohen, R. (2007). A familiarity-based trust model for effective selection of sellers in multiagent e-commerce systems. *International Journal of Information Security*, 6 (5), 333. doi:http://dx.doi.org/10.1007/s10207-007-0025-y.

EVALUATING STARTLE, SURPRISE, AND DISTRACTION: AN ANALYSIS OF AIRCRAFT INCIDENT AND ACCIDENT REPORTS

Andrew B. Talone, Javier Rivera, Camilo Jimenez, & Florian Jentsch
University of Central Florida
Orlando, Florida

Over the years, startle, surprise, and distraction have been frequently cited as potentially having negative effects on aircraft flightcrew performance. This paper aims to build upon and extend our prior research (Rivera, Talone, Boesser, Jentsch, & Yeh, 2014) in which we found evidence that (a) startle may be less problematic to flight deck performance than surprise, and (b) negative flight deck performance following startle is most likely due to concurrent distraction or surprise. The current research examined the theoretical foundations underlying these concepts and analyzed two accident/incident databases to identify potential trends and assess the prevalence of startle, surprise, and distraction on the flight deck. Results indicated that across the entire 20-year period, distraction was the most prevalent, followed by surprise and startle.

In a recent analysis from 2004-2013, the Boeing Company's Statistical Summary of Commercial Jet Airplane Accidents identified loss of control in-flight (LOC-I) as one of the leading categories of fatal accidents (16 fatal accidents) and total number of fatalities (Boeing, 2014). LOC-I accidents have been found to be influenced by physiological and psychological factors such as startle, surprise, and distraction, which have been frequently cited as potentially having negative effects on aircraft flightcrew performance. As such, it is important to understand the psychological and behavioral similarities and differences between these concepts, and investigate their prevalence, especially considering changes in flight deck systems.

This effort aims to build upon and extend our prior research (Rivera, Talone, Boesser, Jentsch, & Yeh, 2014) in which we found evidence that (a) startle may be less problematic to flight deck performance than surprise, and (b) negative flight deck performance following startle is most likely due to concurrent distraction or surprise. Therefore, the current study examined the theoretical foundations underlying these concepts and analyzed records from two accident/incident databases (the Aviation Safety Reporting System [ASRS] database and the National Transportation Safety Board [NTSB] aviation accident database) to identify potential trends and assess the prevalence of startle, surprise, and distraction. Based on the results of this analysis, we discuss the benefits of incorporating these factors into flightcrew training programs, such as line-oriented flight training (LOFT), to improve crew resource management (CRM). Then, we present potential approaches to mitigate the negative effects of startle, surprise, and distraction.

Background

Startle

Startle is an involuntary physiological reflex elicited by sudden exposure to intense stimulation (Koch, 1999). Startle consists of two components: an involuntary, immediate *startle reflex* and a conditioned, behavioral *startle response*. The startle reflex typically consists of physiological responses such as eyes blink, head ducks, and shoulders crouched up to protect the body against adverse situations (Grillon & Baas, 2003). In addition, the startle reflex can have substantial, negative effects on immediate, gross motor performance, but these effects are typically brief (no longer than 1 to 3 s) (Ekman, Friesen, & Simons, 1985; Landis & Hunt, 1939). The conditioned startle response involves a pattern of behavioral and physiological responses, which can result in (a) task interruption and (b) substantial cognitive impairments (i.e., deficiencies in information processing) that can last significantly longer than the startle reflex (up to 15 s to 1 min) (Thackray & Touchstone, 1970; Vlasak, 1969).

During flight, the flightcrew may encounter a variety of situations that could elicit startle. For instance, the impact of laser illuminations, especially during final approach, has been found to have the potential to disorient and startle flightcrew (Nakagawara, Montgomery, Dillard, McLin, & Connor, 2004). However, based on our review and analyses, we believe that it is unlikely that the immediate psychomotor impact of the startle reflex typically results in control inputs that have a catastrophic influence on flight maneuvers (with the exception, perhaps, of startle during critical flight phases close to the ground/terrain, such as during takeoff or final approach/landing).

Surprise

In contrast to the startle reflex and the startle response, which manifest themselves primarily physiologically and behaviorally, surprise is a cognitive-emotional response to mismatches between mental expectations and perceptual representations of the actual environment (Meyer, Niepel, Rudolph, & Schützwohl, 1991; Schützwohl & Borgstedt, 2005). As has been stated before (Kochan, Breiter, & Jentsch, 2004; Rivera et al., 2014), surprise can occur at the onset/appearance of an unexpected stimulus, but also by the absence of an expected stimulus. The main concern regarding surprises in any task environment is that the feeling of surprise is generally strong and noticeable, and thus it interrupts the execution of an ongoing task. For example, Horstmann (2006) found that the inclusion of a surprising event during the execution of a continuous motor task caused 78% of the study's participants to interrupt the task. Horstmann also found that, on average, the interruption to the execution of the motor task lasted about 1 second. The length of an interruption from surprise, however, can vary based on the magnitude of the expectation mismatch. That is, a surprising event that vastly differs from what was expected can produce longer interruption durations.

Surprises occurring on the flight deck during line operations can affect a range of the flightcrew's responses, including their physiological, cognitive, and behavioral responses, manifesting themselves, for example, in increased heart rate, increased blood pressure, the feeling of being unable to comprehend/analyze the situation or to remember appropriate operating standards (Rivera et al., 2014). In extreme cases, behavioral "freezing" and loss of situation awareness have been reported (Bürki-Cohen, 2010). Additionally, surprise has been found to play a role in LOC-I accidents such as Air France Flight 447, in which the flightcrew seemed to be confused about multiple failure indications and a disconnection of the automated systems (Bureau d'Enquêtes et d'Analyses [BEA], 2012). Automation surprises, specifically, have been investigated extensively through the years to better understand the impact of flight deck designs on flightcrew performance (Sarter & Woods, 1995; Sarter, Woods, & Billings, 1997).

In a previous study exploring the impact of surprise in aviation, Kochan et al. (2004) analyzed incident and accident reports from the NTSB's aviation accident database, the National Aeronautics and Space Administration's (NASA's) ASRS database, and the National Aviation Safety Data Analysis Center (NASDAC) database. From their analysis, Kochan et al. concluded that factors eliciting surprise could be grouped into clusters, which included the aircraft's state (e.g., automation, system alerts), environmental conditions (e.g., turbulence, low visibility), instructions or actions from others (e.g., air traffic control [ATC] directing holding), and the sudden appearance of other aircraft. In addition, Kochan et al. stated that surprises did not need to be unusual or rare to be unexpected or surprising, given that most of the reports they reviewed involved a relatively routine flight procedure that became a surprising event by occurring, however, in an unusual situation or context (e.g., temporary runway closure due to debris on the runway, or wildlife in the vicinity of the runway).

Distraction

Distraction refers to the diversion of attention away from activities that are required for the accomplishment of a primary goal to other competing sensory (e.g., visual, auditory, biomechanical) and cognitive activities. Airbus (2004) and Dismukes, Young, and Sumwalt (1998) found that factors such as communication, heads-down work, responding to an non-normal/unexpected event, searching for traffic out-the-window, flight deck ergonomics, flight deck noise level, language proficiency (from both the pilots and controllers), airport infrastructure, and flightcrew fatigue can have an impact on flightcrew performance. In flight, distraction can contribute to the development of dangerous situations, such as runway incursions/excursions, late responses to ATC instructions, late retraction of the landing gear, altitude deviations, inadequate energy management, and controlled flight into terrain.

Method

To better understand the prevalence of startle, surprise, and distraction on the flight deck, we reviewed accident/incidents reports from two databases: the ASRS database and the NTSB's aviation accident database. Below, we describe the procedure used to search each database for reports in which startle, surprise, and distraction were involved. For both of these databases, our searches focused on (a) the terms *startle*, *surprise*, or *distraction* (and their derivatives), (b) the time period January 1994 to December 2013, and (c) Part 121 and 135 operations (i.e., air carriers and commuters). In addition to qualitative (narrative/content) analyses (which we will report in the future), we focused on identifying numerical trends. We asked, for example, whether the two decades (1994-2003

vs. 2004-2013) yielded comparable numbers of reports; whether there were changes in the number of reports across the three phenomena, etc.

Aviation Safety Reporting System (ASRS) Database

The ASRS database allows one to use a wildcard (%) to yield derivations of searched terms. For example, using the search term *distract%* will yield reports containing *distract*, *distraction*, *distracting*, etc. This wildcard was used with all three constructs to ensure our search captured all reports in which the phenomena were discussed. Therefore, three separate searches were conducted: *startl%*, *surprise%*, and *distract%*.

NTSB Aviation Accident Database

The NTSB aviation accident database does not have a wildcard option, therefore, separate searches were conducted for each derivative of a particular term (e.g., *distraction* and *distracted* have to be used within their own search). The database also does not allow one to search more than one type of flight operation at the same time (e.g., Part 121 and Part 135). Therefore, six searches were completed for startle (*startle*, *startling*, *startled* for each type of operation), six for surprise (*surprise*, *surprised*, *surprising* for each type of operation), and eight for distraction (*distract*, *distraction*, *distracting*, *distracted* for each type of operation). After each search was completed, the resulting raw number of reports were analyzed to identify those in which more than one derivative of a word was used (e.g., *distraction* and *distracted*). This was done to prevent counting a report more than once when adding the number of reports across searches. After this process was completed, all of the unique reports were added up to yield the total number of reports in which each construct was mentioned.

Results

Taken together, our searches yielded an initial total of 4,781 reports in which the words *distract%*, *startl%*, or *surpris%* were used (here we used % to imply all derivatives of these words). After removing duplicate NTSB reports, our total was reduced to 4,773 reports. Our prior investigation (Rivera et al., 2014) had thoroughly examined the terminological usage of the terms startle and surprise within the reviewed ASRS reports. In the study reported here, in contrast, we were interested in establishing how often these three factors have been used to describe an incident or accident within aviation operations over the past 20 years. Table 1 displays the total numbers of reports in which *distract%*, *startl%*, or *surpris%* were used to describe an incident or accident. From the results of our investigation, it appears that distraction was much more prevalent than both surprise and startle across the 20-year period investigated. Furthermore, this trend was prevalent across both databases.

Table 1.
Number of ASRS and NTSB Reports Returned.

Factor	ASRS	NTSB	Total
Startl%	181	4	185
Surpri%	1,736	26	1,762
Distract%	2,770	56	2,826
Total	4,687	86	4,773

Note. The total number of ASRS and NTSB reports identified in which a derivative of *startl%*, *surpris%*, or *distract%* was mentioned in the incident/accident description.

Besides comparing the prevalence rates across constructs, we were also interested in assessing whether the prevalence of each construct had changed over time. As indicated in Tables 2 and 3, the total number of reports involving startle decreased during the second decade (Jan. 2004 – Dec. 2013). This trend was also found for surprise. In contrast, distraction increased during the second decade. These trends were prevalent across both databases.

Table 2.
Decade Comparison for the ASRS Reports.

Factor	Jan. 1994 – Dec 2003	Jan. 2004 – Dec. 2013	Difference
Startl%	110	71	-39
Surpri%	956	780	-176
Distract%	1,365	1,405	+40

Note. A comparison is provided between the total number of ASRS reports identified during the period Jan. 1994 – Dec. 2003 and Jan. 2004 – Dec. 2013.

Table 3.
Decade Comparison for the NTSB Reports.

Factor	Jan. 1994 – Dec 2003	Jan. 2004 – Dec. 2013	Difference
Startl%	3	1	-2
Surpri%	16	10	-6
Distract%	22	34	+12

Note. A comparison is provided between the total number of NTSB reports identified during the period Jan. 1994 – Dec. 2003 and Jan. 2004 – Dec. 2013.

Discussion

In this investigation, we found evidence that, across the entire 20-year period and consistently in both decades we investigated, distraction was the most prevalent phenomenon cited in accident and incident reports, followed by surprise and (far-distant) startle. In fact, our findings further supported our prior contention (cf. Rivera et al., 2014) that true startle events, in which the physiological and behavioral responses of flightcrew result in negative consequences to the safety of flight, are actually very, very rare. Instead, we believe that startle, to the degree that it is caused by events such as bird strikes, compressor stalls, and laser illuminations, mostly results in surprise and distraction – which, in turn, may cause incidents and accidents.

We also noticed that the preponderance of distraction as a causal or contributing element in safety events was consistent across both databases, despite each one serving different purposes. The ASRS database consists of voluntarily reported incidents, whereas the NTSB database consists of governmentally mandated reports of major accidents/incidents. Another interesting finding was the fact that, while the number of both startle and surprise reports had decreased from the first to the second decade, reports of distraction not only failed to decrease, but actually increased. This finding was also consistent across the two different databases. One possible explanation for the increased prevalence of distraction may be the introduction of electronic information management devices, such as portable electronic devices (PEDs) and electronic flight bags (EFBs), onto the flight deck; the narratives of the more recent reports certainly seemed to support this. Additionally, this mirrors a finding by Chase and Hiltunen (2014), who, in their own accident/incident database investigation, found evidence that pilots report that they are sometimes distracted by a PED or EFB. Another possible explanation is that increased airport and airspace congestion has placed higher attentional demands on flightcrew; especially during high workload periods (see also Loukopoulos, Dismukes, & Barshi, 2009). Taken together, these findings indicate that, while all three of these constructs are still prevalent on the flight deck, distraction may be the issue that is both most concerning but also most likely to be ameliorated through design of systems and procedures, as well as through training.

Despite the fact that distraction was found to be the most prevalent, our results showcase the fact that all three of these psychological phenomenon are still prevalent within flight deck operations. Given this, we believe that it is important to improve pilot training so that the negative impact of these psychological constructs on flightcrew performance can be mitigated. To this end, we suggest that these factors should be included within LOFT scenarios

conducted during CRM training. This suggestion falls in line with a Notice of Proposed Amendment published by the European Aviation Safety Agency (EASA; 2014) in which it was proposed that startle and surprise should be incorporated into CRM training. In regards to startle, the EASA recommended that CRM training cover the acquisition and maintenance of adequate automatic behavioral responses during crisis to be used during unexpected, unusual, and/or stressful situations. In regards to surprise, the EASA recommended that CRM training should encourage the development of pilot resilience (e.g., mental flexibility and the ability to adapt performance to address current conditions). Other recommendations for addressing surprise include using in-flight discussions of "what if" scenarios (Martin, Murray, & Bates, 2011) and mental simulation (Roth & Andre, 2004) to promote better decision-making. As for distraction mitigation strategies, these include establishing distinctive roles (e.g., pilot flying, pilot not flying), scheduling/rescheduling activities to minimize distraction due to less important activities, avoiding task-irrelevant conversations during high workload phases of flight (e.g., takeoff, approach, and landing), and incorporating scenarios that require pilots to manage distractions into simulator training (Australian Transportation Safety Bureau [ATSB], 2005; Dismukes et al, 1998).

It is, however, important to note several limitations of our investigation. For one, the numbers reported here should be taken as conservative estimates of the true prevalence of distraction, startle, and surprise on the flight deck. This can be attributed to the voluntary aspect of ASRS reports. It is likely that there have been startle, surprise, or distraction occurrences that simply were not reported. Another limitation of this investigation is that we did not include other terms that arguably mean the same thing (e.g., *surprise* and *confuse*). Finally, there may have been reports in which the aforementioned constructs impacted other aviation professionals and not pilots on the flight deck. For example, reports describing distracted air traffic controllers, ground control personnel, or flight attendants may have been included in the numbers reported here. Despite these limitations, this investigation still provides initial insight into the frequency with which these constructs have been, and are currently, occurring on the flight deck. Furthermore, it makes evident the need for training interventions and strategies targeted at mitigating the negative impact of startle, surprise, and distraction on flightcrew performance.

Acknowledgments

This research was supported by FAA collaborative research agreement 13-G-007 (Human Factors Division, ANG-C1). The views and opinions expressed, however, are those of the authors and do not necessarily represent those of the Federal Aviation Administration (FAA) or of the institutions with whom the authors are affiliated.

References

Airbus. (2004, October). Human performance: Managing interruptions and distractions (FOBN Reference: FLT_OPS – HUM_PER – SEQ03 – REV03 – OCT. 2003). In *Flight Operations Briefing Notes*. Blagnac, France: Author. Retrieved from http://www.airbus.com

Australian Transport Safety Bureau [ATSB] (2005). *Dangerous distraction: An examination of accidents and incidents involving pilot distraction in Australia between 1997 and 2004* (Aviation Research Investigation Report B2004/0324). Retrieved from http://www.atsb.gov.au

Boeing. (2014, August). *Statistical summary of commercial jet airplane accidents-Worldwide operations, 1959-2013*. Seattle, WA: Aviation Safety Boeing Commercial Airplanes. Retrieved from http://www.boeing.com

Bureau d'Enquêtes et d'Analyses [BEA] (2012). Pour la Sécurité de l'Aviation Civile. Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France, flight AF 447 Rio de Janerio–Paris.

Bürki-Cohen, J. (2010). Technical challenges of upset recovery training: Simulating the element of surprise. *Proceedings of the AAIA Modeling and Simulation Technologies Conference*. doi:10.2514/6.2010-8008

Chase, S. G., & Hiltunen, D. (2014). *An examination of safety reports involving Electronic Flight Bags and Portable Electronic Devices* (Report No. DOT-VNTSC-FAA-14-12). Washington, DC: Federal Aviation Administration. Retrieved from http://ntl.bts.gov/

Dismukes, R. K, Young, G. E., & Sumwalt, R. L., III. (1998). *Cockpit interruptions and distractions: Effective management requires a careful balancing act* (Document ID 2002006298). Retrieved from http://ntrs.nasa.gov/

Ekman, P., Friesen, W. V., & Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology*, 49(5), 1416-1426. doi:10.1037/0022-3514.49.5.1416

European Aviation Safety Agency [EASA] (2014). *Notice of proposed amendment for crew resource management: NPA 2014-17*. Cologne, Germany: Author. Retrieved from https://www.easa.europa.eu/

Grillon, C., & Baas, J. (2003). A review of the modulation of the startle reflex by affective states and its application in psychiatry. *Clinical Neurophysiology*, *114*(9), 1557-1579. doi:10.1016/S1388-2457(03)00202-5

Horstmann, G. (2006). Latency and duration of the action interruption in surprise. *Cognition & Emotion*, *20*(2), 242-273. doi:10.1080/02699930500262878

Koch, M. (1999). The neurobiology of startle. *Progress in Neurobiology*, *59*(2), 107-128. doi:10.1016/S0301-0082(98)00098-7

Kochan, J. A., Breiter, E. G., & Jentsch, F. (2004). Surprise and unexpectedness in flying: Database reviews and analyses. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 335-339). Santa Monica, CA: Human Factors and Ergonomics Society. doi:10.1177/154193120404800313

Landis, C., & Hunt, W.A. (1939). *The startle pattern*. Oxford, England: Farrar and Rinehart.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009). *The multitasking myth: Handling complexity in real-world operations*. Farnham, Surrey, England: Ashgate.

Martin, W. L., Murray, P. S., & Bates, P. R. (2011). What would you do if…? Improving pilot performance during unexpected events through in-flight scenario discussions. *Aeronautica*, *1*(1), 8-22. Retrieved from https://www104.griffith.edu.au/index.php/aviation

Meyer, W.-U., Niepel, M., Rudolph, U., & Schützwohl, A. (1991). An experimental analysis of surprise. *Cognition & Emotion*, *5*(4), 295-311. doi:10.1080/02699939108411042

Nakagawara, V. B., Montgomery, R. W., Dillard, A. E., McLin, L. N., & Connor, C. W. (2004). *The effects of laser illumination on operational and visual performance of pilots during final approach* (Report No. FAA-AM-04-09). Washington, DC: Federal Aviation Administration Retrieved from: http://www.laserstrikeprotection.com/bulletins-n-reports.html

Rivera, J., Talone, A. B., Boesser, C. T., Jentsch, F., & Yeh, M. (2014). Startle and surprise on the flight deck: Similarities, differences, and prevalence. *Proceedings of the 58th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1047-1051). Santa Monica, CA: Human Factors and Ergonomics Society. doi:10.1177/1541931214581219

Roth, T., & Andre, T. S. (2004). Improving performance in pilot training by using the chair flying technique. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)* (Vol. 2004, No. 1). Arlington, VA: National Training Systems Association.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5-19. doi:10.1518/001872095779049516

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (2nd ed., pp. 1926-1943). New York, NY: Wiley.

Schützwohl, A., & Borgstedt, K. (2005). The processing of affectively valenced stimuli: The role of surprise. *Cognition & Emotion*, *19*(4), 583-600. doi:10.1080/02699930441000337

Thackray, R. I., & Touchstone, R. M. (1970). Recovery of motor performance following startle. *Perceptual and Motor Skills*, *30*(1), 279-292. doi:10.2466/pms.1970.30.1.279

Vlasak, M. (1969). Effect of startle stimuli on performance. *Aerospace Medicine*, *40*(2), 124-128.

CONSUMER TRUST RATINGS AFTER AN AIRLINE ACCIDENT: AN AFFECTIVE PERSPECTIVE

Scott R. Winter, Stephen Rice, Ismael Cremer, and Rian Mehta
Florida Institute of Technology
Melbourne, FL USA

Fortunately airline accidents are rare; however when one occurs it usually results in widespread media attention. The purpose of this study was to examine how consumers' trust ratings were impacted when one airline suffered an accident. The findings indicate that System Wide Trust (SWT) theory applies resulting in a trust reduction for all airlines, not just the accident airline. Affect was shown to act as a mediator in only three of the cases, which suggest that consumer's responses may not be strongly influenced by emotions. Practical implications and limitations of this study are provided.

Consumers have a unique role when it comes to trust in airlines. When a person boards an aircraft for flight, they are placing their trust in the airline. This may be from prior experience or even the experience(s) of another person. However, if an accident or incident were to occur, it is possible that a consumer's trust in that airline may be reduced. Additionally, if one airline has an accident, it is possible that a passenger's trust across multiple airlines may be diminished.

Literature Review

There are many definitions of trust. As it relates to airlines, Meyer, Davis, & Schoorman (1995) provides an applicable definition. In that context, trust is defined as one's ability to give up or relinquish control to another person or object. When it comes to passengers boarding airliners, they are giving up control to the airline and more specifically the flight crew operating the flight. The passengers may or may not have any direct contact with these individuals, yet they are trusting in them and the company to get them to their destination safely. Additional authors (Barber, 1983; Rampel et al., 1985; Rotter, 1967) suggest that trust is the result of expectations of certain events occurring. This expectation allows a person to predict what is most likely to occur, which may lead to a positive experience (Lee & See, 2004).

System-Wide Trust

System-Wide Trust (SWT) theory proposes that when operators or consumers view automated devices, they view them as one system. While the devices are independent, research suggests that both operators and consumers are unable to differentiate between devices and when one fails, they lose trust in other automated devices (Geels-Blair, Rice, & Schwark, 2013; Keller & Rice, 2010; Rice & Geels, 2010; Winter, Rice, & Reid, 2014). The current study sought to examine if SWT theory would apply to consumers rating trust in an airline after one suffered an accident.

Cultural Differences

In previous studies that have examined system-wide trust theory, there have been noticeable differences across cultures, specifically those that are individualistic and collectivistic. A definition of culture is the common societal norms, values, and practices in which one choses to participate (Helmreich, 2000). Those from individualistic cultures have strong views of themselves while persons from collectivistic cultures have a more interdependent view towards one another (Markus & Kitayama, 1991). Persons from collectivistic cultures are more likely to be trusting, especially of new people (Hofstede, 1980). On the rating scale of individualistic/collectivist, the United States scores as the most individualistic country, while India was much more collectivistic (Robbins & Judge, 2009).

Affect

Human beings may or may not be able to remove emotions from their decision-making processes, and it has been demonstrated that affective and cognitive components are separate (Trafimow, & Sheeran, 1998, 2004; Trafimow et al., 2004). When present, emotions can have an influencing effect on the decision-making process. During situations when there is a shortage of time or decisions have to be made quickly without much time to

cognitively process, humans may respond in an effective manner. However, it is possible that these emotional responses may not always be the most appropriate. Earlier studies (Remy, Winter, & Rice, 2014; Winter, Rice, & Mehta, 2014) have examined consumer's perceptions towards pilots and found that emotions played a significant role in which pilots they trusted more or less based on various demographic features (e.g. weight, gender, age).

Current Study

The purpose of the current study was to examine how consumers trust ratings would be affected after one airline experienced a fatal accident. Individuals from India, a collectivistic country, and the United States, an individualistic country were selected to participate in the study. Affect measures were gathered to attempt and determine if affect acted as a mediator between the flight conditions and trust ratings. The study has the following hypotheses:

H_1: Participants would be less trusting of the airline that suffered the fatal accident.

H_2: American participants would have more extreme trust ratings (both positive and negative) due to cultural differences (Rice et al., 2014; Remy, Winter, Rice, 2014; Winter, Rice, & Mehta, 2014).

H_3: Affect will at least partially mediate the relationship between the flight condition and trust ratings.

Methodology

Participants

Four hundred and two (145 females) participants from India and the United States took part in the study. The mean age was 31.55 (SD = 9.84).

Materials and Recruitment

FluidSurveys ®, a web-based survey program was used to create and develop the survey. The researchers recruited participants via Amazon's ® Mechanical Turk ® (MTurk). MTurk is a global online service that enables participants (Turkers) to participate in Human Intelligence Tasks (HITs) in exchange for monetary compensation (typically .10 to .30 cents). Participation in any HIT is voluntary and anonymous. Research (Buhrmester, Kwang, & Gosling, 2011; Germine, et al., 2012) has shown that data collected via MTurk is as reliable as normal laboratory data.

Procedure

The study received ethics board approval. Participants began the study by completing an electronic consent form. They were then presented with the following scenario: "*On June 1st, 2009, Air France Flight 447 was operating on a flight from Rio de Janeiro, Brazil to Paris, France. Approximately 3 hours after departure, contact with the aircraft was lost. It was later determined that this aircraft crashed into the Atlantic Ocean killing all on-board.*" In a separate control condition, participants were presented with the following scenario "*On August 1st, 2009, Air France Flight 445 was operating on a flight from Rio de Janeiro, Brazil to Paris, France. Approximately 11 hours after departure, the flight arrived safely in Paris without incident.*" Participants were then asked to rate their trust in airlines (Asiana, Lufthansa, American, TAM, QANTAS, Ethiopian and Air France) on a Likert-type scale from -3 (extremely distrust) to +3 (extremely trust) with a neutral option of zero (neither trust nor distrust). Participants were asked to provide information on selected demographics and were then dismissed.

Design

This research study employed a mixed design with different participants in the experimental and control conditions, different participants from the two countries, and all participants providing ratings for all the different airlines.

Results

A three-way 2 x 2 x 7 analysis of variance (ANOVA) was performed on the data, with Country and FailureNonfailure being the between-participants conditions and Airline being the within-participants condition. There was an overall 3-way interaction in the data, $F(6, 2388) = 2.60$, $p < .05$, $\eta p^2 = .01$. The interaction between

Country and Airline was significant, $F(6, 2388) = 8.70$, $p < .001$, $\eta p^2 = .02$, as was the interaction between FailureNonfailure and Airline, $F(6, 2388) = 13.76$, $p < .001$, $\eta p^2 = .03$. The main effect of Airline was significant, $F(6, 2388) = 47.52$, $p < .001$, $\eta p^2 = .11$. The main effect of FailureNonfailure was significant, $F(1, 398) = 33.83$, $p < .001$, $\eta p^2 = .08$, however the main effect of Country was not significant, $F(1, 398) = 3.01$, $p = .08$, $\eta p^2 = .01$. Figures 1 and 2 present the data for Indian and American participants, respectively.

This data suggests the following conclusions. First, there is a significant effect of trust based on the airline being rated and the difference was not significant between Indians and Americans. Second, there was a clear drop in trust of airlines in the failure condition when compared to the control condition. Third, the interactions suggest that trust ratings were significantly influenced based on the participant's country, the country in which the airline was based, and the amount of trust lost after the accident scenario. In general, Indian participants were more trusting of airlines in the control condition, and also demonstrated less of a trust drop after the accident when compared to the American participants.

Mediation Analyses

A mediation analysis was completed for Indian and American participants. Affect was not shown to mediate the relationship between trust in any of the airlines and the condition for Indian participants. For American participants, Affect was shown to mediate, at least partially, with three of the airlines: Asiana, TAM, and Air France. Figure 3 depicts the significant mediation relationships for the American participants. In order to conduct the mediation analysis, the correlation between Condition and Trust in Asiana was first found to be significant, $r = -.228$, $p = .001$, showing that the initial variable correlated with the outcome variable. The standardized path coefficients were: condition to affect (-.887, $p < .001$); affect to trust in Asiana (.369, $p = .015$); condition to trust in Asiana controlling for affect (.099; $p = .509$). These data show that Affect had complete mediation on the relationship between Condition and Trust in Asiana. The correlation between Condition and Trust in TAM was first found to be significant, $r = -.378$, $p < .001$, showing that the initial variable correlated with the outcome variable. The standardized path coefficients were: condition to affect (-.887, $p < .001$); affect to trust in TAM (.641, $p < .001$); condition to trust in TAM controlling for affect (.191; $p = .166$). These data show that Affect had some mediation on the relationship between Condition and Trust in TAM. Finally, the correlation between Condition and Trust in Air France was first found to be significant, $r = -.588$, $p = .001$, showing that the initial variable correlated with the outcome variable. The standardized path coefficients were: condition to affect (-.887, $p < .001$); affect to trust in Air France (.966, $p < .001$); condition to trust in Air France controlling for affect (.269; $p = .012$). These data show that Affect had some mediation on the relationship between Condition and Trust in Air France.

Discussion

The purpose of this study was to investigate whether consumers' trust ratings would be affected after one airline experienced a fatal accident. A unique aspect of this study was to investigate differences in culture and whether affect played a mediating role. Participants from America and India were presented with two scenarios wherein one scenario involved a fatal accident, and the other scenario had a safe flight.

The results supported the first hypothesis. The condition that suffered the fatal accident experienced a significant drop in airlines overall trust between both cultural groups. This finding supports the idea that consumers measure their trust on the system as a whole and perhaps have a hard time differentiating between separate independent components, or events in this case, when it comes to failures; the same results found in earlier studies (Geels,-Blair, Rice, & Schwark, 2013; Keller & Rice, 2010; Rice & Geels, 2010). Aviation as a whole is a safe industry. Airlines dedicate themselves to the safety of their operations, from Safety Management Systems (SMS), to recurrent training for their pilots. Consumers may not know the background information regarding the steps taken by the airlines to ensure a safe operation of their fleet. This may be a driving factor in the large drop in safety across different airlines regardless of the airline that had the accident. It is perhaps difficult for consumers to consider the fact that the accident was an unfortunate series of events that would be very unlikely to reoccur with another airline.

The results also indicated that Indians had a higher level of trust in the control condition, and also demonstrated a lower drop in trust scores compared to the Indian participants. There are some different aspects between the Indian collectivistic cultures compared to the American individualistic culture. It has been identified that the collectivistic culture is more trusting and rely on decisions and statements made by an authority figure. In

this case, it could be argued that the aviation governing bodies deem the airlines to be safe, and therefore Indian participants that relate themselves with a collectivist culture would therefore have higher trust levels. Furthermore, even though a fatal accident may occur, they may experience a lower drop in trust due to the same reasoning regarding governing bodies and the authority that may indicate that the accident was something that would not happen again (Markus and Kitayama, 1991; Robbins and Judge, 2009). This also related to previous research that has proposed that the collectivistic nature instills one not be a challenger. This therefore can influence ratings of trust, comfort, and willingness (Gaines et al., 1997; Omodei and McLennan, 2000; Shikishima et al., 2006).

It is also seen that Affect plays a role as a mediator between the condition and the trust levels of an individual, at least for a few of the conditions. This may be due to the fact that humans are emotional beings; however, for many of the conditions Affect did not mediate the relationship between trust and the condition. When faced with an accident, participants may feel their lack of trust is justified and triggered more from a cognitive perspective rather than an affective domain. Further research should be completed to identify how consumer trust levels are motivated after an airline accident occurs.

Practical Implications, Limitations, and Future Research

The findings of this study offer some interesting practical implications. First, despite any logical connection between the airline that had the fatal accident and airlines as a whole, participants from both experimental conditions expressed an overall decrease in their trust level of all airlines. The aviation industry, and in this case airlines, plays a significant role in the transportation of passengers if it were to be compared to light rail, and other forms of ground transportation. Its ability to travel far distances in short amounts of time, and the fair prices for tickets to be able to do that plays a role in why there are so many air passengers per year. Accidents do happen, but very rarely. However, due to the nature of the accident and the number of fatalities that are involved with such tragic events, the public tends to react very strongly as indicated in this research. While it can be argued that people will still fly even after an accident, the public reaction does play a role in the economics of the airlines and the aviation industry.

By understanding the cultural differences and the fact that affect plays a role, the findings of this study can be used by the airlines to see how they should promote their practices that are dedicated to their customer's safety. This study indicates that all airlines become affected with a fatal accident. This should demonstrate the need for airlines to work together to ensure safety throughout all aspects of aviation. Furthermore, the general airline consumer that reads this may see that their response to airlines as a whole after a fatal accident is purely an emotional reaction as opposed to a logical one. It would be interesting for future research to investigate whether trust levels decrease across another industry when an accident occurs. For example, it would be interesting to see if a particular an accident that involved a particular train company, such as the EuroStar, affects consumer trust across all forms of light rail. Similarly, this could be applied with cruise liners as well.

A limitation to this study was that scenarios had to be used instead of conducting the experiment in a real-world situation. However, if it were to have been conducted during the time frame of a real-world aviation accident, the results would have been skewed and affected through the history effect. Lastly, there appear to be cultural influences impacting the results of the study. Additional research could investigate this possible explanation in greater detail and attempt to examine how these and other cultural components may influence SWT. Finally, participant's familiarity with the aviation industry may play a role in their level of SWT. Previous experience with aviation or any industry that highly promotes safety may not demonstrate SWT issues. Further research should be conducted to determine how participant familiarity effects SWT and to verify the accuracy of this study's findings.

Conclusions

The purpose of this study was to examine how an airline accident at one airline could influence the trust ratings of participants across multiple airlines. System-wide trust theory suggests that operators and consumers may view multiple organizations as one system, and a negative event could pull down trust ratings in all organization, not just the accident organization. The data from this study supports that idea as trust levels across all airlines were reduced in the accident condition. Trust levels also varied by the airlines themselves. Affect was only found to mediate the relationship between the condition and trust in three of the cases with American participants and none of the cases with Indian participants suggesting that Affect may not explain much of the reason for the loss of consumer trust after an airline accident.

References

Barber, B. (1983). *The logic and limits of trust* (Vol. 96). New Brunswick, NJ: Rutgers University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(3), 3-5.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *The International Journal of Aviation Psychology, 23*(3), 245-266, DOI: 10.1080/10508414.2013.799355

Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., & Wilmer, J.B. (2012) Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5), 847-857.

Helmreich, R. L. (2000). Culture and error in space: Implications from analog environments. *Aviation, Space, and Environmental Medicine, 71(*9-11), 133-139.

Hofstede, G. (1980). Motivation, leadership and organization: do American theories apply abroad? Organizational Dynamics, 9(1): 42-63.

Keller, D. & Rice, S. (2010). System-wide versus component-specific trust using multiple aids. *The Journal of General Psychology, 137*(1), 114-128.

Lee, J. D., & See, A. K. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46,* 50-80.

Markus, H. R. & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2): 224-253.

Meyer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. Journal of Personality and Social Psychology, 49(1), 95-112. doi:http://dx.doi.org/10.1037/0022- 3514.49.1.95.

Remy, B., Winter, S. R., & Rice, S. (2014, April). *American aviation consumer's trust in pilots.* Presentation at the 7th annual Human Factors and Applied Psychology Student Conference, Daytona Beach, FL.

Rice, S. & Geels, K. (2010). Using system-wide trust theory to make predictions about dependence on four diagnostic aids. *The Journal of General Psychology, 137*(4), 362-375.

Rice, S., Kraemer, K., Winter, S. R., Mehta, R., Dunbar, V., Rosser, T. G., & Moore, J. C. (2014). Passengers from India and the United States have differential opinions about autonomous auto-pilots for commercial flights. *International Journal of Aviation, Aeronautics, and Aerospace, 1(*1), 1-12.

Robbins, S. P. & Judge, T. A. (2009). Organizational behavior (13th Ed.). Upper Saddle River NJ: Prentice Hall.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of personality*, *35*(4), 651-665.

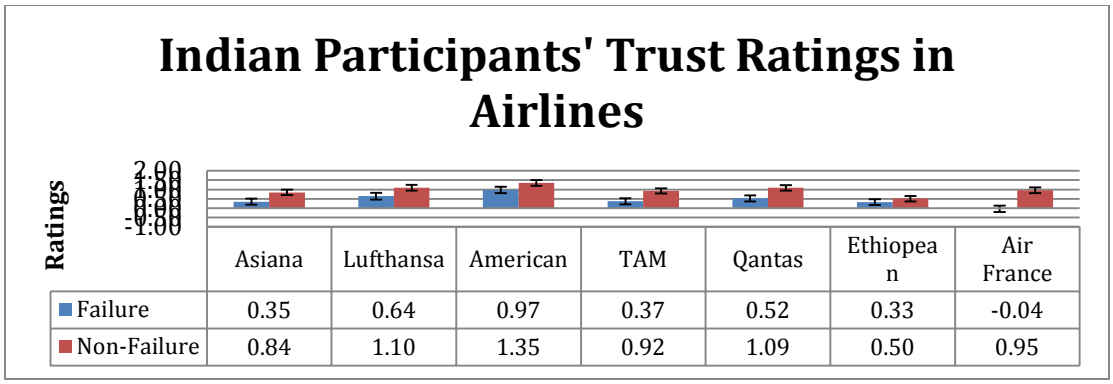
Trafimow, D., & Sheeran, P. (1998). Some tests of the distinction between cognitive and affective beliefs. *Journal of Experimental Social Psychology, 34*, 378-397.

Trafimow, D., & Sheeran, P. (2004). A theory about the translation of cognition into affect and behavior. In G. Maio & G. Haddock (Eds.), Contemporary perspectives in the psychology of attitudes: The Cardiff Symposium (pp. 57-76). London: Psychology Press.

Trafimow, D., Sheeran, P., Lombardo, B., Finlay, K. A., Brown, J., & Armitage, C.J. (2004). Affective and cognitive control of persons and behaviors. *British Journal of Social Psychology, 43*, 207-224.

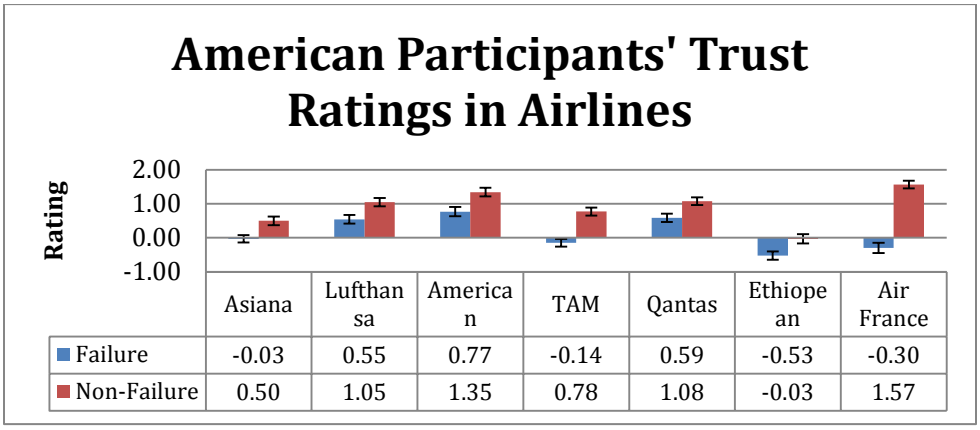
Winter, S. R., Rice, S., & Mehta, R. (2014). Aviation consumers' trust in pilots: A cognitive or emotional function. *International Journal of Aviation, Aeronautics, and Aerospace, 1(*1), 1-18.

Winter, S. R., Rice, S., & Reid, K. M. (2014, July). Using system-wide trust theory to analyze passenger loss of trust in aircraft automation. Proceedings of the 2nd International Conference on Human Factors in Transportation, Krakow, Poland.



	Asiana	Lufthansa	American	TAM	Qantas	Ethiopean	Air France
Failure	0.35	0.64	0.97	0.37	0.52	0.33	-0.04
Non-Failure	0.84	1.10	1.35	0.92	1.09	0.50	0.95

Figure 1. Trust ratings for Indian participants (SE bars included).



	Asiana	Lufthansa	American	TAM	Qantas	Ethiopean	Air France
Failure	-0.03	0.55	0.77	-0.14	0.59	-0.53	-0.30
Non-Failure	0.50	1.05	1.35	0.78	1.08	-0.03	1.57

Figure 2. Trust ratings for American participants (SE bars included).

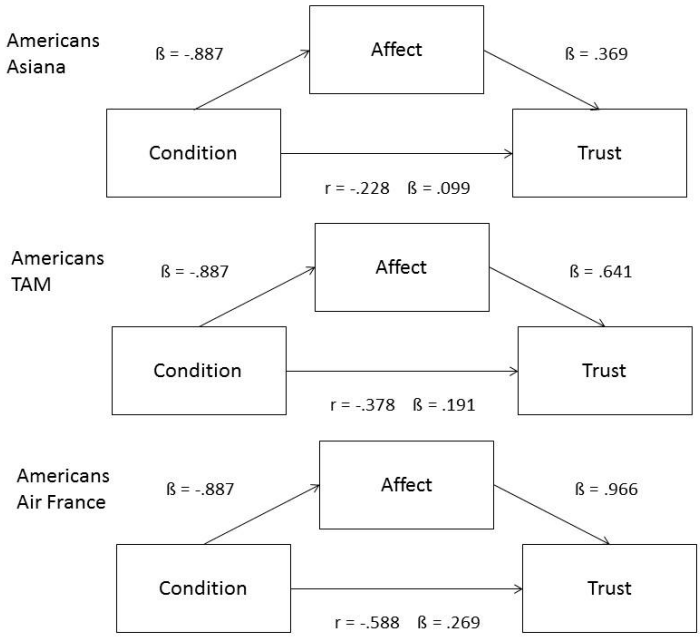


Figure 3: Path analysis for American participants.

TECHNIQUES FOR THE HUMAN CENTERED EVALUATION OF DESIGNS FOR THE FUTURE AVIATION SYSTEM

Philip J. Smith
The Ohio State University
Columbus OH

Kathy Abbott
Federal Aviation Administration
Washington D.C.

Lawrence J. Prinzel
NASA Langley Research Center
Hampton VA

Amy Pritchett
School of Industrial and Systems Engineering, Georgia Institute of Technology
Atlanta GA

Tanya Yuditsky
FAA William J. Hughes Technical Center
Atlantic City NJ

In order to evaluate new operational concepts, system designs, procedures and technologies for the future aviation system, we need to develop and validate a range of techniques to ensure the safe and effective performance of human-machine systems. This becomes increasingly important as such systems incorporate increasing levels of automation and autonomy for technologies, and as they attempt to integrate increasingly complex subsystems. It is challenging to evaluate the individual components of such systems relative to meeting their design requirements. It is orders of magnitude more challenging to evaluate performance when they are embedded in the larger system context. While there is no perfect method for such an assessment, a number of complementary techniques have been developed, applied and evaluated and will be discussed. Some can be applied early in the design process, while others focus on assessment as a system has been released for field trials or actual operations.

One theme of this panel is the need to apply a range of techniques over the development life cycle for a new system or subsystem in order to increase comprehensiveness and provide converging evidence. Methods across this range are outlined below, using examples from concrete aviation systems to help communicate the nature of the assessment methods and their actual use.

A second theme is the need to understand the strengths and weaknesses of such methods, individually and together, addressing questions such as:

- What is the state of the art?
- How good is it?
- What are the weaknesses of each individual method?
- When are they practical?
- What are the barriers to their use?

A third theme focuses on how to get better: What are the most promising directions for further developing our repertoire of techniques for verification and validation of human-machine systems, not just at the individual level but at the level of the complex, distributed work systems in aviation with a wide range of embedded technologies and forms of "automation"? Below is a summary of the topics to be addressed in this panel discussion.

Evaluating Design-Related Pilot Error

In 2013, the Federal Aviation Administration published a new regulation that requires evaluation of new aircraft flight deck systems/equipment for design-related pilot error. Good design standards must be applied as described in the regulation. In addition, the regulation recognizes that even well qualified pilots using well designed systems will make errors, so the systems/equipment designs must incorporate means to enable the pilots to manage those errors. The extent to which the system design needs to be evaluated depends on the novelty, complexity, and level of integration of the systems/equipment. Thus, discussion of future designs need to be framed in terms of regulations, methods used for complying with them, and challenges in applying them.

NASA Research Techniques for Future Aviation Systems: The Case of Synthetic and Enhanced Vision Systems

The NASA Synthetic and Enhanced Vision System (S/EFVS) is one of the enabling technologies that can provide additional margins of safety and aircrew performance in low-visibility surface, arrival, and departure operations. This work provides a case study of research techniques often employed in NASA human factors research.

Synthetic Vision Systems (SVS) use terrain/obstruction databases to present a computer rendered view of the outside world, often on a Head-Down Display (HDD). Enhanced Flight Vision Systems (EFVS) use real-time sensor input to present an enhanced visual image of the outside view on a Heads-Up-Display (HUD) or "equivalent" display, such as a Head-Worn Display (HWD). Synthetic Vision Systems (SVS) use terrain/obstruction databases to present a computer rendered view of the outside world, often on a Head-Down Display (HDD). Enhanced Flight Vision Systems (EFVS) use real-time sensor input to present an enhanced visual image of the outside view on a Heads-Up-Display (HUD) or "equivalent" display, such as a Head-Worn Display (HWD).

Research on such systems has provided an opportunity to study the use of a number of techniques, flight and simulator assets and resources, and newly developed and/or non-traditional

aviation human factors approaches for evaluating new aviation technologies and systems based on research techniques often employed in NASA human factors research. The panel presentation shall outline the various methodological approaches taken to evaluate NASA SVS and EFVS technologies.

Preventing Human Factors Problems Early in Design

Too often, human factors concerns are latent within a design because of some aspect of the underlying concept of operations. Thus, human in the loop testing late in the design cycle may find that a decision made early in the design cycle will lead to, for example, a workload spike where the pilot must quickly execute a large number of key presses to respond to unexpected air traffic controller instructions, or a situation where the pilot performing interval management will need to continuously monitor a task during already-high-workload phases of an arrival and approach. At these late stages in the design and implementation cycle, such human factors issues are often labeled as problems in the interface or with training, even when their genesis is more fundamental in the design.

Thus, it is important to consider how we can examine, early in design, what the fundamental impacts on workload, teamwork and information requirements will be in response to a new concept of operation, to new function allocations between humans and automation and/or between air and ground, and the implementation of new technologies. In particular, at the early stages of design, our models should not seek to predict what a human operator *will* do, but instead should first be checking for what the new design *will ask* the human operator to do. Particularly in the dynamic contexts inherent to aviation, this analysis needs to include computational fast-time simulation to predict when tasks will be demanded of the human operator. Such analyses can then highlight to all the designers involved where the concept of operation or underlying technological functions need to be changed.

Human Factors in the Wild

Traffic Managers continually evaluate the future status of the National Airspace System and make decisions that greatly impact its efficiency. Future systems for Traffic Flow Management will provide increased support to drive those decisions to be more precise in where they affect traffic flows and by how much. Understanding the decisions that are made today and what drives them is critical to the design of future systems, but this has been quite challenging. We typically learn about today's processes by conducting "Human Factors in the Wild:" we go to operational facilities and observe the experts in their natural habitat. This works well when we are counting steps or key presses, but not as well for deconstructing decisions. The environment is so dynamic, the options so varied, and so many factors are in play, that deconstructing the decision process becomes messy. So much of the decision making occurs in the Traffic Manager's head that observation alone is not sufficient. We propose using a modified observation approach where a Subject Matter Expert is part of the observation team and provides an interpretation of what drove the subject's decision. A version of this technique has been used successfully for identifying the drivers of operational errors by having subject matter experts review replays of traffic scenarios.

Structured Knowledge Elicitation to Envision the Impact of Future Designs

There are a variety of complementary approaches to identify potential issues human-automation design concerns associated with integration of some new component with the broader aviation system. Some involve computer modeling and some involve empirical testing or observation.

Another approach is to take advantage of the knowledge of a team of human experts to envision potential safety critical scenarios. This approach focuses on knowledge elicitation from Subject Matter Experts (SMEs) to predict potential incidents or accidents by developing scenarios where the automation embedded within some new technology could contribute to incidents or accidents.

There are several important features defining this method: First, a sequence of stages is used to progressively expose the SMEs to different types of prompts to help stimulate scenario generation. These stages use probes that are increasingly more detailed and suggestive. In the first stage, only nominal scenarios (success stories) are presented in order to avoid any biasing of the SMEs as they generate scenarios. In addition, the SMEs work individually in order to avoid having one SME influence the scenario generation by another.

In the next two stages, increasingly specific probes are presented to stimulate additional ideas for scenarios. The first set of probes uses fairly general categories from the Threat and Error Management literature; the second set provides very specific prompts for the SMEs to consider in generating scenarios, based on system design features and cognitive processes such as the potential impacts of:

- alarm prioritization
- autonomous mode changes
- inadequate knowledge of intent
- slips (errors of omission and commission).

The fourth stage finally brings the SMEs together in a focus group (individuals with relevant operational experience, human factors experts and experts in the underlying technology for the human-automation system of interest) and asks them to work together to identify additional critical scenarios. This focus group uses a variety of structured probes as well, including the presentation of historical accidents and abstract characterizations of these accidents in terms of contributing factors.

The end result is a very concrete set of scenarios predicting potential incidents or accidents for consideration by system designers.

PLANNING FOR THE FUTURE: HUMAN FACTORS IN NEXTGEN AIR TRAFFIC MANAGEMENT

Edward M. Austrian, Katherine A. Berry, Michael W. Sawyer, and Alyssa DeHaas
Fort Hill Group LLC
Washington, DC

The National Airspace System (NAS) Enterprise Architecture (EA) describes Next Generation Air Transportation System (NextGen) goals, operational changes, and guidance materials. While the primary focus of the NAS EA is on infrastructure delivery, the function of human factors is to assess and respond to the impacts of planned changes on end-users. The Federal Aviation Administration's Human Factors Research and Engineering Division has strengthened the presence of human factors activities in NextGen products in the Human Systems Integration (HSI) and other Roadmaps. This paper will present the HSI Roadmap and explore NextGen human factors integration opportunities in tower operations. Opportunities have been identified through the analysis of operational improvements (OIs), decision points, and information obtained through stakeholder interviews. When examining the tower domain and surface operations, 35 OI-actor pairings were identified with 15 describing automation enhancing situation awareness, two describing decision-support tools, one describing procedural changes, and 17 describing mixed changes.

The Federal Aviation Administration (FAA) is executing a transformation of the NAS through the implementation of NextGen. NextGen aims to increase safety, capacity, and efficiency through the introduction of new capabilities to controllers, maintainers, and other NAS users (FAA, 2012). Guiding the implementation of these NextGen capabilities is the NAS EA. The FAA's NAS EA serves as a blueprint for top-down operational and NAS infrastructure improvements. It establishes "a foundation from which the evolution of the NAS can be explicitly understood and modeled" (FAA, 2015). Within the NAS EA are two sets of roadmaps – Service and Infrastructure Roadmaps. Service Roadmaps depict the evolution of current air traffic services to meet future NAS demands through the implementation of OIs. Also to meet future NAS demands, Infrastructure Roadmaps depict the evolution of NAS infrastructure through decision points and regulatory milestones (FAA, 2015). Together, this information enables users to develop a comprehensive, integrated understanding of NextGen changes as well as potential air-ground human factors opportunities (Austrian & Piccione, 2013).

The HSI Roadmap (sample in Figure 1) is the only actor-centric roadmap in the NAS EA. The HSI Roadmap Version 8.0 (FAA, 2014) depicts the evolution of air traffic control (ATC), technical operations, and aviation industry NAS actors by highlighting changes to user-specific technologies and procedures over time. These changes are illustrated through the depiction of key NextGen decisions, milestones, and strategic activities. The identification of actor-NAS EA data element relationships enables the Human Factors Research and Engineering Division (ANG-C1) to define new opportunities for future NextGen human factors research in support of

FAA infrastructure and NextGen capability delivery. Additionally, the HSI Roadmap supports the need to "ensure that human factors issues are fully integrated throughout the development of NextGen systems" by providing a tool to coordinate key enterprise-level human factors activities and needs with relevant stakeholders (GAO, 2010).

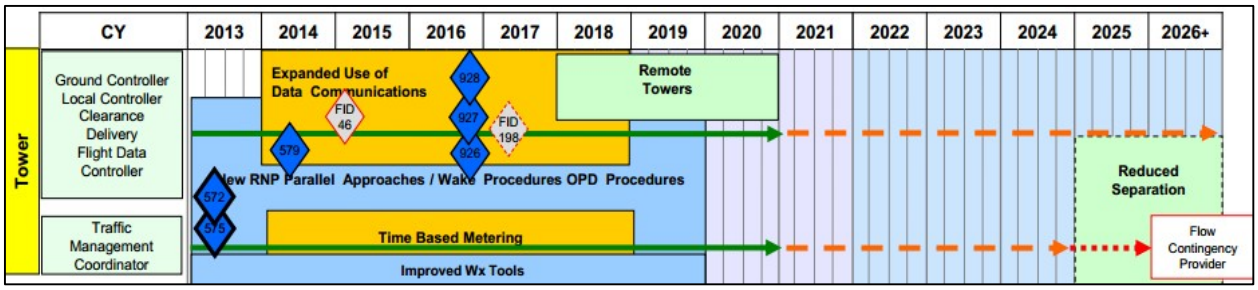


Figure 1. ATCT Portion of the HSI Roadmap Version 8.0 (FAA, 2014). Modified for printing.

Purpose

The concurrent development and implementation of NextGen changes must consider the system- wide impacts to all NAS actors (Zemrowski & Sawyer, 2010; Berry & Pace 2011). This paper aims to utilize the relationships defined in the HSI Roadmap to identify and classify mid- and far-term NextGen human factors research opportunities. These opportunities can be employed to prioritize potential human factors NextGen contributions. As a part of a larger initiative, this paper will present the findings of the HSI Roadmap analysis for the airport traffic control tower (ATCT) domain.

Methodology

During the annual HSI Roadmap update process, data was gathered from the NAS EA Portal and through stakeholder interviews to support HSI Roadmap development and derivation of potential future research opportunities. From the 2015 NAS EA Portal data, 77 OIs were analyzed and classified by a panel of ATC, flight deck, and human factors subject matter experts who utilized a consensus methodology to determine the impacts of OIs on ATCT operations. From those OIs directly impacting ATCT operations, the panel first identified the NAS actor from the ATCT domain (ground controller, local controller, ATCT traffic management coordinator (TMC), and pilot) directly impacted by the NextGen improvement being introduced to the NAS. The panel then determined the specific human factors change to current operations associated with each OI. Those human factors changes were then classified for each ATCT actor as either:

- Situation Awareness (SA) Automation
- Decision Support (DS) Automation
- Procedure Change
- Mixed Change (a combination of two or more of the above change classifications)
- No Human Factors Impact

In addition to the OIs, 139 decision points and regulatory milestones were analyzed through one- on-one working sessions with stakeholders. Each of the 139 decision points and regulatory milestones were linked to the ATCT actors by NextGen timeframe. Decision points

and regulatory milestones represent key infrastructure acquisitions or regulatory changes that could impact ATCT operations and NAS actors at specific points in time.

NAS EA Data Elements

NextGen OIs capture a collection of capabilities that will be incrementally deployed to deliver a variety of benefits to users. To accurately capture cross-cutting NextGen capabilities, OI descriptions are service-focused and lack a direct linkage to NAS infrastructure or systems. Complimenting the OIs and other NextGen data elements are NextGen decision points and regulatory milestones, which capture specific NAS infrastructure investments, acquisitions, or related operational activities that have a clearer linkage to NextGen capabilities. Based on these relationships, it is assumed that OIs may assist in the definition of future NextGen capabilities. These capabilities may drive future NAS infrastructure investments and related changes. As such, both data elements were included in this analysis to obtain a comprehensive understanding of potential NextGen impacts on the ATCT domain and related actors.

Results

Table 1 shows the number of classified OIs that have potential to introduce changes to ATCT actors in the NextGen mid- and far-term.

Table 1.
ATCT OI Classification Analysis Findings

NAS Actor	Total OIs			Human Factors OI Classifications			
	Mid-Term	Far-Term	Total	SA Automation	DS Automation	Procedure Change	Mixed Change
Ground Controller	5	2	**7**	4	1	0	2
Local Controller	9	4	**13**	3	1	1	8
ATCT TMC	8	2	**10**	7	0	0	3
Pilot	3	2	**5**	1	0	0	4

Table 2 shows the number of decision points and regulatory milestones that may impact ATCT actors in the NextGen mid- and far-terms.

Table 2.
ATCT Decision Point and Regulatory Milestone Findings

NAS Actor	Total Decisions / Regulatory Milestones		
	Mid-Term	Far-Term	Total
Ground Controller	5	2	**7**
Local Controller	10	2	**12**
ATCT TMC	11	2	**13**
Pilot	24	10	**34**

Discussion

The ATCT OI classification analysis results revealed that seven OIs will impact the ground controller, 13 OIs will impact the local controller, 10 OIs will impact the ATCT TMC, and five ATCT OIs will impact the pilot. Below are examples of impact classification results by

actor. The listed examples also show how an individual OI (e.g., Improved Parallel Runway Operations) may simultaneously impact more than one NAS actor.

OI: Initial Integration of Weather Information into NAS Automation and Decision Making
Related ATCT Actor: ATCT TMC
HF Impact Classification: Situation Awareness Automation
Classification Rationale: This OI proposes the introduction of improved weather data quality and availability to controllers and NAS stakeholders. Users will have the ability to access tailored weather information that enables informed, collaborative decision-making that supports the timely initiation of group or individual flight re-planning actions.

OI: Initial Surface Traffic Management
Related ATCT Actor: Ground Controller
HF Impact Classification: Decision Support Automation
Classification Rationale: This OI proposes the introduction of automation enhancements that support controller surface movement decisions. Automation will share surface information across NAS systems and integrate departure sequencing times with surface movement information to support the prioritization of aircraft staging.

OI: Improved Parallel Runway Operations with Airborne Applications
Related ATCT Actor: Local Controller
HF Impact Classification: Procedure Change
Classification Rationale: This OI proposes the introduction of policies, procedures, and standards that support the use of advanced aircraft avionics to fly dependent approaches to closely spaced parallel runways while maintaining designated spacing intervals.

OI: Improved Parallel Runway Operations with Airborne Applications
Related ATCT Actor: Pilot
HF Impact Classification: Mixed Change
Classification Rationale: This OI proposes the introduction of policies, procedures, and standards that support use of advanced aircraft avionics to fly dependent approaches to closely spaced parallel runways while maintaining designated spacing intervals.

Collectively, the implementation of the analyzed OIs could equip ATCT NAS actors with the tools, capabilities, and information to maintain an increasingly accurate and up-to-date view of high density airport and system-wide NAS operations. Further supporting this assumption are specific NAS infrastructure changes that are detailed through NextGen decision points and regulatory milestones. Below are sample decision points that were included in this assessment and linked to ATCT NAS actors:

Decision Point 46: Final Investment Decision (FID) for Terminal Flight Data Manager (TFDM)
Decision Point 198: FID for TFDM Segment 2
Related ATCT Actors: Ground Controller, Local Controller
Rationale: TFDM is a phased FAA acquisition program that will deliver NextGen capabilities and decision support tools to ATCT NAS actors (ground, local, etc.) throughout the NextGen mid- and far-terms. Decision support capabilities will enable the integration of surface, flight, and traffic management information. Additionally, TFDM will introduce electronic flight strips, surface traffic management, and scheduling capabilities to ATCT NAS actors.

Decision Point 927: Decision on the Implementation Strategy of NextGen ATC Alarms, Alerts, and Notification Guidance
Related ATCT Actors: Ground Controller, Local Controller, ATCT TMC
Rationale: Decision point 927 represents a collection of human factors products that are actively being conducted in support of future NAS infrastructure delivery. Specifically, the development of alarms and alerts standards and guidance materials could influence Terminal and En Route system designs.

Potential NextGen Human Factors Opportunities

Many NextGen improvements aim to improve airport and airspace capacity through the implementation of increasingly complex air-ground procedures. These procedures are dependent on NAS infrastructure improvements, conditional amendments to legacy separation requirements, upgraded aircraft technologies, flight crew eligibility, and strict aircraft procedural conformance. Potential NextGen human factors research opportunities have been identified through the analysis of NextGen OIs, NAS EA decision points and regulatory milestones, and information obtained through stakeholder interviews. These potential research opportunities have been categorized as either surface operations or closely spaced runway operations based on operational relevance.

Surface Operations. Several NextGen changes leverage improved air-ground surveillance technologies, decision support automation, and new procedures to further increase surface efficiencies during adverse weather conditions. To-date, a large portion of NextGen research has focused on mid-term concepts. For far-term concepts, research opportunities will be examined as the concepts mature. Potential research areas may include the development of information integration needs, air-ground information management strategies, and identification of controller-to-controller and controller-pilot information needs. Potential concepts may include data communications during surface operations, utilization of air-ground automation and related technologies to support surface conformance monitoring, examination and prioritization of new air-ground alerting functions, and identification of individual and integrated far-term off-nominal conditions.

Closely Spaced Runway Operations. Several NextGen changes aim to conditionally reduce Terminal aircraft-to-aircraft separation requirements to increase high density airport throughput. Many of these mid-term concepts have been examined individually. However, opportunities exist to understand the integrated impacts of these concepts on human performance. Potential research areas may include the development of controller-controller and controller-pilot information needs to support individual and collaborative decision-making during critical phases of flight. Research aimed at understanding perceived air-ground workload and procedural complexity during closely spaced parallel runway operations could support the implementation of NextGen concepts. Terminal information needs will also be required to support the seamless integration of unmanned aircraft systems (UAS) into daily NAS operations. Potential concepts that may benefit from this research include interval management-spacing, UAS surface operations, and UAS terminal airspace operations. This research could support the integration and implementation of multiple operational concepts to enable closely spaced runway operations.

Conclusion

The 2014 HSI Roadmap is the only roadmap in the NAS EA to be actor driven. As such, the HSI Roadmap may be used as a tool to develop NextGen human factors dependencies and a method to drive the identification of future potential human factors opportunities. Multiple NextGen human factors opportunities exist to support the successful delivery of NextGen infrastructure and capabilities throughout the mid- and far-terms. Functionally, the HSI Roadmap may be used as means to identify those opportunities and proactively close NAS-wide human factors gaps.

Acknowledgements

We would like to acknowledge the FAA's Human Factors Research and Engineering Division (ANG-C1) for funding this project and similar work. Additionally, we would like to acknowledge the air traffic control and human factors subject matter experts who provided the valuable insight necessary to developing these results. The results presented herein represent the results of this research project and do not necessarily represent the view of the Federal Aviation Administration.

References

Austrian, E. & Piccione, D. (2013). The FAA's human factors air traffic control/technical operations strategic research plan. In *Proceedings of the International Symposium on Aviation Psychology,* Dayton, OH.

Berry, K. & Pace, J. (2011). Examining the actors and functions of an airline operations center. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Las Vegas, NV.

FAA. (2012). *Destination 2025*. Retrieved from http://www.faa.gov/about/plans_reports/media/Destination2025.pdf

FAA. (2015). *NAS Enterprise Architecture.* Retrieved from https://nasea.faa.gov.

FAA. (2014). *NAS EA Human Systems Integration Roadmap, Version 8.0*. Retrieved from https://nasea.faa.gov/products/roadmap/main/display/9

GAO (2010). *Next Generation Air Transportation System: FAA and NASA Have Improved Human Factors Research Coordination, but Stronger Leadership Needed.* Retrieved from http://www.gao.gov/ products/GAO-10-824

Zemrowski, K., & Sawyer, M. (2010). Impacts of increasing reliance on automation in air traffic control systems. *Journal of Air Traffic Control, 52*(4), 49-55.

ASSESSING POTENTIAL HUMAN PERFORMANCE SAFETY IMPACTS ASSOCIATED WITH INTEGRATING MULTIPLE TIME-BASED FLOW MANAGEMENT CONCEPTS

Michael W. Sawyer, Katherine A. Berry, Devin Liskey, and Richard Rohde
Fort Hill Group, LLC
Washington, DC

Time-Based Flow Management (TBFM) seeks to enhance system efficiency by improving scheduling and interval management tools that expand and enhance the flow of traffic. This paper presents the results of an integrated human performance safety assessment of TBFM concepts planned for implementation between 2016 and 2020. The assessment utilized the Human and Organizational Safety Technique (HOST) to provide a structured method for identifying potential human error modes and estimating their contribution to the risk profile. HOST further defined graphical human-system interaction models for each proposed change and an integrated interaction model across all assessed TBFM changes. The HOST assessment resulted in multiple potential human error modes across the individual TBFM changes. As no concept will be implemented in isolation, key interactions among error mode interactions were assessed utilizing the integrated human and system interaction model. These findings directly influenced the development of Next Generation Air Transportation System (NextGen) safety requirements.

The Federal Aviation Administration (FAA) is currently executing a considerable transformation of the National Airspace System (NAS). NextGen aims to improve the convenience and dependability of air travel while increasing safety and reducing environmental impacts. NextGen plans to meet these goals by introducing a variety of new systems and capabilities (FAA, 2013b). While NextGen may produce many positive safety improvements, the introduction of each new system and capability also offers the possibility of increasing the human contribution to risk in the NAS (Berry & Sawyer, 2013). This is especially true when considering the system-wide impact and concurrent development of many of the systems (Sawyer, Berry, & Blanding, 2011; Zemrowski & Sawyer 2010). From a risk management perspective, research into these effects is needed to address the potential for both positive and negative impacts on the safety of the NAS (FAA, 2013a).

This paper presents the results of an integrated human performance and safety assessment of TBFM concepts planned for implementation between 2016 and 2020. Proactive assessment of NextGen concepts is critical to understanding the cross-cutting impacts of proposed changes to the human-system interactions among all NAS actors (Austrian, Berry, Sawyer & DeHaas, 2015; Austrian & Sierra, 2013).The resulting hazards and overall human performance risk profile are provided to support the development of targeted mitigation strategies to ensure that new NextGen capabilities support human performance.

Time-Based Flow Management (TBFM)

TBFM proposes to enhance system efficiency by leveraging the capabilities of the Traffic Management Advisor (TMA) decision-support tool. TMA has already been deployed across Air Route Traffic Control Centers in the contiguous United States. Proposed NextGen improvements to TMA will improve its trajectory modeler, enhance TMA's departure capabilities, and optimize demand and capacity. Improvements will be made to enable controllers to more accurately deliver aircraft to the Terminal Radar Approach Control (TRACON) facility while also providing the opportunity for aircraft to fly optimized descents.

In the NextGen mid-term timeframe, this portfolio focuses on scheduling and interval management tools that further expand Time-Based Metering benefits to assure the smooth flow of traffic and increase the efficient use of airspace. These changes will be implemented through a series of improvements such as Point-in-Space Metering, Time-Based Metering in the Terminal Environment, and Improved Management of Arrival/Surface/Departure Flow. These changes are designed to extend, enhance, and increase metering operations; improve the accuracy of schedules and demand predictions for more efficient and predictable NAS operations; and continue the path toward trajectory-based operations. These changes also introduce the use of Interval Management-Spacing operations, using a combination of ground- and flight deck-based capabilities. These changes are described through a series of Operational Improvements (OIs).

Human Performance Hazard Assessment

A human performance safety assessment was conducted on these OIs to proactively identify potential positive or negative impacts to a controller's ability to provide safe air traffic control services. For the purpose of this assessment, four OIs from the TBFM portfolio were examined. The OIs listed below in Table 1 were retrieved from the FAA's NAS System Enterprise Architecture (FAA, 2014).

Table 1.
Time-Based Flow Management NextGen Operational Improvements

Operational Improvement
102118: Interval Management - Spacing
104120: Point-in-Space Metering
104123: Time-Based Metering Using RNAV and RNP Route Assignments
104128: Time-Based Metering in the Terminal Environment

Human and Organizational Safety Technique (HOST)

The assessment utilized HOST to assess planned TBFM changes. HOST provides a structured method for identifying potential human error modes and estimating their severity, likelihood, and recovery potential. As part of the HOST analysis, a team of air traffic control, commercial aviation, and human factors subject matter experts reviewed each OI to first identify the controller and pilot tasks impacted by the proposed change. Identified task impacts were then

used to develop Human System Interaction Models (HSIMs) for each OI. Each HSIM depicts the interactions of actors and systems for a given change as identified by the working group.

Following the development of the HSIMs, the working group reviewed each impacted actor/system interaction point to identify potential human performance hazards associated with not executing the action, completing the interaction in an unsafe way, and completing the action too soon or too late. Worst credible outcomes for each potential hazard were then identified and assessed based on the potential severity, likelihood, and recovery impact of each hazard. Resulting hazards were then prioritized based on potential impact to human operators in the NAS.

Results and Discussion

The HOST analysis of the TBFM portfolio resulted in two primary results. Phase one of HOST yielded HSIMs for each of the OIs and an integrated HSIM showing the interactions among the four OIs in the portfolio. Each HSIM was then used as the basis for identifying potential hazard conditions associated with each impacted actor-system interaction. Selected results will be presented for OI 104123 followed by aggregated results across all TBFM OIs.

104123: Time-Based Metering Using RNAV/RNP Route Assignments

Provided below in Figure 1 is the completed HSIM for 104123: Time-Based Metering Using RNAV and RNP Route Assignments. The development of the HSIM outlined the key impacts of the proposed change on the en route controller. Primary impacts were seen on the interaction between the en route controller and the en route automation system and flight crew.

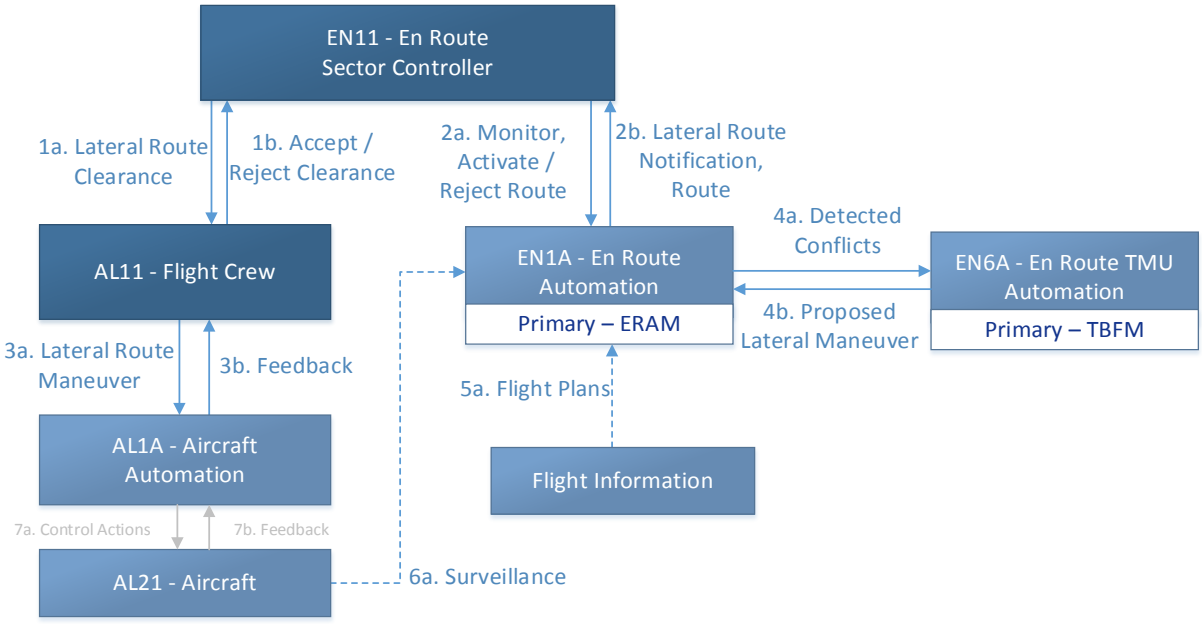


Figure 1. HSIM for 104123: Time-Based Metering Using RNAV and RNP Route Assignments

Following the development of the HSIM for each OI, the human performance hazard assessment was completed to identify potential human performance hazards introduced or

impacted by each OI. The working group identified 10 potential human performance hazards associated with OI 104123. An example of an identified human performance hazard and the associated impact data is included below in Table 2.

Table 2.

Example Hazard identified for OI 104123

Impacted Task	2b. En Route Automation provides controller with lateral route instruction generated to help aircraft meet meter time		
Potential Error Mode	Automation provides controller with path stretching instruction with inadequate return point		
Worst Credible Outcome	During path stretching operation, actual aircraft performance and airspace conditions differ from predicted traffic flow and conditions. Return point no longer provides required separation minima for path stretching aircraft.		
Hazard Actor	Automation	**Hazard Activity**	Non-Controller Task
Outcome Actor	En Route Controller	**Outcome Activity**	A6: Manage Traffic Flows and Sequences
Effect Type	Safety	**Human Factors Priority**	Moderate

Aggregated Human Performance Safety Impacts

Across the four OIs included in this assessment, 48 human performance hazards were identified. These hazards and their evaluated priorities are included below in Figure 2. OI 102118: Interval Management – Spacing showed the largest number of potential human performance hazards. This included the one high human factors priority hazard which related to the impact of a deviation by the lead aircraft in an interval management pair. Identified hazards demonstrated the potential impact to the controller's ability to detect and resolve spacing issues when aircraft are actively managing the interval spacing behind another aircraft.

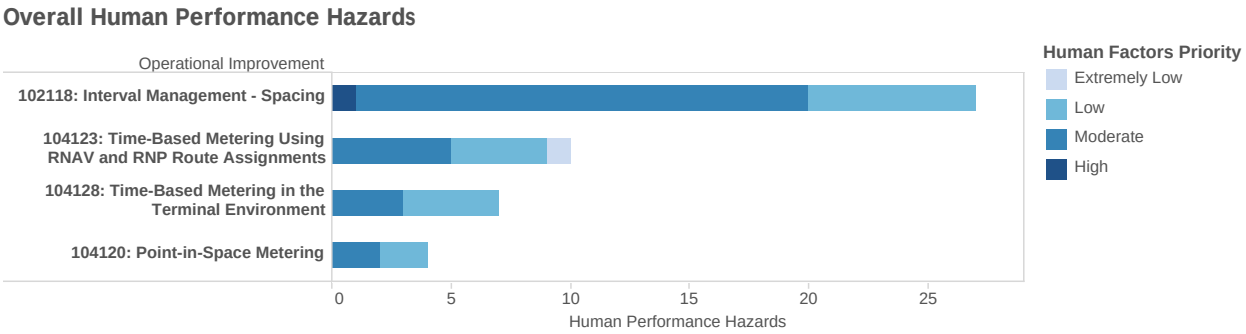


Figure 2. Overall Human Performance Hazards

Impacted Controller Tasks. The controller tasks associated with introducing a hazard (Hazard Task) and the tasks associated with mitigating each hazard (Outcome Task) were also identified for each of the 48 hazards. Impacted hazard tasks for TBFM OIs are included in Figure 3. Results indicated that the majority of identified hazards were introduced, not through controller tasks, but by actions initiated either by automation or by the actions of the flight crew. Of the impacted controller tasks, most hazards related to the controller developing and managing traffic flows and sequences. Many of these hazards relate to the controller interpreting the information provided by the automation to develop and implement a traffic flow.

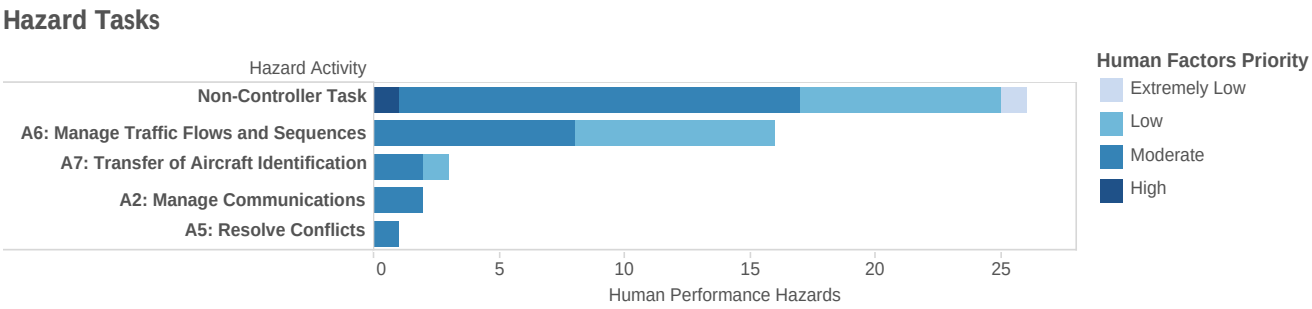


Figure 3. Identified Hazard Tasks for TBFM OIs

The outcome tasks associated with resolving or recovering from each human performance hazard are shown in Figure 4. While many of the hazards were initiated by non-controller tasks, 44 of the 48 hazards will require controller actions to resolve or recover from the impact of the hazard task. Many of the potential hazards will require controllers to resolve traffic flow.

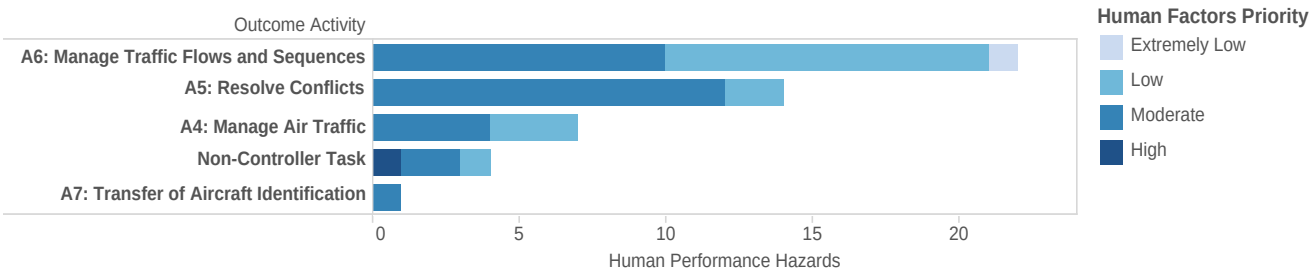


Figure 4. Identified Outcome Tasks for TBFM OIs

Inter-Actor Relationships. A comparison of hazard actor to outcome actor is shown in Figure 5. Breaking down the relationships between the hazard actor and outcome actor provides an overview of the critical human-system interactions necessary to support the implementation of these planned changes. As previously identified, the majority of hazards potentially introduced with these hazards will require the controller to mitigate. En route and TRACON controllers will be responsible for resolving the majority of the identified automation hazards. As many of these OIs propose additional automation tools to support scheduling and sequencing, system designers will need to ensure controllers are provided the necessary information to recover from hazards once they occur. Designers should ensure controllers consistently have access to the information needed to update their traffic flows for cases where automation has provided an adequate sequence.

A second key finding of this assessment revolved around the impact of flight crew actions on the en route controller's ability to implement TBFM concepts. As the flight crew begins playing a more active role in meeting meter times with concepts like interval management, the impact of an error by one flight crew can now impact other aircraft and the controllers monitoring their performance. The potential for vigilance decrements associated with decreasing controller-pilot interactions and reduced controller workload could further increase the consequences of a pilot error.

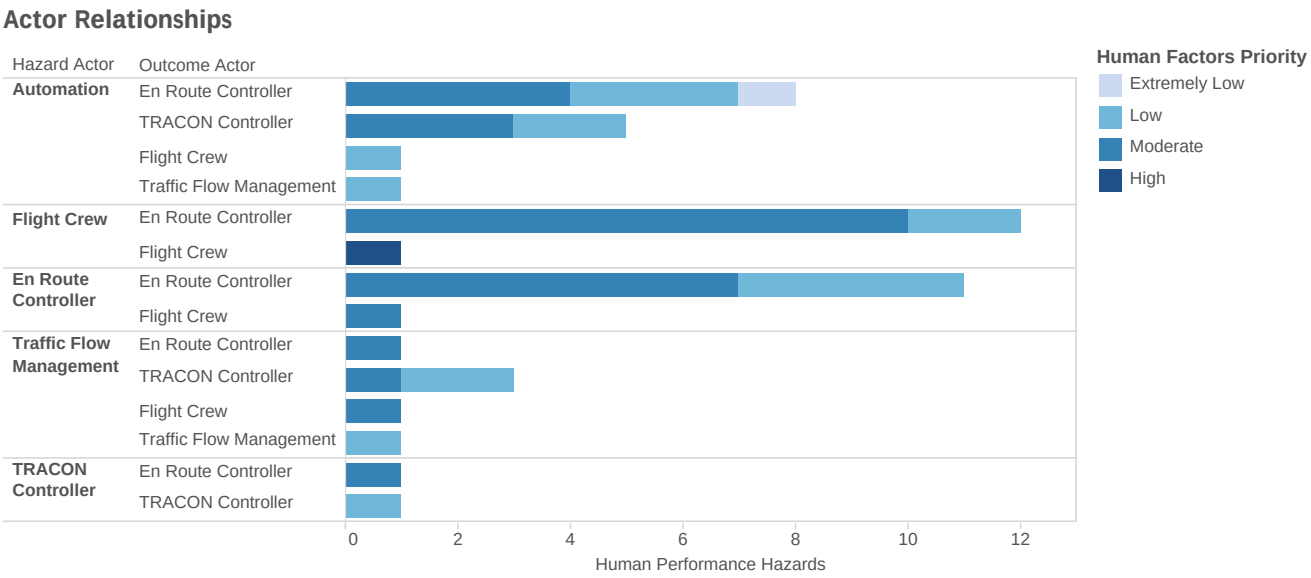


Figure 5. Actor Relationships for TBFM OIs

Acknowledgements

We would like to acknowledge the FAA's Human Factors Research and Engineering Division (ANG-C1) for funding this project and similar work. Additionally, we would like to acknowledge the air traffic control and human factors subject matter experts who provided the valuable insight necessary to developing these results. The results presented herein represent the results of this research project and do not necessarily represent the view of the FAA.

References

Austrian, E., Berry, K., Sawyer, M., & DeHaas, A. (2015). Planning for the Future: Human Factors Human Factors in NextGen Air Traffic Management. In *Proceedings of International Symposium on Aviation Psychology*, Dayton, OH.

Austrian E. & Sierra, E. (2013). Applied Human Factors Research For The Technical Operations Organization of the Federal Aviation Administration. In *Proceedings of International Symposium on Aviation Psychology*, Dayton, OH.

Berry, K. & Sawyer, M. (2013). Assessing the Changing Human Performance Risk Profile in the NextGen Mid-term. In *Proceedings of International Symposium on Aviation Psychology*, Dayton, OH.

FAA. (2014). NAS Enterprise Architecture. Retrieved from https://nasea.faa.gov.

FAA. (2013). Destination 2025. Retrieved 2013 from www.faa.gov/about/plans_reports/media/Destination2025.pdf.

Sawyer, M, Berry, K., & Blanding, R. (2011). Assessing the Human Contribution to Risk in NextGen. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, Las Vegas, NV.

Zemrowski, K., & Sawyer, M. (2010). Impacts of Increasing Reliance on Automation in Air Traffic Control Systems. *Journal of Air Traffic Control, 52*(4), 49-55.

COMPUTATIONAL SIMULATION OF AUTHORITY-RESPONSIBILITY MISMATCHES IN AIR-GROUND FUNCTION ALLOCATION

Martijn Ijtsma
Delft University of Technology
Delft, the Netherlands

Amy R. Pritchett and Raunak P. Bhattacharyya
Georgia Institute of Technology
Atlanta, GA USA

Authority-responsibility mismatches are created when one agent is authorized (has authority) to perform an activity, but a different agent is responsible for its outcome. An authority-responsibility mismatch demands monitoring by the responsible agent that itself requires additional information transfer and taskload. This paper demonstrates a computational simulation methodology that identifies when mismatches will occur in complex, multi-agent aviation operations, and their implications for information transfer between agents and task demands on each agent. A case study examines 25 authority and responsibility allocations in a NextGen/SESAR scenario in a terminal area where authority and responsibility for activities involving optimal profile descents, merging and spacing can be fluidly allocated to the aircraft (pilot/flight management system) or to the ground (air traffic controller/controller decision aids and automation).

Human factors needs to be involved early in design. We propose the early intervention of analyzing for, and preventing, air traffic and flight deck concepts of operation that place unreasonable demands on any agent, particularly the pilot and/or air traffic controller. Such demands may include requiring too much taskload, or assigning tasks that require substantial information transfer between agents, or implicitly creating additional monitoring tasks.

A concept of operation defines which actions must be performed in complex multi-agent systems and which agents – human or automated – have authority and responsibility to perform these actions. In this paper the following definitions are used: *Authority* is the requirement for an agent to execute a task, and *Responsibility* is the designation of accountability for the outcome of a task, in an organizational, regulatory and legal sense. Authority and responsibility do not always need to be aligned. Authority-responsibility mismatches, as first identified by Woods (1985), occur whenever one agent is authorized to execute a task, but a different agent is responsible for the outcome. As a result of the mismatch, the responsible agent needs to get information about the task outcome (and perhaps performance), monitor the authorized agent, and perhaps intervene. Thus, when the function allocation within a concept of operation generates authority-responsibility mismatches, it also implicitly creates additional information transfer and monitoring-taskload beyond that visible when only the authority allocation is examined.

Feigh and Pritchett (2014) distinguish between taskwork (required to achieve common work goals regardless of function allocation) and teamwork (required to coordinate between

agents within a specific function allocation). Function allocation methods to date typically look at the allocation of authority, typically focusing on the taskwork (e.g. Wing et al., 2010; Scallen & Hancock, 2001), but the allocation of responsibility must also be considered to properly predict the teamwork demands that will emerge during the actual operation. This teamwork includes the information transfer and monitoring resulting from authority-responsibility mismatches.

Predicting the demands on any person in a novel, complex, multi-agent concept of operation is difficult. For example, earlier studies have shown that, in a chain of aircraft performing flightdeck interval management, the timing of information transfer and taskload changes from the first aircraft in the chain to subsequent aircraft that have to respond to the aircraft ahead of them (IJtsma, Bhattacharyya, Pritchett & Hoekstra, Submitted; Bhattacharyya & Pritchett, 2014). Thus, in this paper we demonstrate how simulation can predict such emergent effects. Here, we focus on authority-responsibility mismatches and their commensurate task load. We demonstrate the general method in the specific context of a terminal area where 25 different allocations of authority and responsibility are fluidly made for activities involving optimal profile descents, merging and spacing, changing whether they are allocated to the aircraft (pilot/flight management system) or to the ground (air traffic controller/decision aids and automation).

Computational Simulation of Authority and Responsibility Allocation

Work Models that Compute (WMC) is an open-source simulation platform written in C++ that can dynamically model complex, multi-agent concepts of operation (Pritchett, Feigh, Kim & Kannan, 2014). WMC is unique in the sense that the model of work is independent of the agent models, allowing for the fluid allocation of activity to different agents.

Work models describe the collection of tasks that together achieve common goals. The tasks are modeled such that each represents an action that can be completed by a single agent at a single point in time. In this case study, to isolate the effect of function allocation, actions are executed the same way regardless of the authority allocation. Additionally, to isolate the demands placed on each agent by the function allocation, actions are executed without errors and delays. After this preliminary evaluation, more detailed analysis can evaluate human performance in the concept of operation (Pritchett, Feigh, Mamessier & Gelman, 2014).

A function allocation is represented by which actions are allocated to which agents for both authority and responsibility in any simulation run – or at any particular instant within a run. Mismatches in authority and responsibility manifest themselves through extra monitoring actions. In real operations these monitoring actions are created implicitly when the need for them emerges; correspondingly, they are created automatically during a simulation whenever the simulation framework detects an authority-responsibility mismatch. In this paper the monitoring actions are empty placeholders that serve to identify taskload and information transfer requirements, but the simulation framework also allows for any action to specify functions that represent more elaborate monitoring activities appropriate to its own situation.

Table 1.
Authority and responsibility allocations (A = Air, G = Ground).

Functional blocks	Authority allocations (AA)					Responsibility allocations (RA)				
	1	2	3	4	5	1	2	3	4	5
Vertical profile control	G	A	A	A	A	G	A	A	A	A
Aircraft configuration management	G	A	A	A	A	G	A	A	A	A
Lateral control	G	A	A	A	A	G	A	A	A	A
Speed control	G	G	A	A	A	G	G	A	A	A
Lateral profile management	G	G	G	A	A	G	G	G	A	A
Vertical profile management	G	G	G	G	A	G	G	G	G	A
Speed management	G	G	G	G	A	G	G	G	G	A
Non-nominal situation management	G	G	G	G	A	G	G	G	G	A

WMC logs the exact time instances when an action is performed, and the executing agent for that action. Additionally, the simulation logs any time instances when an agent requires knowledge of information that is set by a different agent: these instances reflect a requirement for information transfer, and are from here on referred to as information transfer requirements.

Case Study

This case study builds on an earlier study of authority allocation between air- and ground-based operators in a NextGen/SESAR terminal area (IJtsma et al, 2014). Three aircraft are arriving into Schiphol Airport RWY18R with the lead aircraft performing an Optimum Profile Descent (OPD) and subsequent aircraft performing in-trail and merging interval management (IM). One aircraft enters the airspace from the West and initially performs an OPD along the RIVER arrival route. Two other aircraft enter from the East and follow the ARTIP route, where the first aircraft initially performs an OPD and the second aircraft follows at a 60 second time interval through IM. The two traffic streams later merge. An off-nominal situation can be introduced wherein the RIVER aircraft requests priority to land (e.g. medevac flight) and the other two aircraft need to maneuver to sequence behind it at the merge point. Thus, four agents are simulated: the three flight crews (FC) and one air traffic controller (ATC). The agent models are deliberately "perfect" in that they execute actions immediately and without error, so that any concerns with the underlying concept of operation can first be clearly isolated.

The model groups similar actions together into functional blocks (IJtsma et al., 2014). The functional blocks are allocated to either the FC or ATC agents. Five authority (AA) and five responsibility allocations (RA), both conventional and non-conventional, are analyzed, as shown in Table 1. Each authority allocation is tested with each responsibility allocation, thereby resulting in 25 complete function allocations. Monitoring actions are automatically spawned whenever an authority-responsibility mismatch is present and is assumed to be perfect in the sense that whenever an action is executed, the responsible agent will instantly monitor the executing agent.

Results

To illustrate the detailed analysis that WMC affords, Figure 1 shows an action time trace for the air traffic controller with AA3 and RA2. The ATC agent experiences high task load in peaks, particularly between 350 and 480 seconds into the simulation. Additionally, there are three moments in time when heavy monitoring is required, starting at 100 s, 180s and 480 s.

To provide more aggregate results, summing up all taskwork and monitoring actions within each combination of allocations of authority and responsibility results in Figure 2. As may be predicted, increasing authority allocation to the flight crew results in a higher task load for the flight crews and a lower task load for the ATC agent. The monitoring required of each agent, on the other hand, results from the combination of allocations of authority and responsibility: where mismatches occur, monitoring results. Put together, the total demands on the agent – explicit taskload and implicit monitoring – is driven more by responsibility allocation than by authority.

Similarly, Figure 3 shows the total amount of information transfer, discriminating between transfers stemming from taskwork versus monitoring. A wave pattern can be observed in the information transfer stemming from the taskwork, wherein authority allocations that divvy up the work equally between air and ground result in high information transfer wherever their assigned actions need to coordinate; conversely, an agent that is allocated authority for everything doesn't need to ask for information set by others' activities. On the other hand, information transfer for monitoring shows a similar trend as the taskload results in Figure 2: it is driven by mismatches.

Conclusion

This paper demonstrates that function allocation should not just consider the distribution of authority, but also of responsibility, particularly to identify authority-responsibility mismatches. These mismatches implicitly create additional monitoring tasks for the responsible agent, and should be included in human factors analysis. Thus, computational simulation of concepts of operation can provide quantitative insight in the task load, monitoring and information transfer demands resulting from function allocations, including authority-responsibility mismatches.

This methodology can be used to objectively assess function allocations early in the design process and subsequently guide the further design of the concept of operation to prevent human performance issues. These results, thus, can highlight situations where a so-called "human factors issue" is actually inherent in a concept of operation, regardless of training or

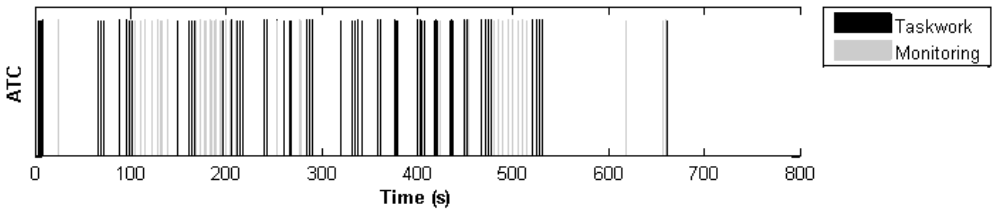


Figure 1. Action time trace for the ATC agent with AA2 and RA3.

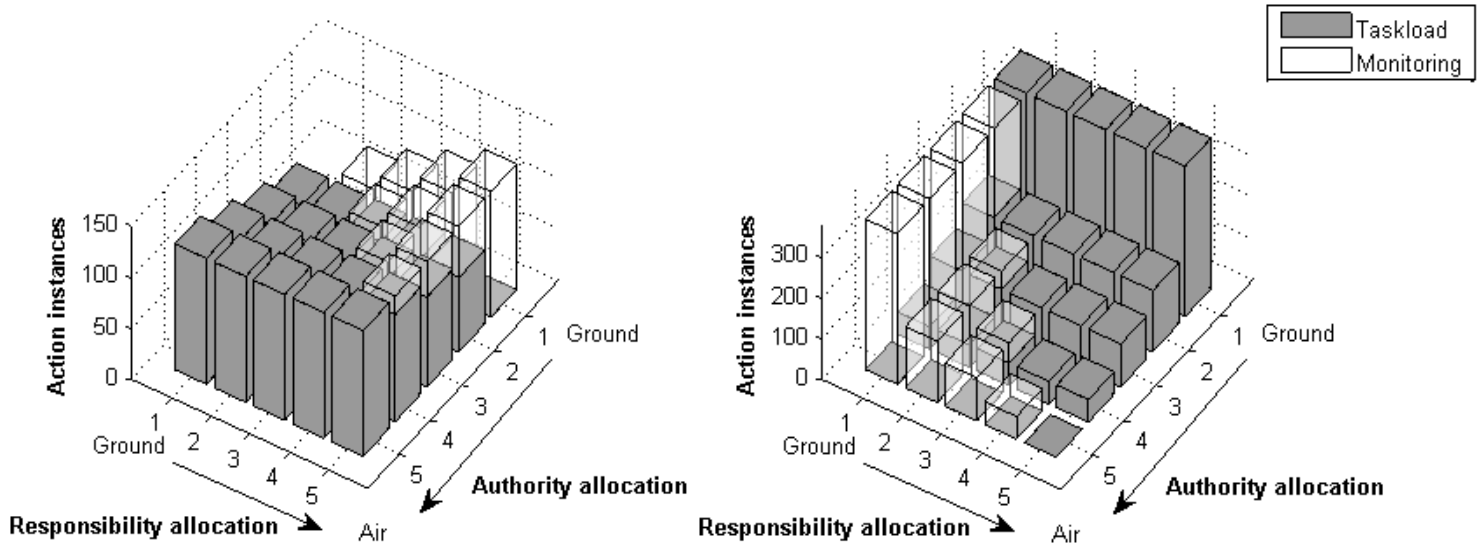


Figure 2. Total taskwork and monitoring actions for (left) averaged over the three flight crew agents and (right) the ATC agent.

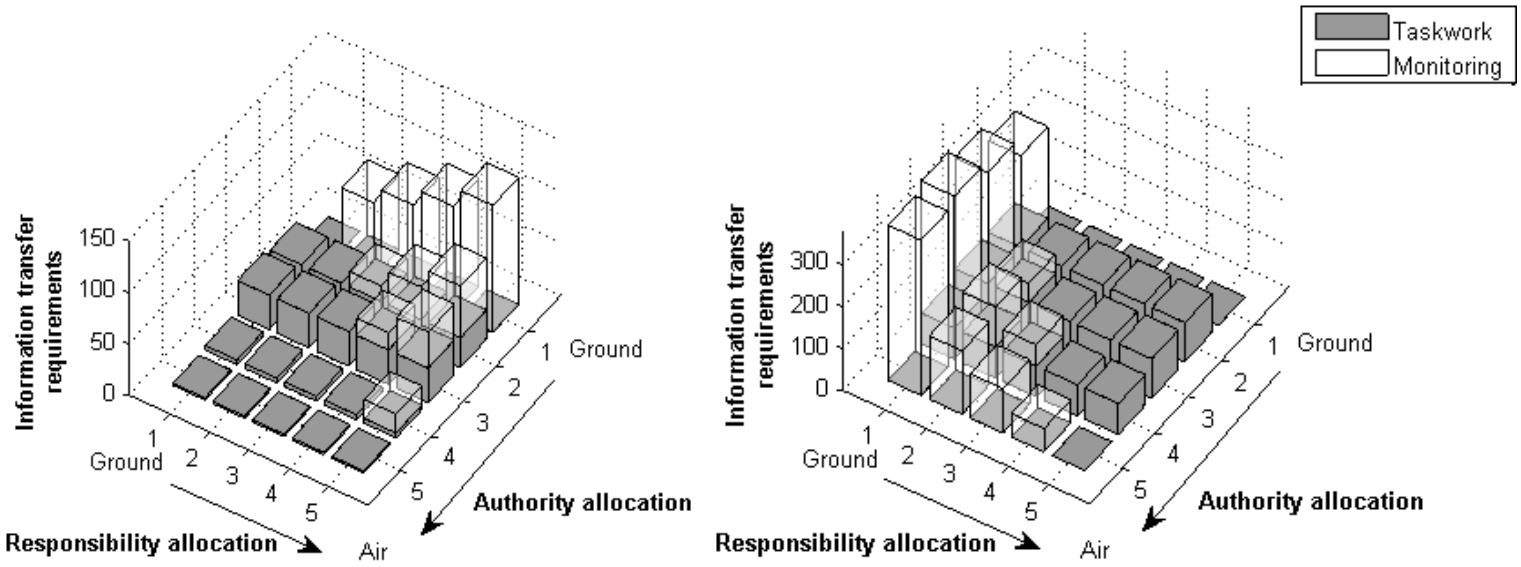


Figure 3. Information transfer requirements for (left) averaged over the three flight crew agents and (right) the ATC agent.

operator capability – we hope also that such issues can then be designed out of the concept of operation before it is entrenched through the implementation of automation and interfaces that are costly to re-design.

Once this fundamental assessment is performed, subsequent computational simulations can also examine human performance issues in greater detail. For example, different methods for performing each task can be examined, and the sensitivity of the operation to response time or variation in performance analyzed in detail.

Acknowledgements

This work is sponsored by the NASA Aviation Safety Program with Dr. Guillaume Brat serving as Technical Monitor under grant number NNX13AB71A S04. The authors also thank the other WMC developers for their ongoing mutual support.

References

Bhattacharyya, R.P. and Pritchett, A.R. (2014). A computational study of autonomy and authority in air traffic control, *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference*, Colorado Springs, CO.

Feigh, K.M. and Pritchett, A.R. (2014) Requirements for effective function allocation: A critical review. *Journal of Cognitive Engineering and Decision Making, 8*(1), 23-32. doi: 10.1177/1555343413490945

IJtsma, M., Bhattacharyya, R.P., Pritchett, A.R. and Hoekstra J. (Submitted). Computational assessment of different air-ground function allocations. Submitted for presentation at the *2015 FAA/Eurocontrol ATM R&D Seminar*.

Pritchett, A.R., Feigh, K.M., Kim, S.Y. and Kannan, S.K. (2014). Work models that compute to describe multiagent concepts of operation. *Journal of Aerospace Information Systems, 11*(10), 610-622. doi: 10.2514/1.I010146

Pritchett, A.R., Feigh, K.M., Mamessier, S. and Gelman, G (2014). Generic agent models for simulations of concepts of operation. *Journal of Aerospace Information Systems, 11*(10), 623-631. doi: 10.2514/1.I010147

Scallen, S.F. and Hancock, P.A. (2001). Implementing adaptive function allocation. *The International Journal of Aviation Psychology, 11*(2), 197-221. doi: 10.1207/S15327108IJAP1102_05

Wing, D.J. et al. (2010). Comparison of ground-based and airborne function allocation concepts for NextGen using human-in-the-loop simulations. *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference,* Fort Worth. Reston, VA: AIAA. doi: 10.2514/6.2010-9293

Woods, D.D. (1985). Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems. *AI Magazine*, 6(4), pp. 86-92.

INTEGRATING UAS OPERATIONS IN CLASS C AIRSPACE

Todd R. Truitt, Ph.D. & Randy L. Sollenberger, Ph.D.
Federal Aviation Administration
Pomona, NJ

The Federal Aviation Administration (FAA) is investigating Unmanned Aircraft Systems (UAS) operations in the National Airspace System because military, commercial, and civil users want to fly UAS for a broad range of purposes. Our research addresses the potential impact to Air Traffic Control Specialists (ATCS) due to UAS pilots' inability to comply with FAA regulations and air traffic control clearances and instructions that require direct visual means. UAS pilots cannot maintain visual separation from other aircraft, report aircraft *in sight*, or conduct visual approaches. The inability of UAS pilots to rely on visual means may affect ATCS workload, performance, and airspace efficiency. Twelve ATCS participated in teams of two in a high-fidelity, human-in-the-loop simulation. The participants controlled simulated traffic in two complex Class C airspace sectors under three conditions: Manned aircraft only, Low UAS activity, and High UAS activity. We collected measures of airspace efficiency, radio communications, and workload.

The safe and efficient integration of Unmanned Aircraft Systems (UAS) into the National Airspace System (NAS) is a primary goal for the Federal Aviation Administration (FAA) as well as UAS manufacturers and operators. UAS operations have increased in both the public and private sectors and the eventual goal is to enable UAS to fly routinely in the NAS as manned aircraft currently do. To achieve this goal, Air Traffic Control Specialists (ATCS) must be able to safely separate UAS from manned aircraft during all phases of flight. However, according to FAA Notice N8900.207, *Unmanned Aircraft Systems Operational Approval*, UAS are not compliant with sections of Title 14 of the Code of Federal Regulations (14 CFR) that pertain to aircraft (FAA, 2013). For instance, the see and avoid provisions of 14 CFR part 91, § 91.113b (FAA, 2014) cannot be satisfied by UAS operators due to the absence of an onboard pilot. For ATC operations requiring visual means of maintaining inflight separation, the lack of an onboard pilot does not permit ATCS to issue all standard clearances and instructions. Consequently, to ensure an equivalent level of safety, UAS operations require an alternative method of compliance or procedural risk mitigation to address the see and avoid limitations. In the future, a permanent and consistent method of visual compliance is needed for UAS operations in the NAS without the need for waivers or exemptions (FAA, 2013).

The research presented here addresses the potential impact on the NAS due to the inability of UAS pilots to comply with regulations and ATCS clearances and instructions that require the use of direct visual observation. Without the use of direct visual observation, UAS pilots cannot see and avoid other aircraft, maintain visual separation from other aircraft, or execute visual approaches. These limitations have the potential to increase ATCS workload and communications and decrease airspace efficiency. We conducted a series of experiments to examine the integration of UAS operations in complex Class C airspace that contained commercial and general aviation Instrument Flight Rules (IFR) controlled traffic and Visual Flight Rules (VFR) uncontrolled traffic (Truitt, Zingale, & Konkel, 2015). The experiment presented here examined UAS integration in a busy Terminal Radar Approach Control (TRACON) arrival stream to Oakland International Airport (OAK).

Method

We collected data from a total of six groups of two participants each for a total sample size of $N = 12$. Each group of participants spent five days in the laboratory. The experiment comprised a single factor (Condition – No UAS vs. Low UAS Integration vs. High UAS Integration) within-subjects repeated measures design. During the No UAS condition, the air traffic scenario contained only manned aircraft. During the Low UAS Integration condition, eight UAS operations were integrated with manned aircraft operations. In the High UAS Integration condition, thirteen UAS operations were integrated with manned aircraft operations. We counterbalanced the order of conditions and participant/sector combinations.

Participants

Twelve ATCS from Level 10-12 TRACON facilities served as participants. The participants were Certified Professional Controllers (CPC) from Boston, Charlotte, Dallas/Fort Worth, Houston, Philadelphia, Seattle, and

Minneapolis TRACONs. All of the participants were males between 26 and 55 years of age (*M* = 43.3, *SD* = 11.3, *Mdn* = 48.5). The participants had worked as ATCS from 6.3 years to 33.2 years (*M* = 20.7, *SD* = 10.2, *Mdn* = 24.0) and had worked as a CPC for the FAA from 5.9 years to 29.2 years (*M* = 19.5, *SD* = 9.0, *Mdn* = 23.5). The participants had controlled traffic in a TRACON facility for 5.9 years to 24.1 years (*M* = 13.9, *SD* = 6.6, *Mdn* = 12.0) and had controlled traffic for 12 months within the past year. The participants' experience using the Standard Terminal Automation Radar System (STARS) ranged from 0 years to 14 years (*M* = 6.9, *SD* = 4.9, *Mdn* = 7.7). None of the participants had previous experience with UAS operations.

Apparatus

Hardware. Each ATCS workstation included a Barco 2K x 2K Liquid Crystal Display (LCD), a STARS keyboard and trackball, and an emulated Terminal Voice Switching and Communication System (see Figure 1). Above each radar display was an emulation of an Information Display System (IDS) presented on a 21.3″ touchscreen. A Workload Assessment Keypad (WAK; Stein, 1985) was located at each workstation. Ceiling-mounted color video cameras were located above and behind each workstation. Simulation pilots and Air Traffic Control (ATC) Subject Matter Experts (SMEs) used workstations to affect simulated aircraft movements and communications. Each simulation pilot workstation included a computer, keyboard, mouse, display of aircraft information, and communications system. The SME workstations were similar to the participant workstations.

Figure 1. Air Traffic Control Specialist (ATCS) workstations in the Research Development and Human Factors Laboratory (RDHFL).

Software. We used the Distributed Environment for Simulation, Rapid Engineering, and Experimentation (DESIREE) to enable the STARS interface and functionality. We used the Target Generation Facility (TGF) to provide aircraft performance models, to generate aircraft tracks based on predefined flight plans, and to enable the simulation pilot workstations. Both DESIREE and TGF provided data collection capabilities.

Airspace. The airspace comprised sectors and surrounding airspace based on the Mulford and Grove sectors of Northern California TRACON (NCT). Modification of the airspace was necessary because we recruited participants from TRACON facilities across the NAS (with the exception of NCT) and the participants had to learn the airspace in about two days. SMEs simplified the airspace by consolidating the multiple sectors that surround the Mulford and Grove sectors into North and South sectors. The airspace modification reduced the number of sector handoff symbols and radio frequencies that participants had to memorize. SMEs also removed the complex altitude shelf structure of the sectors to further simplify operations. Figure 2 depicts the airspace.

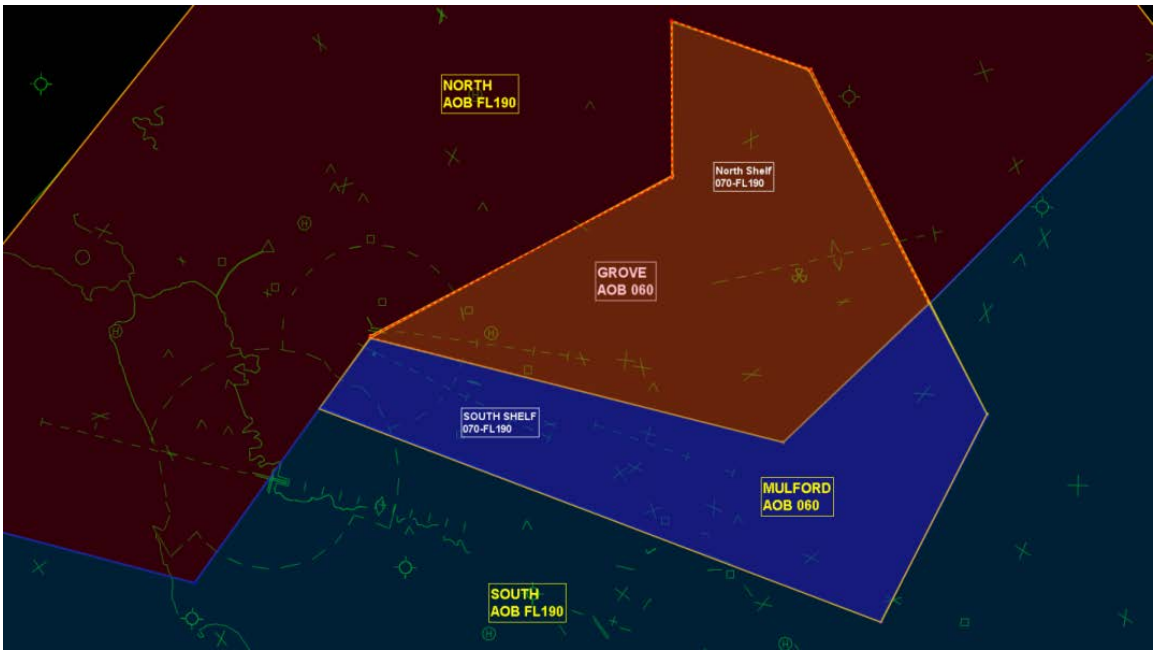


Figure 2. Mulford and Grove sectors with surrounding North and South sectors.

The participants controlled traffic in the Mulford and Grove sectors and managed arrivals into Oakland International Airport (OAK). We implemented a "West" configuration that required arrivals to use OAK runways 30, 28L, and 28R. We did not use runway 33/15. The Grove sector included airspace at or above (AOB) 6,000 ft Mean Sea Level (MSL) up to, but not including, Flight Level (FL) 190 (i.e., 19,000 ft MSL). The Grove sector was located above the final approach course to OAK runways 28L and 28R and was responsible for directing arrivals to those runways. The Mulford sector included airspace AOB 6,000 ft MSL up to, but not including, FL190. The Mulford sector was located above the final approach course to OAK runway 30, and the final approach course to Hayward Executive Airport (HWD) runway 28L, and was responsible for directing arrivals to those runways. The North and South sectors were "ghost" sectors. Each ghost sector was operated by an ATC SME. The North sector managed traffic AOB FL190 and between 7,000 ft MSL and FL190 over the Grove sector. The South sector managed traffic AOB FL190 and between 7,000 ft MSL and FL190 over the Mulford sector.

Air Traffic Scenarios. The No UAS (baseline) scenario comprised 91 total aircraft (56 arrivals, 25 departures, and 10 overflights). Twenty-two of the aircraft were uncontrolled VFR flights, and 69 were controlled IFR flights. There were 20 arrivals at OAK 30 and 13 arrivals and 5 departures at OAK 28R. Background traffic that impacted the participants' sectors were arrivals at San Francisco (SFO) 28R (10) and SFO 28L (13) as well as departures at SFO 28R (1), Buchanan Field (3), San Carlos (4), Reid-Hillview (2), Palo Alto (5), Livermore Municipal (1), Tracy Municipal (1), and San Jose (1) airports. The Low UAS Integration scenario was the same as the No UAS scenario, except 8 of the manned aircraft arriving at OAK 30 were replaced with UASs. The High UAS Integration scenario was the same as the No UAS scenario, except 13 of the manned aircraft arriving at OAK 30 were replaced with UASs. When UAS were present, they were evenly spaced throughout the scenario and were intermingled with other arrivals at OAK 30. All scenarios were 30 minutes in length. We created multiple versions of each scenario by changing only the aircraft callsigns to minimize the participants' ability to recognize traffic patterns within each scenario.

Results

We analyzed each data set using the appropriate repeated measures Analysis of Variance and calculated effect sizes using partial eta-squared (η_p^2). We analyzed significant main effects and interactions using Tukey's Honestly Significant Difference (HSD) test.

Aircraft Time and Distance in Sector

We used geographical sector boundaries to measure the total time and distance flown within each sector. We also counted the number of unique aircraft that flew through each sector. We measured the mean time (s) and distance flown (nm) by each unique aircraft in the Mulford and Grove sectors. An aircraft that flew into the Grove sector, then flew into the Mulford sector, and then flew back into the Grove sector was counted as a single unique operation for time and distance calculations per aircraft. There were no significant differences for any of the measures in the Grove sector.

The mean total time flown in the Mulford sector was significantly longer in the High UAS Integration condition (M = 15529 s, SD = 659 s) compared to the No UAS condition (M = 14675 s, SD = 638 s), $F(2, 22) = 7.45$, $p = .003$, $\eta_p^2 = .40$, $HSD(22) = 3069.48$. There was no significant difference in mean total time flown between the Low UAS Integration condition (M = 15003 s, SD = 791 s) and the other two conditions. The mean time flown per aircraft in the Mulford sector was significantly longer in the High UAS Integration condition (M = 345 s, SD = 15 s) compared to the No UAS condition (M = 327 s, SD = 16 s), $F(2, 22) = 6.70$, $p = .005$, $\eta_p^2 = .38$, $HSD(22) = 12.10$. There were no significant differences for mean time flown in the sector between the Low UAS Integration condition (M = 336 s, SD = 17 s) and the other two conditions.

The mean total distance flown in the Mulford sector was significantly greater in the High UAS Integration condition (M = 624 nm, SD = 30 nm) compared to the No UAS condition (M = 591 nm, SD = 27 nm), $F(2, 22) = 4.28$, $p = .027$, $\eta_p^2 = .28$, $HSD(22) = 29.90$. There was no significant difference in mean total distance flown between the Low UAS Integration condition (M = 597 nm, SD = 26 nm) and the other two conditions. The mean distance flown per aircraft in the Mulford sector was significantly greater in the High UAS Integration condition (M = 13.8 nm, SD = 0.7 nm) compared to the No UAS condition (M = 13.2 nm, SD = 0.7 nm), $F(2, 22) = 3.82$, $p = .038$, $\eta_p^2 = .26$, $HSD(22) = 0.63$. There were no significant differences in mean distance flown per aircraft between the Low UAS Integration condition (M = 13.4 nm, SD = 0.5 nm) and the other two conditions.

Communications

We recorded all voice communications to evaluate the number and duration of air-ground (pilot-to-Mulford/Grove) and ground-air (Mulford/Grove-to-pilot) Push-to-Talk (PTT) transmissions for the Mulford and Grove sectors. For the Grove sector, we found no statistically significant differences across conditions for either the number or duration of PTT transmissions, indicating that there were no differences in communication when UAS were in the Grove sector.

We measured the number and duration of the ground-air PTT transmissions from the Mulford controllers to the pilots and the air-ground PTT transmissions from the pilots to the Mulford controllers. The number of air-ground PTT transmissions differed significantly by condition, $F(2, 22) = 4.59$, $p = .022$, $\eta_p^2 = .29$. The post-hoc analysis indicated that there were more ground-air PTT transmissions at the Mulford sector in the High UAS Integration condition than in the No UAS condition, $HSD(22) = 13.13$. There was no difference in the number of ground-air PTT transmissions between the No UAS condition and the Low UAS Integration condition at the Mulford sector (see Figure 3).

Ground-air PTT transmission durations differed significantly between conditions in the Mulford sector, $F(2, 22) = 12.05$, $p = .03$, $\eta_p^2 = .52$. The post-hoc analysis indicated that ground-air PTT transmission durations in the Mulford sector were shorter in the High UAS Integration condition compared to the Low UAS Integration condition and the No UAS condition, $HSD(22) = 0.17$. There was no difference between the No UAS condition and the Low UAS Integration condition (see Figure 4).

The number of air-ground PTT transmissions in the Mulford sector did not differ significantly between conditions. However, the duration of air-ground PTT transmissions did differ significantly between conditions, $F(2, 22) = 3.95$, $p = .03$, $\eta_p^2 = .26$. The post-hoc analysis indicated that air-ground PTT transmission durations at the Mulford sector were shorter in the High UAS Integration condition compared to the No UAS condition, $HSD(22) = 0.18$ (see Figure 5). Therefore, both the controllers and the pilots made shorter transmissions when UAS were in the Mulford sector.

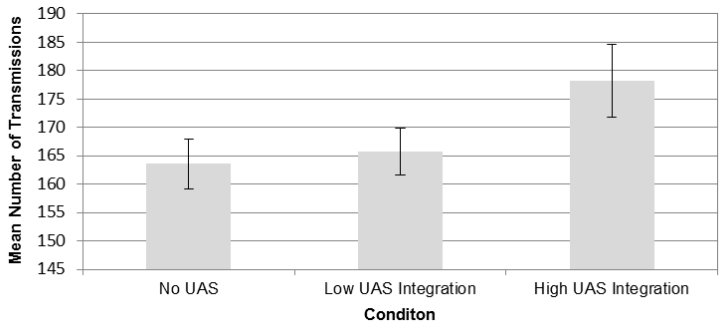


Figure 3. Mean number of ground-air PTT transmissions by Condition in the Mulford sector.

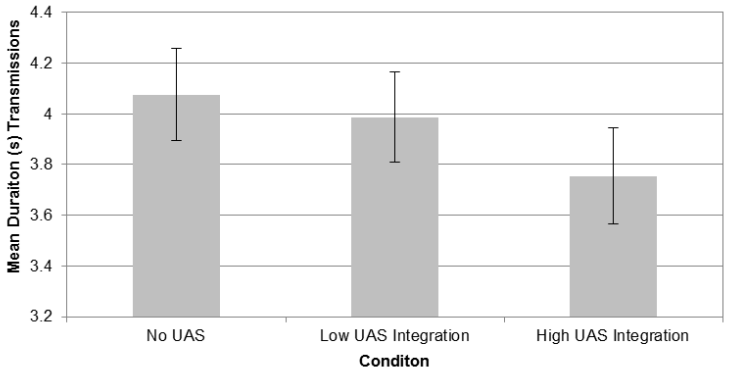


Figure 4. Mean duration of ground-air PTT transmissions by Condition in the Mulford sector.

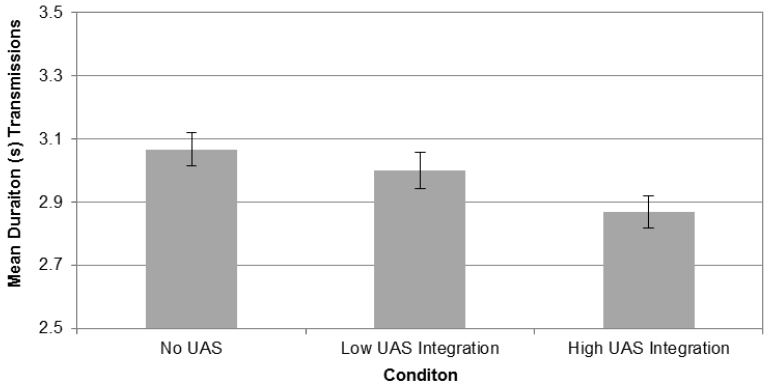


Figure 5. Mean duration of air-ground PTT transmissions by Condition in the Mulford sector.

Workload

Participants rated their subjective level of workload using the 10-button WAK (Stein 1985). If the participant did not respond within 20 seconds, the response was coded as "missing." We coded the failed responses as missing data because it is unknown if the participant was too busy to respond or simply did not notice the WAK prompt. We replaced missing responses (12/504 = 2.4%) with the mean WAK rating for the respective condition and time interval. The missing responses were randomly distributed across interval, condition, and sector.

There was a significant effect of Interval for WAK ratings at the Grove sector, $F(6, 66) = 14.77$, $p < .001$, $\eta_p^2 = .57$. Ratings increased from the first interval (4 min) to the second interval (8 min) and then increased again in

the fourth interval (16 min) before decreasing in the final interval (28 min), *HSD*(66) = 1.70. There was also a significant effect of Interval for WAK ratings at the Mulford sector, $F(6, 66) = 13.99$, $p < .001$, $\eta_p^2 = .56$. Ratings increased from the first interval (4 min) to the third interval (12 min), and then remained level until the final interval (28 min), *HSD*(66) = 1.56. The Interval effects were most likely due to the design of the air traffic scenarios, which included a "ramp up" of traffic in the beginning of each scenario.

There was a significant main effect of Condition for WAK ratings in the Mulford sector due to increased subjective workload as UAS were added to the scenarios, $F(2, 22) = 5.31$, $p = .013$, $\eta_p^2 = .33$ (see Figure 6). Although differences were relatively small, WAK ratings were significantly higher in the High UAS Integration condition ($M = 4.44$, $SD = 1.87$) compared to the No UAS condition ($M = 3.62$, $SD = 1.90$) and the Low UAS Integration condition ($M = 4.00$, $SD = 1.76$), *HSD*(22) = 1.68. There was no statistical difference between WAK ratings in the No UAS condition and the Low UAS Integration condition.

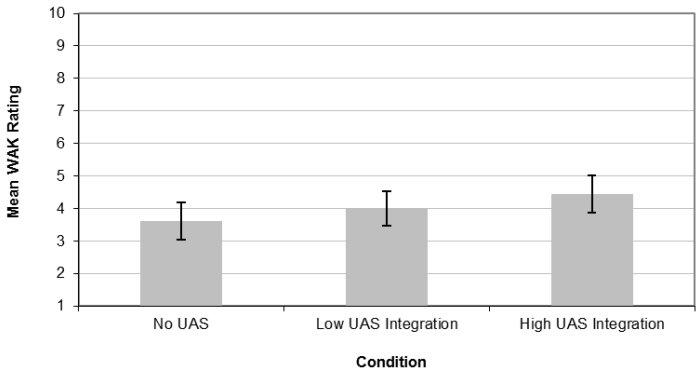


Figure 6. Mean WAK rating by Condition at the Mulford sector.

Conclusion

The Low UAS Integration condition had small but insignificant effects compared to the No UAS condition. The High UAS Integration condition had significant effects on efficiency, communications, and workload in the Mulford sector. In the High UAS Integration condition, there was an increase in the time and distance flown; there were a greater number of ground-air communications and shorter ground-air and air-ground communications; and there were higher ratings of subjective workload. Overall, a low volume of UAS operations in Class C airspace may be tenable and have relatively small effects on the airspace and ATCS.

References

Federal Aviation Administration. (2013). *Unmanned Aircraft Systems (UAS) operational approval* (FAA Notice N8900.207). Washington, DC: FAA.

Right-of-way rules; Except water operations, 14 C.F.R. § 91.113b (2014).

Stein, E. (1985). *Air traffic controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: FAA William J. Hughes Technical Center.

Truitt, T. R., Zingale, C., & Konkel, A. (2015). *A human-in-the-loop experiment assessing UAS integration in Class C airspace.* Atlantic City International Airport, NJ: FAA William J. Hughes Technical Center. Manuscript in preparation.

UAS AIR TRAFFIC CONTROLLER ACCEPTABILITY STUDY-2: EFFECTS OF COMMUNICATIONS DELAYS AND WINDS IN SIMULATION

James R. Comstock, Jr., Rania W. Ghatas, Maria C. Consiglio, James P. Chamberlain
NASA Langley Research Center
Hampton, Virginia

Keith D. Hoffler
Adaptive Aerospace Group
Hampton, Virginia

This study evaluated the effects of Communications Delays and Winds on Air Traffic Controller ratings of acceptability of horizontal miss distances (HMDs) for encounters between UAS and manned aircraft in a simulation of the Dallas-Ft. Worth East-side airspace. Fourteen encounters per hour were staged in the presence of moderate background traffic. Seven recently retired controllers with experience at DFW served as subjects. Guidance provided to the UAS pilots for maintaining a given HMD was provided by information from self-separation algorithms displayed on the Multi-Aircraft Simulation System. Winds tested did not affect the acceptability ratings. Communications delays tested included 0, 400, 1200, and 1800 msec. For longer communications delays, there were changes in strategy and communications flow that were observed and reported by the controllers. The aim of this work is to provide useful information for guiding future rules and regulations applicable to flying UAS in the NAS.

One of the major barriers to integrating UAS in the National Airspace System (NAS) is the requirement to see and avoid other aircraft per CFR 14, Parts 91.111 and 91.113 and other applicable regulations and accepted practices. In today's operations pilots are required to follow right of way rules and remain well clear of other aircraft. There is also an obvious collision avoidance requirement. In an Air Traffic Services (ATS) environment, pilots are expected to comply with these see and avoid requirements while also complying with Air Traffic Control (ATC) instructions and clearances or to negotiate changes to these instructions and/or clearances as necessary. See-and-avoid capable pilots are generally expected to maneuver and communicate in predictable ways and in a manner that preserves the safety, orderliness, and efficiency of the ATS environment. UAS will likely be expected to operate in a similar manner, but with Detect and Avoid (DAA) replacing the see-and-avoid capability of a manned aircraft. The acceptable design space and capabilities for DAA systems in this environment are largely undefined. This controller-in-the-loop simulation experiment sought to illuminate the DAA design space for UAS operating in an ATS environment.

Detect and Avoid implementations must be designed in a way that minimizes issuance of corrective Resolution Advisories (RAs) by TCAS (Traffic Collision Avoidance System) equipped intruders. RAs are alerts with recommended vertical escape maneuvers, to maintain or increase vertical separation with intruders that are collision threats. Corrective RAs that cause evasive maneuvers can be disruptive to the air traffic system and are a last resort maneuver when all other means of separation have failed. The DAA concept evaluated in this experiment was designed to detect encounter geometries that will cause an RA, and provide guidance for action that may be taken early enough to avoid an RA.

This study is the second in the Controller Acceptability Study (CAS) experiment series and is based largely on CAS-1 experiment design, scenarios, and results. The primary goals of this study were to address the impact of communication delays and wind conditions on the execution of Ground Control Station self-separation tasks and how the resulting maneuvers are rated by Air Traffic Controllers. The communications delays used in this study include four different ATC-pilot communication latencies or delays that might be expected in operations of UAS

controlled by combinations of ground or satellite command and control links. These include 0, 400, 1200, and 1800 msec one-way communications delays.

One of the goals of the earlier CAS-1 study was to establish a generally acceptable Horizontal Miss Distance when there were encounters between DAA equipped UAS and transponder equipped manned General Aviation aircraft that were not communicating with ATC. The results indicated that horizontal miss distances (HMDs) of 1.0 and 1.5 nautical miles (nmi) appeared to be optimal for ATC acceptability, when the traffic encounters are away from the airport vicinity. In that study HMDs of 0.5, 1.0, 1.5, 2.0, 2.5 and 3.0 nmi were evaluated for encounters that were Opposite Direction (Head-on), Overtakes (same direction with UAS faster), and Crossings.

Objectives

The overall focus of this experiment (CAS-2) was on determining the effect of simulated DAA equipped UAS on Air Traffic Controller workload and acceptability of maneuvers with differing spacing parameters used in the DAA algorithms and with Winds and Communications delays. Based on the results of CAS-1, the set of Horizontal Miss Distances (HMD) for crossing traffic encounters was reduced to include 0.5, 1.0, and 1.5 nmi. An important difference, however, was that in CAS-1, crossing geometry HMDs of 1.5 nmi or less were designed to require no maneuver by the UAS to maintain the desired HMD. In this study, there were instances of crossing geometries of both 1.0 and 1.5 nmi that required maneuvers and concomitant communications with ATC. All opposite direction (head-on) encounters and overtaking encounters required communications with ATC and maneuvering.

Research questions

A. Given wind and communications delay conditions, were DAA self-separation (SS) maneuvers too small/too late, resulting in issuance of traffic safety alerts or controller perceptions of unsafe conditions? Tested by traffic encounters with smaller HMDs requiring maneuvers.
B. Given wind and communications delay conditions, were DAA SS maneuvers too large (excessive “well clear” distances), resulting in behavior the controller would not expect and/or disruptions to traffic flow? Tested by traffic encounters with larger HMDs.
C. Given wind and communications delay conditions, were there acceptable, in terms of ATC ratings, workload, and closest point of approach data, DAA miss distances that can be applied to the development of future DAA algorithms?
D. Do communications delays for the UAS in the airspace result in an impact on the Air Traffic Controllers communications flow? Are the delays disruptive in terms of transmissions being “stepped-on” (simultaneous transmissions by several aircraft), and/or are additional repeats of information required with delays.

Methodology

Subjects

Seven recently retired Air Traffic Controllers with experience at the Dallas-Ft. Worth (DFW) East-side facility performed traffic separation tasks for the scenarios developed. Most of the Controllers were currently instructors in the training center at DFW. Each of the controllers performed ATC tasks in the simulated DFW East side environment over two days of testing. There were 14 UAS traffic encounters each hour for six test hours and these UAS were controlled by two pseudo-pilots each having access to Ground Control Station displays showing the self-separation guidance information in real-time. Background traffic, to maintain the environment and workload close to that of actual DFW traffic, was controlled by pseudo-pilots at two additional pilot stations. Controllers who participated in CAS-1, about four months earlier, were eligible to serve in CAS-2.

Independent Variables

To get at the Research Questions noted above, the first independent variable of interest was the HMD. Related to the first variable is the encounter geometry between the aircraft in the encounter situation and the speed differentials between the encountering aircraft. Additional variables of interest include two levels of wind (calm and moderate) and four levels of communications delay. The parameters of these variables are shown in Table 1.

Table 1.
Parameters of Research Variables

- Horizontal Miss Distances (HMD), 3 values: 0.5, 1.0, 1.5 nautical miles
- Wind Conditions, 2 values: Calm (~7 knots) and Moderate (~22 knots)
- Communications Delay, 4 values: 0, 400, 1200, and 1800 msec (one-way times)
- Encounter Geometry, 3 cases: Opposite-direction, Overtake, Crossing
    - Intruder Opposite-direction at 180 degrees +/- 15 degrees (Non-crossing)
    - Intruder at 90 degrees +/- 15 degrees (Crossing)
    - Intruder ahead at 0 degrees +/- 15 degrees (Overtaking, Non-crossing)
    - All geometries without vertical separation (but may include climbing/descending trajectories)
    - UAS requests passes to right of intruder for non-crossing geometries
    - UAS passes in front of intruder for crossing geometries
    - Intruder Speed Differential (5 values for Crossings: 0, +/- 40, +/- 80 knots)
- 42 test conditions: 6 Opposite-direction, 6 Overtake, 30 Crossing
- 14 encounters per hour, 6 hours of testing over two days, 84 total encounters
- Background (non-encounter) traffic communicating with ATC: Approximately 40 per hour

Scenarios

The airspace modeled for this experiment is a portion of airspace delegated to Dallas-Ft. Worth TRACON (D10). Specifically, Sector DN/AR-7 South Flow. The majority of UAS traffic arrived or departed the Collin County Airport (KTKI). The scenarios were designed and situated in this airspace so as to enable various encounter geometries between the UAS and intruder aircraft while manned aircraft traffic was handled in order to achieve realistic levels of workload for the Controllers. A chart of the area is shown in Figure 1.

Communications, Navigation, and Surveillance Assumptions

The experiment assumed Communication, Navigation and Surveillance (CNS) architectures and capabilities appropriate for current-day operations in the applicable airspace classes and that these capabilities were available to all aircraft (manned and unmanned) in the simulation environment. The intruders were not communicating with ATC. UAS command, control, and communication capability was assumed available between Unmanned Aircraft (UA) and their respective GCS. The UA was assumed to be capable of receiving/transmitting voice communications to and from ATC facilities and proximate “party-line” aircraft via VHF frequencies in the same manner as manned aircraft in the same airspace, and of relaying these voice communications to/from the GCS pilot via one or more UA-GCS links. It was further assumed that, in addition to the relayed voice communications, the UA-GCS link(s) carried all command/control data between the UA and GCS. This study assumed large UAS. The UAS GCS pilots were confederate participants (not subjects). It was assumed that surveillance sensors applicable to support SAA were available and functioned without failures.

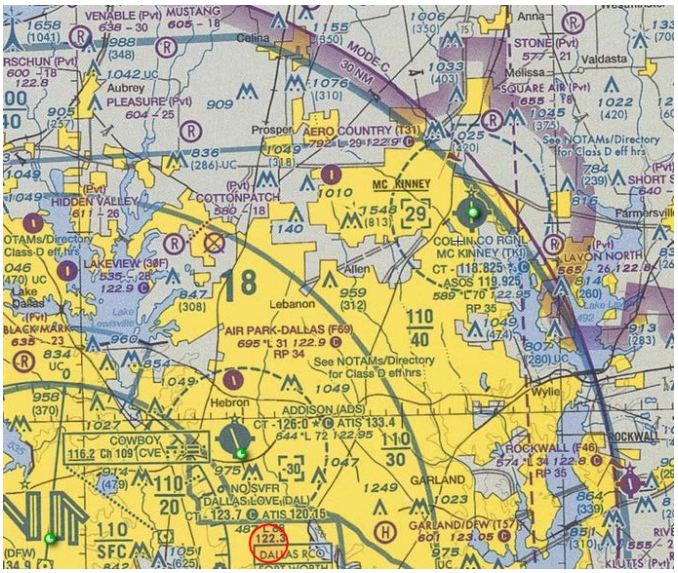


Figure 1. Chart showing Collin County Airport (McKenny, KTKI), upper right; DFW is in the lower left.

Facilities, Software, and Hardware

The study was run in a dedicated facility housed at Stinger Ghaffarian Technologies (SGT), near the NASA Langley Research Center. The displays for the UAS and manned aircraft control stations and the ATC displays were driven by modified versions of the MACS (Multi Aircraft Control System) software (Prevot, 2002). Modifications included incorporation of Stratway+ algorithms to drive Navigation display "bands" which indicated a range of headings that would result in a loss of well clear with one or more traffic aircraft. Information on the self-separation algorithms may be found in Hagen, Butler, and Maddalon, 2011, and Muñoz, Narkawicz, Chamberlain, Consiglio, and Upchurch, 2014. The hardware, software, and operations implementation team included personnel from SGT, Adaptive Aerospace Group (AAG), and Intelligent Automation Inc. (IAI).

Dependent Variables

Horizontal Miss Distance Ratings. After each traffic encounter, an ATC subject matter expert seated next to the Controller subject asked: "How was the spacing of that last encounter?" or "How Acceptable was the miss distance in the previous encounter?" Subjects had a copy of the information in Table 2 available to them during the test sessions. They were briefed that fractional responses, such as 1.5 or 3.5, were completely acceptable. If time permitted, an explanation for the rating was asked and noted.

Table 2. *Rating scale used for encounter assessment. (Fractional values, e.g., 1.5, were acceptable)*

1	Much too close; unsafe or potentially so; cause or potential cause for issuance of a traffic alert
2	Somewhat close, some cause for concern
3	Neither unsafely close nor disruptively large, did not perceive the encounter to be an issue
4	Somewhat wide, a bit unexpected; might be disruptive or potentially disruptive in congested airspace and/or with high workload
5	Excessively wide, unexpected; disruptive or potentially disruptive in congested airspace and/or with high workload

Workload assessment. About every five minutes during each hour long test session a workload rating was requested. This was done similar to the ATWIT (Air Traffic Workload Input Technique) method of Workload assessment (Stein, 1985). A scale with numbers from 1 to 6 was presented at the top of the ATC display and the subject clicked on one of the numbers when an aurally presented (through headphones) "Ding" occurred and the rating scale turned yellow. ATC Test subjects were briefed on definitions of the 1 to 6 scale during training and also had the scale definitions available during the test sessions. For this study the scale definitions were: 1 - Minimal mental effort required; 2 - Low mental effort required; 3 - Moderate mental effort required; 4 - High mental effort required; 5 - Maximal mental effort required; and 6 - Intense mental effort required.

System Performance Metrics. Data concerning the encounter aircraft were recorded and included Aircraft-to-Aircraft separation distances and time to the closest point of approach (CPA). For the communications time delay conditions, the communications system that permitted incorporating delays also recorded the push to talk status of all parties communicating so that "step-ons" (two stations transmitting at the same time) could be recorded.

Post-run questionnaires. After each one-hour test session a questionnaire was administered to record ratings and comments on the preceding test session. Specific topics addressed included: 1 – Effects of communications delay; 2 – Realism of traffic density; 3 – Realism of workload; and 4 – Realism of communications rate.

Results

Horizontal Miss Distances. Figure 2 shows the mean ratings by the Controllers for each of the Horizontal Miss Distances (HMDs) tested for the crossing traffic encounters. The Geometric CPA (Closest Point of Approach) is how close the two aircraft would pass if no maneuver was made. If HMD was equal to Geometric CPA, no maneuver would be called for by the self-separation algorithms, and no communications with ATC to request a maneuver was required. To see if the Controller's rating was affected by whether the UAS had to contact ATC to request a maneuver to maintain the HMD, the encounter geometry was also set up such that the HMD was greater than the Geometric CPA for the 1.0 and 1.5 nmi HMDs. As can be seen from Figure 2, the Controllers ratings of HMD were not affected by whether communications and a maneuver were required by the UAS.

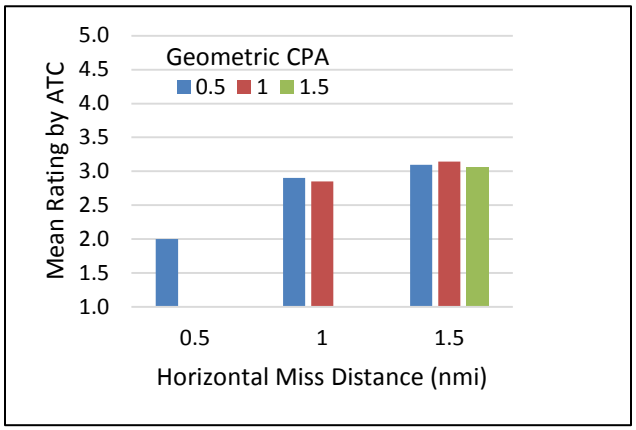


Figure 2. Mean Ratings by encounter distance (Crossings). Rating definitions are in Table 2.

Figure 3 shows the Controller rating data for crossing encounters and shows the highest percentages for a rating of 3 (*Neither unsafely close nor disruptively large, did not perceive the encounter to be an issue*), at the 1.0 and 1.5 nmi HMDs. Ratings shifted for the 0.5 nmi HMD indicating greater concern for that miss distance. Figure 4 shows similar rating data for the Overtake and Opposite Direction encounters, all of which required maneuvers, and communications with ATC. The rating scale used is shown in Table 2.

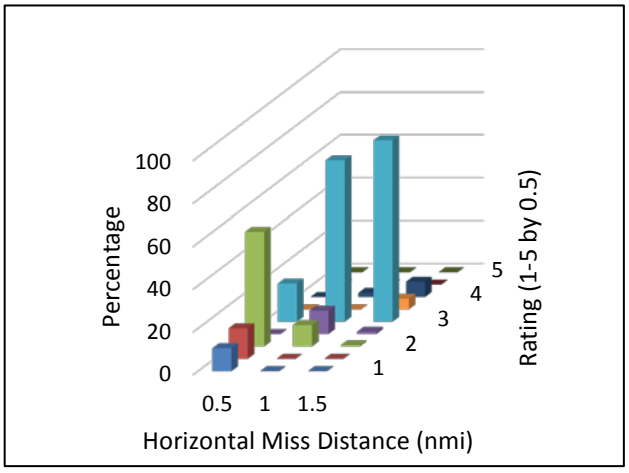


Figure 3. Ratings by HMD (Crossings)

Realism of Traffic Density and Workload. Care was taken in the design of the research scenarios to have traffic densities like those found in the real world. In response to the end of each hour question "*Rate the realism of the Traffic Density of the simulation during the preceding hour*," 66.7% of responses were that "*Traffic Density was about the same as would be found in real world operations;*" and 31.0% of the responses were that "*Traffic Density was somewhat lower than real world operations.*" Workload ratings, based on data collected at 5-minute intervals, showed the following distribution of responses: 32.3% "*Minimal mental effort required;*" 42.9% "*Low mental effort required;*" 18.2% "*Moderate mental effort required;*" and 0.9% "*High mental effort required.*" Workload ratings did not differ across the two wind levels or four communications delay conditions.

Communications Delays and Wind. Communications delays of 0, 400, 1200, and 1800 msec (one-way times) were used for communications with the 14 UAS per hour that had traffic encounters. Manned aircraft in the scenario had no added delays. While no differences in ratings of HMD or workload were noted, selected Controller comments reflect the difficulties long delays introduce: *"The communications delays did cause some a/c to 'step-on each other.' This required extra transmissions to other traffic because they were blocked;" "The delay resulted in many repeats and was irritating;" "Repeats have a major impact on workload of ATC. In a busy environment you can't stand for a lot of them;" "Numerous repeats and step-ons! When in busy environments your transmissions need to flow and repeats/blocks only put you behind."* Also observed was a change in strategy by some controllers in the long delay scenarios to work manned, quicker responding, traffic first then go to the UAS with their delayed responses. The "low" and "moderate" wind levels did not create any issues for the controllers. For the UAS pilots the separation algorithms handled the wind conditions with no problems.

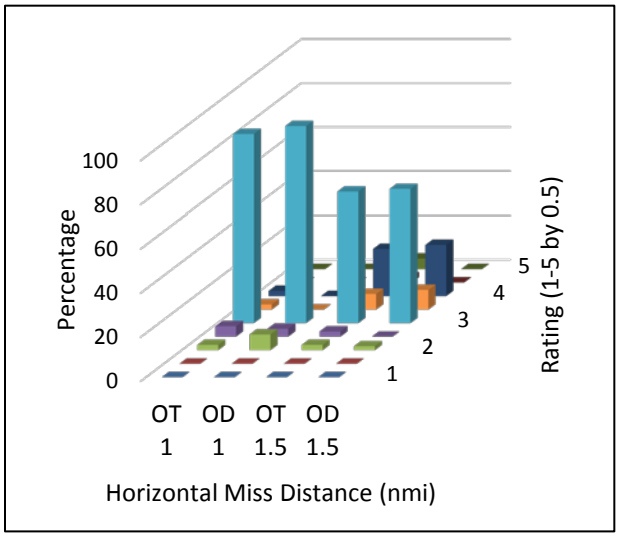


Figure 4. Ratings by HMD (Overtake – OT and Opposite Direction - OD)

Discussion

The present study employed a simulation of the Dallas-Ft. Worth East-side airspace with UAS operating in and out of Collin County airport Northeast of DFW. The results confirm the Controller acceptability of 1.0 and 1.5 nmi HMDs found in the CAS-1 study, even when maneuvers are required to maintain those miss distances, and winds are part of the scenarios. The 7 and 22 knot wind conditions tested were handled by the self-separation algorithms without issues, and presented no issues for the controllers. Long voice communications delays between the UAS and ATC are identified as a problem in a high traffic-density environment such as this.

Since the present study assumed perfect surveillance, future studies should incorporate sensor uncertainty and sensor effective range as variables of interest. Also of interest are simulation of failure modes, and especially from the ATC perspective, the maneuvers that a UAS would perform in a high traffic density environment if the communications link is lost. The aim of this work is to provide useful information for guiding future rules and regulations applicable to flying UAS in the NAS.

References

Hagen, G. E., Butler, R. W., and Maddalon, J. M. (2011). Stratway: A Modular Approach to Strategic Conflict Resolution. *Proceedings of the 11th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, September 20-22, 2011, Virginia Beach, Virginia.

Muñoz, C., Narkawicz, A., Chamberlain, J., Consiglio, M., and Upchurch, J. (2014). A Family of Well-Clear Boundary Models for the Integration of UAS in the NAS. *Proceedings of the 14th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, AIAA-2014-2412, Atlanta, Georgia.

Prevot, T. (2002). Exploring the Many Perspectives of Distributed Air Traffic Management: The Multi Aircraft Control System MACS. *AAAI HCI-02 Proceedings*, 149-154.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe.* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: Federal Aviation Administration.

UAS IN THE NAS AIR TRAFFIC CONTROLLER ACCEPTABILITY STUDY-1: THE EFFECTS OF HORIZONTAL MISS DISTANCES ON SIMULATED UAS AND MANNED AIRCRAFT ENCOUNTERS

Rania W. Ghatas, James R. Comstock, Jr., Maria C. Consiglio, James P. Chamberlain
NASA Langley Research Center
Hampton, Virginia

Keith D. Hoffler
Adaptive Aerospace Group
Hampton, Virginia

This study examined air traffic controller acceptability ratings based on the effects of differing horizontal miss distances (HMDs) for encounters between UAS and manned aircraft. In a simulation of the Dallas/Fort Worth (DFW) East-side airspace, the CAS-1 experiment at NASA Langley Research Center enlisted fourteen recently retired DFW air traffic controllers to rate well-clear volumes based on differing HMDs that ranged from 0.5 NM to 3.0 NM. The controllers were tasked with rating these HMDs from "too small" to "too excessive" on a defined, 1-5, scale and whether these distances caused any disruptions to the controller and/or to the surrounding traffic flow. Results of the study indicated a clear favoring towards a particular HMD range. Controller workload was also measured. Data from this experiment and subsequent experiments will play a crucial role in the FAA's establishment of rules, regulations, and procedures to safely and efficiently integrate UAS into the NAS.

Unmanned Aircraft Systems (UAS) are no longer technological systems of the unforeseeable distant future, but rather of the present and near future. They are systems that are evolving quickly and will soon become commonplace in the National Airspace System (NAS). According to the Federal Aviation Administration (FAA) Modernization and Reform Act of 2012 (2012), the United States Congress mandated the FAA to open the NAS to civil UAS "as soon as practicable, but not later than September 30, 2015." However, opening the NAS to civil UAS is a challenging task, a task that encompasses multiple safety issues of which include detect and avoid (DAA) implementations, self-separation (SS) procedures, and collision avoidance (CA) technologies to remain well-clear of other aircraft. Routine access to the NAS will require UAS to have new equipage, standards, rules and regulations, and procedures, among others, in addition to a slew of supporting research efforts. As a result, the National Aeronautics and Space Administration (NASA) has established a multi-center "UAS in the NAS" project, in collaboration with the FAA and industry, to examine essential safety concerns regarding the integration of UAS in the NAS. Among NASA's guiding research efforts is NASA Langley Research Center's (LaRC) air traffic Controller Acceptability Study (CAS) human-in-the-loop (HITL) experiment series. The first CAS experiment (CAS-1) researched a subset of safety features to examine well-clear volumes by simulating differing horizontal miss distances (HMDs) at the Dallas/Fort Worth (DFW) East-side airspace.

The concepts of remaining well-clear and DAA come from current standards under which pilots currently operate within the NAS. According to Title 14, Part 91, Section 91.111 (a), of the Code of Federal Regulations (14CFR 91.111 (a)), "no person may operate an aircraft so close to another aircraft as to create a collision hazard," and 14CFR 91.113 (b), under right-of-way rules, states "*General*. When weather conditions permit, regardless of whether an operation is conducted under instrument flight rules or visual flight rules, vigilance shall be maintained by each person operating an aircraft so as to see and avoid other aircraft. When a rule of this section gives another aircraft the right-of-way, the pilot shall give way to that aircraft and may not pass over, under, or ahead of it unless well clear." In essence, these standards, among others, require pilots to follow right-of-way rules and remain well-clear, by seeing and avoiding, other aircraft. In an Air Traffic Services (ATS) environment, pilots are expected to comply with those requirements while also complying with Air Traffic Control (ATC) instructions and clearances, or to negotiate changes, as necessary, to those instructions and clearances. Pilots capable of seeing and avoiding other aircraft are mostly expected to maneuver and communicate in predictable ways; ways that preserve the safety, orderliness, and efficiency of the ATS environment. Inherently, UAS pilots will be expected to operate in a similar manner. As such, in October of 2009, the term sense and avoid (SAA), used interchangeably with DAA and comparable to manned aircraft see-and-avoid requirements, was defined as "the combination of UAS Self-Separation (SS) plus Collision Avoidance (CA) as a means of compliance with 14CFR Part 91, §91.111 and §91.113" and published by the FAA-sponsored SAA for UAS Workshop Final Report. The SAA for UAS Workshop Final

Report goes on to define SS and CA as a means to remain well-clear and as a means to avoid Near Mid-Air Collisions (NMACs), respectively. Under Section 6, 7-6-3 (b), of the Aeronautical Information Manual (AIM), the FAA defines NMACs as "an incident associated with the operation of an aircraft in which a possibility of collision occurs as a result of proximity of less than 500 feet to another aircraft..." Figure 1 shows the different volumes and boundaries associated with remaining well-clear. In order to remain well-clear, the Self-Separation Volume (SSV) size should be large enough to avoid corrective Resolution Advisories (RAs) for Traffic Collision Avoidance System (TCAS)-equipped intruders; safety concerns for controllers; and, undue concern for proximate see-and-avoid pilots. Determination of minimum and maximum operationally acceptable SSV sizes will inform the design space for required DAA surveillance accuracy. Current standard NAS operations are the building blocks for which future UAS NAS operations will advance.

Figure 1. NASA's Separation Assurance/Sense-and-Avoid Interoperability (SSI) SSV represents a concept volume of remaining well-clear. *Note*. CAT, SSV and SST boundaries are notional and generally not cylindrical.

Controller Acceptability Study-1 Objectives

The primary focus of the CAS-1 experiment was on determining the effects of self-separation maneuvering tasks, as performed by pilots in a Ground Control Station (GCS) using simulated DAA-equipped UAS, on ATC workload and how the resulting maneuvers impacted ATC acceptability of the differing spacing parameters, also known as HMDs, which were implemented in the DAA algorithms.

The aim of CAS-1 was to address, through data collection and analysis, the following research questions: A) Are DAA SS maneuvers too small/too late, resulting in issuance of traffic safety alerts or air traffic controller perceptions of unsafe conditions?; B) Are DAA SS maneuvers too large (excessive "well clear" distances), resulting in behavior the air traffic controller would not expect and/or disruptions to traffic flow?; and, C) Are there acceptable, in terms of ATC ratings, workload, and closest point of approach data, DAA miss distances that can be applied to the development of DAA algorithms?

In order to address the above research questions, an appropriate experiment design was necessary to achieve the goal of the experiment's primary focus and aim.

Method

Subjects

To keep in line with designing an appropriate experiment, ATC subjects who had real-world experience controlling the East-side area of DFW were sought after, and, as such, fourteen recently retired DFW controllers were utilized for this experiment. ATC experience among subjects ranged between 25.5 years to 33 years with an average of approximately 30.4 years. Subjects also had an average of approximately 20.4 years of DFW experience in a Terminal Radar Approach Control Facility (TRACON). Additionally, of that DFW experience, an average of 18.3 years' worth of experience was in the East-side sector of the DFW TRACON (D10) region. Furthermore, out of the fourteen subjects, none had experience with UAS operations, which allowed for a fresh perspective to controlling UAS traffic encounters, and four of the fourteen controllers were active instructors at the DFW training center. Also, in order to maintain and simulate a close to real-world DFW environment and workload, two pseudo-pilots controlled each UAS GCS and two additional pseudo-pilots controlled background traffic. ATC positions, other than that of the subject controller, were 'controlled' via personnel acting as other DFW TRACON sector controllers. The subject controller was expected to communicate with these other sectors as he normally would in the field, with the exception of some Standard Terminal Automation Replacement System (STARS) functions;

STARS "provides controllers with critical operational information about aircraft positions, flight data, and weather" (FAA, 2012).

Independent Variables

With the aim of acquiring data on ATC acceptability ratings on differing spacing parameters, the primary Independent Variable (IV) of interest was determining the minimum acceptable HMD as a result of a given parameter in the DAA algorithm. The secondary IV of interest was the encounter geometry between the aircraft in the encounter situation.

Horizontal miss distances. CAS-1 researched six different HMD values that included the following spacing parameters measured in nautical miles (NM): 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0. These values were implemented in the DAA algorithm.

Encounter geometry. Three encounter geometries were utilized in CAS-1, which included opposite-direction, overtake, and crossing. Figure 2 visually portrays the different encounter geometries. The following parameters frame the secondary IV:

- Intruder opposite-direction at 180 degrees +/- 15 degrees (non-crossing)
- Intruder to right at 90 degrees +/- 15 degrees (crossing)
- Intruder ahead at 0 degrees +/- 15 degrees (overtaking, non-crossing)
- All geometries without vertical separation (but may include climbing/descending trajectories)
- UAS pilots were instructed to pass to the right of intruder for non-crossing geometries
- UAS pilots were instructed to pass in front of intruder for crossing geometries
- Intruder Speed Differential (5 speed values for crossing: 0, + 40, - 40, + 80, and - 80 knots)
- 42 test conditions: 6 opposite-direction, 6 overtake, 30 crossing
- 14 encounters per hour; 6 one-hour test sessions per subject enabled a replicate for each encounter

The parameters of the primary and secondary IVs are shown in Table 1.

Table 1.
Parameters of the primary and secondary independent variables.

	Horizontal Miss Distances in Separation Algorithm					
Encounter Geometry	*0.5*	*1.0*	*1.5*	*2.0*	*2.5*	*3.0*
Opposite-direction	1 speed	1 speed	1 speed	1 speed	1 speed	1 speed
Overtake	1 speed	1 speed	1 speed	1 speed	1 speed	1 speed
Crossing	5 speeds	5 speeds	5 speeds	5 speeds	5 speeds	5 speeds

Scenarios

The scenarios implemented in CAS-1 simulated ATC Sector DN/AR-7 South Flow, which is a portion of airspace delegated to DFW TRACON (D10). The scenarios were designed and situated in the selected airspace so as to enable various encounter geometries between the UA and intruder aircraft.

Figure 2. Encounter geometries used in CAS-1 included, from left to right, opposite-direction, overtake, and crossing encounters.

Dependent Variables

System Performance Metrics. Aircraft-to-Aircraft separation distances, operational errors and deviations, delays to aircraft in scenario, re-sequencing arrival aircraft, and voice communication errors, which included transposing information, call sign errors, repeats, and "say again" were recorded during each one-hour data collection run.

Human Operator Performance Metrics. Three different human operator performance metrics were examined. Among those three was the assessment of controller workload through the use of the Air Traffic Workload Input Technique (ATWIT) methodology. ATWIT was the tool used to measure mental workload in "real-time" by presenting auditory and visual cues that prompted the controller to press one of six ratings at fixed time intervals to indicate the amount of mental workload experienced at that moment (Stein, 1985). The response scale was built into the controller display software and had ratings from 1 to 6. A rating of 1 suggested "*minimal mental effort required*;" a rating of 2 suggested "*low mental effort required*;" a rating of 3 suggested "*moderate mental effort required*;" a rating of 4 suggested "*high mental effort required*;" a rating of 5 suggested "*maximal mental effort required*;" and, a rating of 6 suggested "*intense mental effort required*." In addition, another performance metric collected involved post encounter verbal queries that were gathered to evaluate controller acceptability of HMD spacing parameters. Controllers were asked to rate HMDs based on a scale from 1-5. Table 2 shows the scaled used and defines each of the acceptability ratings. Lastly, an "end-of-hour questionnaire" was administered to each subject controller at the conclusion of each one-hour data collection session.

Facilities, Software, and Hardware

The experiment was conducted in a dedicated facility located at Stinger Ghaffarian Technologies (SGT), near NASA LaRC in Hampton, Virginia. The facility ran a UAS modified version of the Multi Aircraft Simulation System (MACS) software (Prevot, 2002). MACS is an environment for developing, setting up, and running real-time controller and pilot-in-the-loop simulations; it was configured to emulate the existing Air Traffic Management (ATM) system. The modified version of MACS included incorporation of UAS aircraft models with the addition of Stratway+ algorithms to drive the Electronic Horizontal Situation Indicator (EHSI), known as bands, which indicated a range of headings that would result in a loss of well-clear with one or more intruder aircraft. Muñoz, Narkawicz, Chamberlain, Consiglio, and Upchurch (2014) provide additional information regarding self-separation algorithms. The subject controller's workstation closely resembled the workstations that are currently used in FAA field facilities. STARS functionality was included in this experiment but with limitations. The implementation team included personnel from SGT, Adaptive Aerospace Group (AAG), and Intelligent Automation Inc. (IAI).

Table 2.
Horizontal miss distance acceptability rating scale.

	Horizontal Miss Distance Rating Scale Definition
Rating Scale	
1	Much too close; unsafe or potentially so; cause or potential cause for issuance of a traffic alert
2	Somewhat close; some cause for concern
3	Neither unsafely close nor disruptively large; did not perceive the encounter to be an issue
4	Somewhat wide, a bit unexpected; might be disruptive or potentially disruptive in congested airspace and/or with high workload
5	Excessively wide, unexpected; disruptive or potentially disruptive in congested airspace and/or with high workload

Results

Horizontal Miss Distances

Subject controllers were verbally asked to rate HMDs on a scale from 1-5, as shown in Table 2, based on their acceptability of the HMD spacing parameter.

Opposite-direction encounters. Illustrated in Figure 3, the ratings for the opposite-direction encounter geometry show that HMDs with a spacing parameter of 3.0 NM were considered unacceptable due to either being "*somewhat wide*" or "*excessively wide*." In addition, the graph also shows that the HMDs that the controllers' found to be acceptable were the ones in the 1.0 and 1.5 NM range with 80% of ratings suggesting 1.5 NM being the most acceptable among the two.

Overtake encounters. Figure 4 illustrates the ratings for the overtake encounter geometry. The graph shows that the highest percentages, with a rating of 3, were at the 1.0, 1.5, and 2.0 HMD spacing parameters. In addition, the graph also shows that a rating of more-than 3 was given for HMDs with a 2.5 or 3.0 spacing parameter.

Crossing encounters. Figure 5 illustrates the ratings for the crossing encounter geometry. The graph affirms that the controllers found the 1.0 and 1.5 NM HMD spacing parameters to be the most acceptable by giving a large majority of encounters, with those specific spacing parameters, a rating of 3 indicating that they were "*neither unsafely close nor disruptively large*" and "*did not perceive the encounter to be an issue*." HMDs of 2.5 NM had comparable percentage ratings of 3 and more-than 3. Furthermore, as was the case with the other two encounter geometries, HMDs with 3.0 NM spacing parameters, received a majority of ratings of more-than 3, indicating that those encounters were either "*somewhat wide*," or "*excessively wide*" and "*disruptive*."

Figure 3. Subject controllers' ratings for HMD spacing parameters for the opposite-direction encounter geometry.

Figure 4. Subject controllers' ratings for HMD spacing parameters for the overtake encounter geometry.

Figure 5. Subject controllers' ratings for HMD spacing parameters for the crossing encounter geometry. In this crossing encounter, the UA's speed was faster than the encounter aircraft.

In summary, the analysis of the data collected concludes that 1.0 to 1.5 NM were the most favored HMDs. It also concludes that the majority of subject controllers found that 0.5 NM to be considered "*much too close*" for all three encounter types. Furthermore, a majority of controllers found that 2.0 NM was not unreasonable but that 2.5 NM and above were considered disruptive.

Realism of Traffic Density and Workload Ratings

Careful consideration was taken in the design and realism of the simulation environment. Research was conducted to find the optimal traffic density allowable to achieve the aim of the study while maintaining as close to real-world densities as possible for a realistic simulation of the DFW East-side airspace. At the termination of each one-hour data collection run, an "end-of-hour questionnaire" was administered to each controller. Among the questions asked was one regarding the realism of the traffic density; controllers were asked to "*rate the realism of the traffic density of the simulation during the preceding hour*." The following responses are collective for all subjects for all six one-hour data collection runs: 0% of responses were that "*Traffic Density was significantly higher than in real operations*;" 1.2% of responses were that "*Traffic Density was somewhat higher than real world operations*;" 55.6% of responses were that "*Traffic Density was about the same as would be found in real world operations*;" 42.9% of responses were that "*Traffic Density was somewhat lower than real world operations*;" and

0% of responses were that "*Traffic Density was significantly lower than in real world operations.*" Table 3 shows the average workload ratings, captured at five-minute intervals using the ATWIT methodology, for all subjects and for all data collection runs.

Table 3.
Average Air Traffic Workload Input Technique (ATWIT) Workload Ratings.

ATWIT Time Intervals (*in seconds*)											
	300	*600*	*900*	*1200*	*1500*	*1800*	*2100*	*2400*	*2700*	*3000*	*3300*
Average Rating	1.37	1.79	1.84	1.68	1.93	1.89	2.15	2.37	2.08	1.89	2.01

Discussion

The CAS-1 research experiment employed a close-to-real world simulation of the DFW East-side airspace. The study focused on determining the effect of simulated DAA-equipped UAS on ATC workload, as well as, on the acceptability of maneuvers with differing HMD spacing parameters used in the DAA algorithms. The results of the study confirmed a clear favoring, from the ATC perspective, towards a particular HMD range, which was 1.0 and 1.5 NM; this range was still favored even when maneuvers were required to maintain those horizontal miss distances and appeared to be the optimal range for ATC acceptability. In addition, controllers found the DAA integration concept as presented to be absolutely viable. ATC workload ratings using the ATWIT method showed that the controllers considered the simulated workload to require minimal to low mental effort given their experience with the DFW sector.

Follow-on research studies in this series of experiments will focus on assessing the impact of modeled communication delays on the execution of SS procedures as defined in the CAS-1 experiment and the performance of the Stratway+ generated maneuver guidance in the presence of winds. In continuation of the aforementioned follow-on research, additional research studies will address minimum and maximum acceptable declaration times for projected well clear losses, from the perspectives of both the air traffic controller and the Unmanned Aircraft (UA) pilot. Data from the CAS-1 experiment and subsequent experiments are meant to play a crucial role in the FAA's establishment of rules, regulations, and procedures to safely and efficiently integrate UAS into the NAS.

References

Code of Federal Regulations. Title 14 Aeronautics and Space, Parts 60 to 109. Revised as of January 1, 2011. Retrieved from http://www.gpo.gov/fdsys/pkg/CFR-2011-title14-vol2/pdf/CFR-2011-title14-vol2.pdf

Federal Aviation Administration (2012). Standard Terminal Automation Replacement System (STARS). Retrieved from http://hf.tc.faa.gov/projects/stars.htm

Federal Aviation Administration Modernization and Reform Act of 2012. Retrieved from http://www.gpo.gov/fdsys/pkg/CRPT-112hrpt381/pdf/CRPT-112hrpt381.pdf

Muñoz, C., Narkawicz, A., Chamberlain, J., Consiglio, M., and Upchurch, J. (2014). A Family of Well-Clear Boundary Models for the Integration of UAS in the NAS. *Proceedings of the 14th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, AIAA-2014-2412, Atlanta, Georgia.

Prevot, T. (2002). Exploring the Many Perspectives of Distributed Air Traffic Management: The Multi Aircraft Control System MACS. *AAAI HCI-02 Proceedings*, 149-154.

Stein, E.S. (1985). *Air traffic controller workload: An examination of workload probe*. (Report No. DOT/FAA/CT-TN84/24). Atlantic City, NJ: Federal Aviation Administration Technical Centre.

U.S. Department of Transportation Federal Aviation Administration (2014). Aeronautical Information Manual: Official Guide to Basic Flight Information and ATC Procedures. Retrieved from http://www.faa.gov/air_traffic/publications/media/AIM_Basic_4-03-14.pdf

INCORPORATING NEW METHODS OF CLASSIFYING DOMAIN INFORMATION FOR USE IN SAFETY HAZARD ANALYSIS

Professor Nancy G. Leveson
Massachusetts Institute of Technology
Cambridge, MA
Major Daniel R. Montes
United States Air Force, Massachusetts Institute of Technology
Cambridge, MA
Professor Leia A. Stirling
Massachusetts Institute of Technology
Cambridge, MA

The increase of interacting humans and autonomous components in complex systems necessitates rigorous methods to classify domain information pertaining to controllers in the system. Systems-Theoretic Process Analysis (STPA) was developed at MIT as a method for identifying hazardous scenarios from a system design in order to generate functional system requirements to eliminate or control those scenarios. An STPA analysis, while systems-based and including human operators (e.g., pilots and air-traffic controllers) in the scenarios, is currently limited in the types of human contribution to accidents that it can identify (which are primarily related to situation awareness). This paper extends STPA in three ways: first, the analysis of the controller mental model was updated to include more system features; second, fundamental human-engineering considerations were added; and third, types and sources of decision-making influences that transfer from the planning cycle to the operations cycle were identified.

Humans play an important role in accidents (both positive and negative) and must be included in any hazard analysis performed during the development or field use of the system. Currently, safety investigators discuss human contributions in a simplistic way, if they include them at all (e.g., “pilot failed” or “pilot lost situation awareness”), and then they assign a probability to this “failure”. Such analyses are not very useful in designing to prevent accidents.

The MIT Systems-Theoretic Accident Model and Processes (STAMP) is a use-centered systems-modeling approach to understanding and preventing accidents (Leveson, 2012). Systems-Theoretic Process Analysis (STPA) is a hazard-analysis technique based on STAMP that is used to investigate system designs in order to generate functional safety requirements. STPA goes beyond the tendency to simply state that a human failed and investigates errors in the human’s mental model and errors in decision-making. This method, although advanced in terms of safety analysis, still oversimplifies the human’s role in complex systems because it is currently posed similar to investigating a machine controller’s model and decision algorithm. Human mental models contain more information about the system than a machine’s and develop using more sources of feedback.

This paper extends STPA methods of generating hazardous scenarios by refining how the human (or intelligent controller) is considered in the analysis. The methodology now identifies more system information the controller might use to make decisions during the operation, considers human-specific controller characteristics (e.g., workspace and human variability), and identifies organizational influences to controller behavior that originate before the operation. An overview of STAMP will be presented, followed by STPA extensions for intelligent controllers. The new techniques presented are meant to analyze currently existing systems, although some aspects may apply similarly in concept development and design.

System Safety Modeling

STAMP was inspired by cybernetics (Wiener, 1965) and systems theory (Von Bertalanffy, 1968), as well as the U.S. system-safety standards that evolved during the development of long-range guided missiles (Department of Defense, 2012). Dekker (2006) describes two types of accident models in existence today. The first considers physical-component reliabilities (including people and software) and finds failures that chain together in time and/or space and lead to accidents. This type of model gives rise to fault-tree analyses and failure mode and effects analyses, for example. The second type of accident model, of which STAMP is an example, treats the prevention of undesirable losses as a top-level set of system requirements, and then it generates constraints to meet these goals

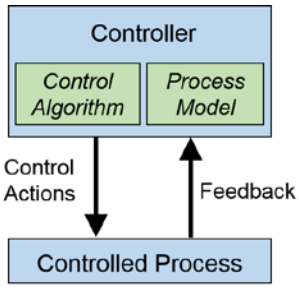


Figure 1. A basic control loop between functional entities in the safety control structure. Although a horizontal decomposition is not shown, disturbances, actions from other controllers, and communications would be modeled when appropriate.

through the functional system behavior. STAMP treats safety as a control problem. The main concepts of STAMP are safety constraints, the hierarchical control structure, and process models (Leveson, 2012).

Organizational stakeholders identify accidents[1], which are typically domain independent (e.g., environmental damage); then hazards[2] are identified, which are domain specific (e.g., chemical toxins exposed to the environment). The list of hazards is typically short because hazards may not include any engineering assumptions from the design (e.g., valve leaks and releases toxin). The prevention of each hazard becomes a *safety constraint*. Each hazard is mapped to one or more accidents. This mapping allows traceability from the findings of an ensuing hazard analysis (like STPA) all the way up to accidents.

The STAMP *hierarchical control structure* is an abstract model of the system design. It is a decomposition, starting at the top with legal/regulatory and organizational entities all the way to the lower-level components of the system operations. Higher levels have more responsibility, authority, and accountability than lower levels, and control-feedback relationships exist between levels. A general form (not pictured here) can be found in Leveson (2012). The control structure is a model of the functional system and not necessarily the physical structure. For example, suppose air-traffic control speaks to a drone operator through a UHF radio signal that travels from a control tower to the drone, is multiplexed and sent to the operator's ground control station via datalink band, and then demuxed into an audio channel in the operator's headset. While a physical schematic would detail the intricate connections just described, a STAMP control structure would show the tower personnel controlling the drone operator, who in turn would be controlling the drone.

In an STPA analysis, each level of the control structure is explored from the top down, and control relationships between entities are examined. A general control-feedback loop is shown in Figure 1. In "Step 1" of STPA, functional behaviors of a controller that violate safety constraints are identified as unsafe control actions (UCA), along with the system or environmental context in which they are hazardous. In "Step 2," causal scenarios that could produce each UCA are generated. This detailed analysis requires domain subject-matter experts (SME) because aspects of the physical design, hardware, software, and humans contribute to scenarios.

Currently, all the portions of the control loop—as well as any additional external information being used by the controller—are investigated to generate causal scenarios, including the controller itself and its *process model*. The process model is the controller's understanding of the states in the system it is trying to control. If the model states do not match the true system states, the controller could execute a UCA. In humans, this is called the mental model, although "process model" may be used generally. The following section discusses extensions to STPA that include refinements to the process model.

Extended Human Controller Analysis

Stringfellow (2010) and Thornberry (2014) previously elaborated on the human as a controller, with the former emphasizing that humans have a model of the organization, not just the controlled process, and the latter introducing a sequence for the Step-2 human-controller analysis. This extension will: 1) build on the existing

[1] Accident: An undesired or unplanned event that results in a loss, including human, property, environmental, mission, etc.
[2] Hazard: A system state or set of conditions that, together with a particular set of environmental conditions, will lead to an accident.

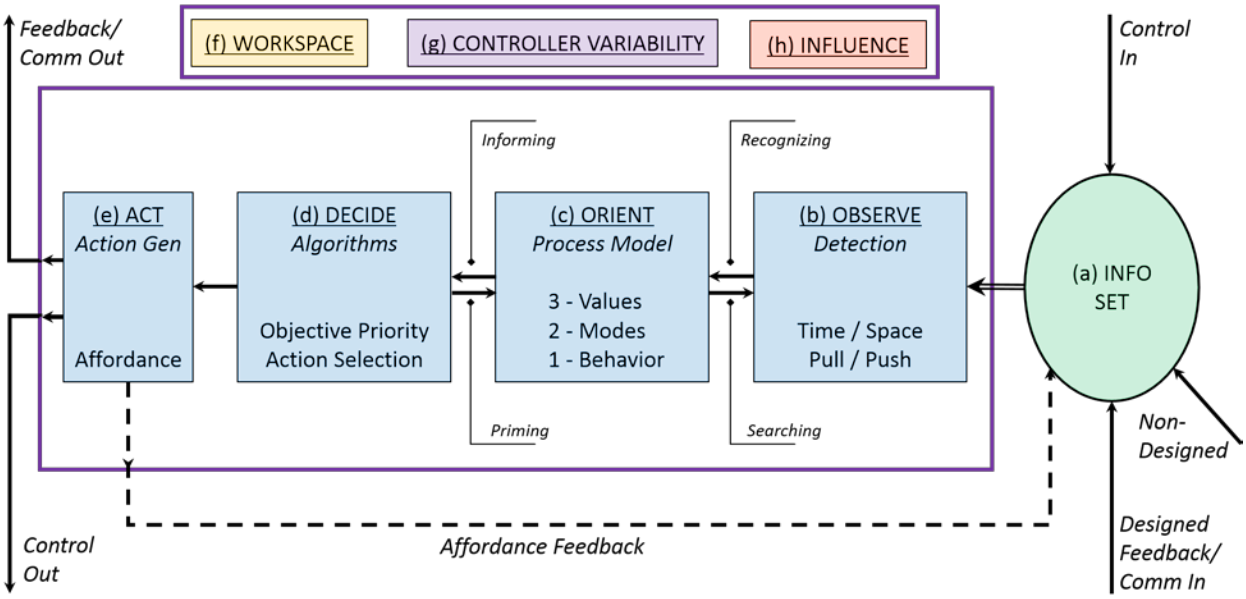


Figure 2. Extended intelligent-controller analysis sequence. Letters (a) through (h) denote areas for investigation.

sequence and add new sections, 2) refine the inquiries in some sections, and 3) introduce a method for identifying organizational influences on the controller. Figure 2 is the updated analysis sequence. It is *not* meant to be interpreted as an information-processing (in-the-head) model of human cognition, but rather as a set of considerations that map the controller to the work domain. The structure of parts (a) through (e) is maintained from Thornberry, while (f) through (h) are new. Parts (a) through (c) have been refined.

Part (a), the information set, corresponds to Dekker's "data availability" (2006) and helps identify all the information presented to the controller, including controls, feedbacks, and communications. This part has now been refined to explicitly differentiate between information available *as originally designed* to arrive at the controller and information outside original design intent that is being used by the controller regardless. An example of the latter would be a copilot looking at how the pilot's hands are displacing a traditional yoke instead of looking at the copilot's own flight displays, or two employees using an informal socio-organizational communication channel. The purpose of making the effort to delineate between design and non-design communications is crucial when new technologies and system upgrades threaten to change the nature of human-system interactions without properly documenting all the connections. Knowing that pilots were using "free" feedback (such as yoke movement) that goes away with an upgrade (such as fly-by-wire) is important.

Additionally, Thornberry (2012) emphasized that feedbacks the controller receives when an affordance is acted upon (such as a switch physically moving or a "bug" being set on an airspeed display) should be identified. This is important for similar reasons as non-designed information. For example, turning and removing a key from a classic (non-electronic) car ignition is sufficient feedback to the driver that the vehicle changed to a shutdown state; feedback from the controlled process (the car) was not required for the human to conclude the change had occurred.

Parts (b)-(e) in Figure 2 are named Observe, Orient, Decide, and Act after Boyd's O-O-D-A loop (2010). Part (b) corresponds to Dekker's "data observability" (2006) and performs inquiries on *if* and *how* data are attended to in time-space. A refinement here adds that data can be pushed to or pulled by the controller in several ways ranging as follows: the controller requests a controlled component or process for missing data, fetches already available data that is not yet displayed, refreshes an obsolete display, attends to a current display, or receives data immediately via her currently attended time-space (or through exogenous cueing).

Boyd (2010) emphasized that *orient*—part (c)—is the most vital investigation area for analyzing decisions in a complex adaptive system. This section of the analysis maps to the *process model* in STAMP. The investigation here has been refined to include three levels: behavior, modes, and values. Behavior represents how the controlled process is interacting with the mission environment. A behavior state variable might be a direct reading from the

Table 1.
Three types of modes (Leveson et al., 1997) and recommended inquiries that were added to the process model analysis.

Supervisory Structure	**The control relationships and communication links in the system hierarchy.**
	Which controllers currently have or share priority over each controlled component?
	Which controlled components may apply authority limits and under what circumstances? Can those limits be overridden? How will conflicts be decided (i.e., who should have the final authority?)
Component Operating Mode	**The set of algorithms that components under my control can use to exert control over their process(es).**
	What are the physical or logical assumptions and constraints associated with the component's current operating mode?
	What data in the information set is the controlled component using to inform its model?
	What input/and output format am I using with my controlled component(s)?
Mission Phase	**The specified set of related behaviors of the controlled system representing its operational state.**
	What mission phase is the system in (e.g., takeoff, cruise, etc.)
	Do all controllers know the current mission phase?
	Does a change in mission phase mode cause a change in supervisory structure and/or component operating modes (including input/output formats)?

information set, or it might first be translated into a more useful variable (e.g., altitude and airspeed displayed as energy). A supervisor might be monitoring a lower-level controller which is managing the process behavior. In this situation the next level of the process model analysis (mode) becomes important.

A *mode* is a mutually exclusive set of system behaviors (Leveson, Pinnel, Sandys, Koga, & Reese, 1997). There are three types of modes: *supervisory structure*, *component operating mode*, and *mission phase*. Stringfellow's (2010) model of the organization, for example, would fall under supervisory structure. Table 1 presents definitions and a minimum set of recommended inquiries that have been added to investigate modes in the process model. *Authority limits*—worth mentioning—are a type of lockout or interlock that controlled components may exercise, by design, to ignore a received control request if they know it to be hazardous to the system. For example, a flight-control computer can limit the angle of attack the pilot demands, or a pilot can disregard an air-traffic control request if she sees visual traffic in the way. These limits must be carefully analyzed to make sure they do not prohibit behavior that might be necessary in some situations. In addition, there must be some determination of who should have the final authority in case of a conflict.

Values contains two lines of inquiry. The first is *external values*, which is an understanding of any values the controller personally maintains outside the system. An example would be the personal pressure behind the classic "get-there-it is" that might prompt a pilot to ignore system-derived objectives and rush a landing. The second is *value mapping*, which is the controller's understanding of how values at higher abstractions of the system's means-ends hierarchy (Rasmussen, Pejtersen, & Goodstein, 1994) map to objectives at the controller's level.

While STPA investigates control algorithms as part of Step 2, there is no specific methodology prescribed for human decision-making (d), and one is not recommended here. Figure 2 highlights that the mental model is affiliated with observation (searching-recognizing) as well as decisions (priming-informing), both which a human controller cognitively attends to. Part (e), appropriateness of response mappings, is not refined here.

Workspace inquiries (f) are new to the analysis and include climate, visual and auditory noise, work physiology (e.g., "pulling Gs"), anthropometric and ergonomic compatibility, and workload. *Controller variability* inquiries (g) are also new and include age, perceptual acuity, attention capability, natural disposition, health, injury/disabilities/disease, psychological/emotional conditions, fatigue/stress/sleep cycles, and drugs/medications.

Findings in both (f) and (g) are human-specific considerations. In addition to being used to evaluate the operator population for current systems and applications, they could be used to create criteria during design development to select a future population. Part (h), called *influence*, is new and it is discussed next.

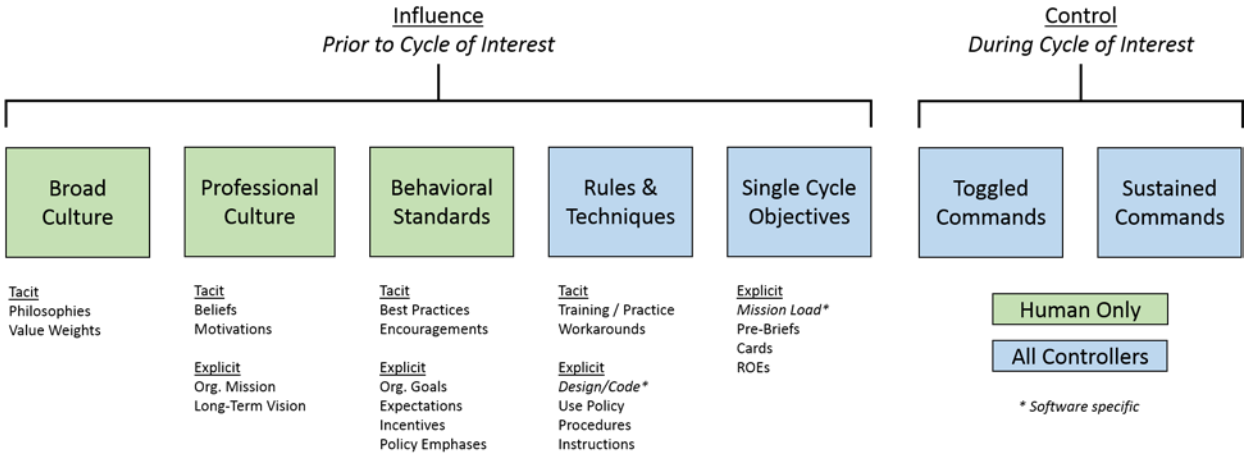


Figure 3. Sources of decision-making influences that evolve prior to the operation cycle. Green sections affect only humans.

Decision-Making Influences to the Operating Process

Influences can affect both human and machine (or software) controllers, although some are specific to only humans. The need to identify influences enforces the necessity of SME involvement during Step 2 of STPA. Socio-technical organizations contain large functional hierarchies, and often the lower levels of operation exhibit a smaller time constant (faster cycling) than higher-level managerial processes. Additionally, there may be planning or maintenance cycles that precede operating cycles, depending on the industry. Influences to operating controllers from these sources evolve before the operating cycle, and they are presented in Figure 3. Influences are controls, but because they do not occur in the real time of the operating cycle they are considered static during the operation (and thus can be incorporated into a section of the controller analysis).

Going from left to right in the figure, each consecutive block represents influences that evolve closer in time to the operating cycle. Influences can be unintentional. Conflicting policies or outdated procedures are an example of the wrong information making its way into the operating cycle. Sources of the influences can be explicit (formal, articulated, and codified) or tacit (not easily transferred via media). In a tacit example, operators while on their lunch breaks might complain about a certain aspect of a computer interface they use on the job, and those sentiments eventually evolve into a proclivity not to use that feature. On the other hand, an explicit influence would be a company policy letter stating to discontinue use of the feature based on data gathered from a formal employee reporting system. Explicit sources might be the only ones easily available to a safety analyst with little experience in the specific domain; SMEs will more readily understand sources of tacit knowledge in the organization.

Examples of Causal Scenarios – Autonomous Cruise Control

Suppose a human driver of a modern car has an option to use an autonomous cruise control (ACC) that maintains a set distance from traffic ahead of it. This feature cannot accurately detect position and velocity of other traffic during inclement weather. A published warning about this exists in the owner's manual, and a small icon that says "ACC-deg" lights up next to the "ACC-on" icon by the speedometer when the ACC is armed but experiencing sensing problems (it will not shut off automatically if this happens). An STPA of the car design begins by referencing top-level hazards, one of which says: "Car violates minimum safe velocity/distance separation to another vehicle on the road." In STPA Step 1, a safety analyst examining the control loop between the human and the ACC (see Figure 1) would generate several UCAs for that hazard, one of which says: "Driver does not disengage ACC during inclement weather."

For STPA Step 2, the analysis sequence in Figure 2 begins at the "Information Set" (data availability). Design feedback listed here includes "ACC-on" and "ACC-deg" icons, speedometer and engine instruments, and visual weather cues. SMEs might offer non-design feedbacks that include engine noise indicative of a struggling

ACC. Affordance feedback would include the feel of the button that engages or disengages ACC. An example of a causal scenario for the UCA would be: "Driver assumes she has disengaged ACC by pushing the ACC button but does not confirm that the 'ACC-on' light has extinguished." A causal scenario informed by the "Observe" section of Figure 2 would be: "'ACC-deg' icon not noticed by driver for [specific design reason(s)]."

The process model identifies behavior states, some of which are: velocity, distance to front traffic, and weather condition. Some mode states are: ACC on/off and ACC degraded/not. The analysis also investigates human mode knowledge, to include that ACC becomes brittle in inclement weather (even cloudy days that might otherwise seem safe), and that ACC communicates its status via the icons on the dashboard. An example of a causal scenario would be: "Driver does not know that the ACC-deg icon indicates a system limitation and continues to use ACC." A further inquiry into "influences" would look at "Rules and Techniques" and identify causal scenarios such as: "Car manual does not sufficiently emphasize the importance of monitoring for an 'ACC-deg' icon, even if weather appears normal." A causal scenario from "Behavioral Standards" would be: "Auto industry does not sufficiently emphasize that drivers of vehicles with autonomy should read the entire warnings section of their manuals." These scenarios are used by engineers to eliminate them from the system design or to control them. For example, the cruise control software could be redesigned or the driver interface might be improved.

Conclusion

This paper extends the generation of hazardous scenarios in STPA by classifying information useful for investigating human (or intelligent) controller contributions to system hazards. The analysis of the controller was improved by refining several sections, including looking at three levels of the process model that contribute to robust and flexible behavior. Fundamental human considerations were also added in the form of adding new sections covering workspace and variability, and an emphasis was added to differentiate design, non-design, and affordance feedback to the controller. Finally, influence was considered to capture organizational ties to operational behavior.

These techniques add to the already improved ability of STPA to go beyond targeting humans or software for making arbitrary errors. Considerations that contribute to hazardous scenarios have been refined. Ideally, these improvements can be used not only to investigate existing systems but to inform system design, particularly as software and autonomy capabilities improve.

Acknowledgements

The views expressed in this document are those of the authors and do not reflect the official position or policies of the Air Force, Department of Defense, or United States Government. This discussion covers part of doctoral research by the second author. Special thanks to Dr. Cody Fleming, Adam Williams, and Dajiang Suo for collaborating on many concepts that became the methods presented here.

References

Boyd, J. R. (2010). *A discourse on winning and losing* (Lecture notes). Retrieved from http://dnipogo.org/john-r-boyd/

Dekker, S. (2014). *The field guide to understanding human error*. Ashgate Publishing, Ltd..

Department of Defense (2012). *System safety* (MIL-STD-882). Retrieved from http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-882E_41682/

Leveson, N. (2012). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.

Leveson, N., Pinnel, L. D., Sandys, S. D., Koga, S., & Reese, J. D. (1997). Analysing software specifications for mode confusion potential. In *Proceedings of a workshop on human error and system development*, Glasgow, Scotland (pp. 132-146).

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. Wiley.

Stringfellow. M. V. (2010). *Accident analysis and hazard analysis for human and organizational factors* (Doctoral dissertation). Retrieved from http://dspace.mit.edu/handle/1721.1/63224

Thornberry, C. L. (2014). *Extending the human-controller methodology in Systems-Theoretic Process Analysis* (Master's thesis). Retrieved from http://dspace.mit.edu/handle/1721.1/90801/

Von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications* (Vol. 55). New York: George Braziller.

Wiener, N. (1965). *Cybernetics or control and communication in the animal and the machine* (Vol. 25). MIT Press.

INDIVIDUAL PROBLEM REPRESENTATIONS IN DISTRIBUTED WORK

Alicia Fernandes, Philip J. Smith, Ken Durham, and Mark Evans
The Ohio State University
Columbus, OH

Human-machine interfaces in distributed work systems provide external problem representations that activate the cognitive processes people use to perform their work. Appropriate design of such representations is an important factor in supporting complex work. In air and surface traffic management, problems are typically framed according to airspace constraints even for practitioners whose domain is the airport surface. Constraints are passed from the en route and terminal domains to the surface in the form of airspace constraints, with the displays available to Air Traffic Control Tower (ATCT) personnel communicating these constraints in airspace terms. However, ATCT personnel use a different mental model to manage departures. An exploratory study found that ATCT personnel very quickly transform airspace-centric constraints into surface-centric constraints, while still discussing the constraints with en route and terminal traffic managers using airspace-centric terms. They must continually perform such transformations due to the representation of the information provided to them.

External problem representations provided to agents in a distributed work system activate the cognitive processes practitioners use to reason about, and ultimately decide upon a solution to, a given problem (Smith, McCoy, & Layton, 1997; Zhang & Norman, 1994). While external problem representations are not necessarily re-created internally by the problem solver, they strongly influence the internal representation used to perform cognitive work and the way in which the practitioner frames the problem at hand. When engaging in coordinated activities, practitioners use these external representations as forms of communication. External representations that are incongruous with practitioners' cognitive work, however, may still prove useful as tools for sharing problem representations in a distributed work environment.

Air traffic managers often address airspace constraints by invoking initiatives that reduce traffic flows through the affected airspace. Such initiatives often take the form of Miles In Trail (MIT) restrictions, defining the longitudinal separation required between two aircraft operating on the affected route. When en route airspace is constrained, the Air Route Traffic Control Center (ARTCC) determines the MIT required to manage affected traffic and is likely to "pass back" the MIT constraint to any Terminal Radar Approach Control (TRACON) handing off aircraft to the ARTCC airspace. For example, "WAVEY 25MIT 1900-2200 ZNY:N90" indicates that the New York ARTCC (ZNY) requires the New York TRACON (N90) to provide 25 miles between any two aircraft using the WAVEY departure fix from 1900Z to 2200Z.

The TRACON, in turn, passes back the constraint to the Air Traffic Control Tower (ATCT). However, the TRACON uses a different separation standard than the ARTCC during normal operations (3 miles versus 5 miles), and so the TRACON may pass back a different MIT restriction for affected aircraft at the first radar hit upon takeoff (e.g., 15 MIT off the ground).

ATCT controllers have to ensure that restricted aircraft have the appropriate separation upon takeoff and try to avoid delaying them more than necessary to achieve the restriction.

Furthermore, these aircraft share departure runways with aircraft that may not be restricted (i.e., aircraft using different routes). ATCT controllers need to stage departures such that they can maximize runway throughput while minimizing the delay experienced by any one aircraft. This paper describes an exploratory study that identified departure management strategies used by ATCT controllers in the face of dynamic weather-related constraints.

Method

Structured interviews were performed with retired ATCT controllers and Traffic Management Coordinators (TMCs) who walked through a dynamic weather scenario and shared the strategies they would use to manage departures on the surface of a hypothetical airport.

Participants

Twelve recently retired ground controllers and TMCs with an average of 23.4 years of experience at busy facilities participated. Eight participants had formal experience as a TMC and 2 had unofficial experience as a TMC (such as filling in while a TMC was on vacation). Four participants had formal experience as an ATCT supervisor and 3 had unofficial experience as an ATCT supervisor. In addition, 7 had worked as TRACON controllers, 2 had worked as ARTCC controllers, 1 had worked as a flight service specialist, and one had worked for 15 years as an Air Traffic Control System Command Center (ATCSCC) traffic management specialist.

Major Airport (MJA)

This study used a hypothetical airport, Major Airport (MJA), shown in Figure 1. The study scenario involved departures from runways 18C and 18L. Note that 18C has two parallel taxiways and 18L has only one. MJA was embedded in the Collaborative Airport Traffic System, known as CATS (Fernandes, Smith, Spencer, Wiley, & Johnson, 2011).

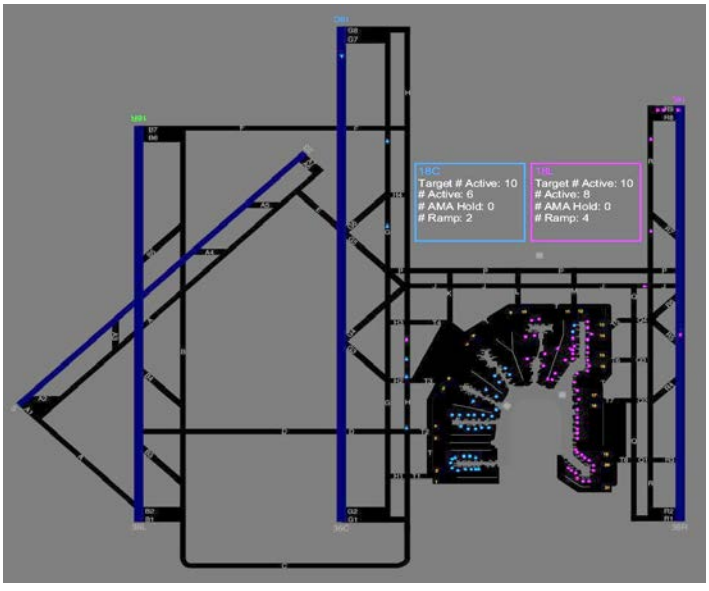


Figure 1. Layout of hypothetical airport used in the study

Simulation Scenario

The weather scenario consisted of actual current and forecast weather from July 26, 2010, in the Dallas, TX area. The researcher walked through the weather scenario from 1500Z to 0000Z, stopping every 30 minutes to allow the participant to view the two-hour forecast in 15-minute increments. At each 30-minute increment, the participant was asked about the surface management strategies they thought they would use in response to the weather. Then the participant was shown the current list of departure restrictions (generated by ARTCC and TRACON air traffic managers in a previous structured interview) and asked whether that information would impact their surface management strategy. The participant also would be shown the scheduled demand over the coming 30-60 minutes. For example, at 1800Z the participant would have access to displays similar to those shown in Figure 2.

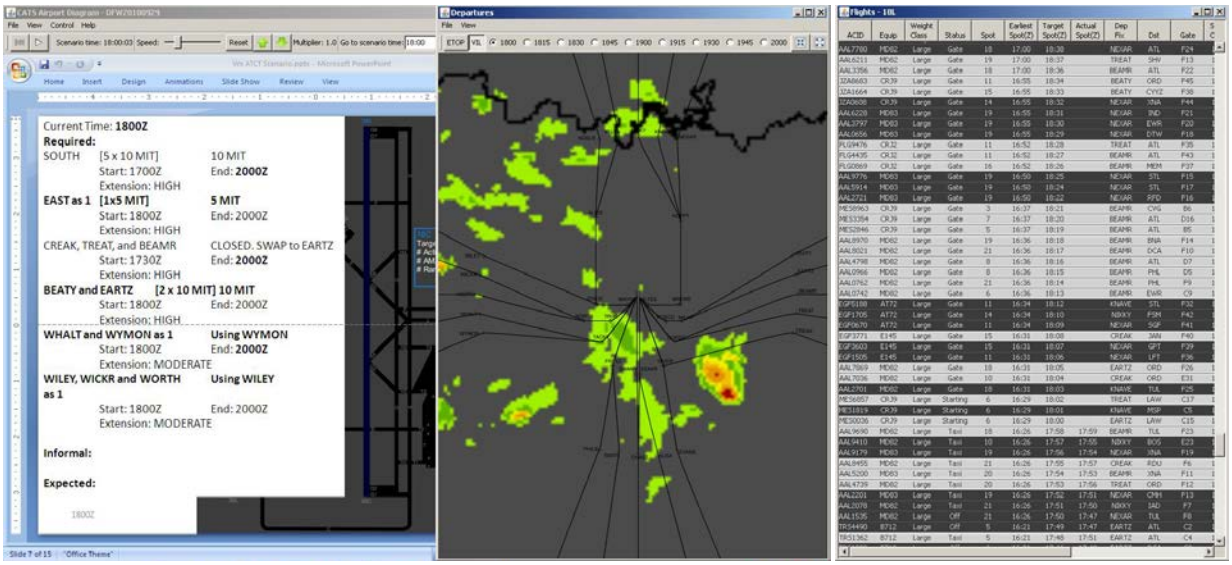


Figure 2. Displays similar to those shown to participants at 1800Z

Figure 2 shows the departure restrictions in place at 1800Z. Restrictions that had changed since 1730Z are shown in boldface. The list of departure restrictions is super-imposed over a map of the hypothetical airport. Participants were shown a map with no aircraft on it to try to avoid biasing their thinking in determining a surface management strategy. Participants also were able to see a list of aircraft already taxiing and scheduled to enter the movement area over the coming 30-45 minutes. In Figure 2, aircraft scheduled to depart to the east are highlighted, enabling the participant to quickly assess the demand for eastbound departure fixes in developing a strategy for staging those aircraft for departure.

At each 30-minute interval, the researcher updated the simulated weather, departure schedule, and departure restrictions. The researcher asked the participant questions such as:

- How would you want to stage flights for each departure runway?
- How many flights would you want in the lineup for each departure runway?
- Do you see anything in the weather that would cause you to change your plan?

- Do you see anything in the updated departure route restrictions that would cause you to change your plan?

Participants described the strategy they would use for staging departures given the weather and departure restrictions. It was hypothesized that participants would consider the weather forecast and the scheduled departure demand in developing a surface management strategy. Participants also were expected to use the taxiways to segregate aircraft by departure fix and direction when there were restrictions in place. In particular, participants were expected to use taxiways G and H and run-up pads G-7 and G-8 to stage aircraft for runway 18C and run-up pads R-8 and R-9 to stage aircraft for departure from runway 18L (see Figure 1 above). Thus, with two taxiways available for runway 18C and only one taxiway for runway 18L, participants were expected to use different strategies for staging departures for the two runways.

Results

Despite the difference in taxiway structure for the two runways, participants used similar strategies to stage departures for each runway. However, they expressed that they had greater flexibility in staging departures for 18C because it had two taxiways as well as an intersection at F from which any aircraft in the scenario could depart if and when it would be advantageous to do so. In addition, the surface management strategies were not so different when there were departure restrictions in place than when there were no departure restrictions. Due to space limitations, only strategies used for assigning taxiways and sequencing departures for runway 18C to accommodate the departure restrictions at 1800Z are discussed here.

At 1800Z, each of the southbound departure fixes had a 10 MIT restriction. Westbound routes Whalt and Wymon would be treated as one route and westbound routes Wiley, Wickr, and Worth would also be treated as one route until 2000Z. All northbound routes were open with no restrictions.

Six participants said they would assign flights to taxiways according to departure fix. Three of these said they would separate taxiways by effective route (i.e., departure fixes Wiley, Worth and Wickr on one taxiway and departure fixes Whalt and Wymon on the other). The other three said they would assign all westbound flights to one taxiway and all "splitters" to the other. "Splitters" is an ATCT term for unrestricted departures sequenced between restricted departures to achieve the required MIT. An important consideration in building a departure queue is the number of splitters to include.

The number of splitters an ATCT controller uses to meet an MIT restriction is an external representation of the translation they mentally perform when presented with the airspace-centric restriction. Participants were asked how they determine the number of splitters to use between any two aircraft subject to a 10, 15, or 20 MIT restriction. Their responses are shown in Figure 3. One participant said, "6,000 feet [and airborne] will get you 2 ½ to 3 miles, you've got 3, 6, 9 miles," referring to the minimum separation requirement for departures whose headings diverge by at least 15 degrees (FAA, 2014).

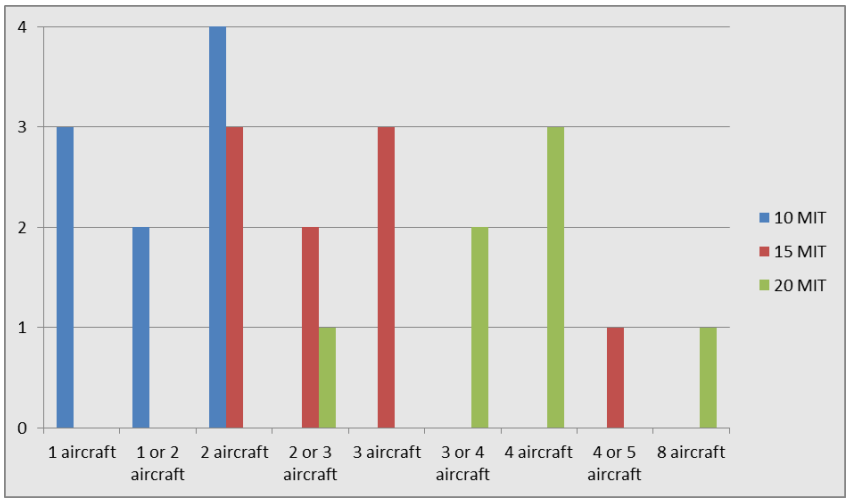


Figure 3. Number of splitters between aircraft subject to 10, 15, or 20 MIT

Some of the participants said they would vary their strategy by aircraft type, but their overriding concern was ensuring that they provided no more than the required MIT because that would represent wasted capacity. One participant said that controllers "don't want to provide any more than that number because if you provide more than that number then you're probably self-imposed restrictions and it's not a good thing. So you want to be right on the dot with that…"

Discussion

ATCT tools describe airspace constraints in terms of air route restrictions such as miles in trail, when in fact ATCT personnel transform these restrictions into surface management strategies involving departure staging locations, aircraft characteristics, and splitters. Such differences in problem representations have consequences for the design of tools to support airport surface management personnel in performing their work as well as supporting inter-facility coordination and collaboration throughout the NAS.

For example, ramp controllers stage aircraft leaving the ramp area in a way that they expect to be efficient for the ground controllers. However, they rarely have explicit information about the strategy the ground controllers are using and so may not actually stage departures in an efficient manner (Borgman & Smith, 2010). In addition, explicit representations of surface management strategy may support Surface Collaborative Decision Making (Fernandes, et al., 2012; FAA, 2013) and other decision support tools (Atkins, Churchill, & Capozzi, 2013; Brinton & Lent, 2012).

Acknowledgements

The FAA Human Factors Research & Engineering Group coordinated the research requirement, and its principal representative acquired, funded, and technically managed execution of the research service. In addition, the authors would like to thank Dr. Rich DeLaura of MIT Lincoln Laboratory, who provided the weather images from the Corridor Integrated Weather System (CIWS). The research was conducted as part of the first author's doctoral dissertation.

References

Atkins, S. C., Churchill, A., & Capozzi, B. J. (2013). Sensitivity of NASA's Spot and Runway Departure Advisor to traffic forecast errors. *Aviation Technology, Integration, and Operations Conference.* Los Angeles, CA.

Borgman, A. D., & Smith, P. J. (2010). *The Integrated Management of Airport Surface and Airspace Constraints for Departures: An Observational Study of JFK, EWR, and IAH.* Columbus, OH: The Ohio State University Technical Report #CSEL 2010-09.

Brinton, C., & Lent, S. (2012). Departure queue management in the presence of traffic management initiatives. *2012 Integrated Communications Navigation and Surveillance (ICNS) Conference.* Herndon, VA.

FAA. (2013). *U.S. airport Surface Collaborative Decision Making (CDM) Concept of Operations (ConOps) in the near-term: Application of Surface CDM at United States airports.* Washington, DC: Federal Aviation Administration.

FAA. (2014). *Order JO7110.65V Air traffic control.* Federal Aviation Administration, Air Traffic Organization. Washington, DC: Department of Transportation.

Fernandes, A. B., Smith, P. J., Spencer, A., Wiley, E., & Johnson, D. (2011). Collaborative Airport Traffic System (CATS) to evaluate design requirements for an airport surface departure management system. *Proceedings of the 55th Annual Meeting of the Human Factors and Ergonomics Society.* Las Vegas, NV.

Fernandes, A. B., Smith, P. J., Weaver, K., Durham, K., Evans, M., & Johnson, D. (2012). Identifying support requirements for airport departure management. *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society.* Boston, MA.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man and Cybernetics, 27*, 360-370.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*, 87-122.

EXPERIMENTAL INVESTIGATION OF FLIGHT CREW STRATEGIES IN HANDLING UNEXPECTED EVENTS

Joris Field[1, 3]
[1]National Aerospace Laboratory NLR,
Amsterdam, the Netherlands
Rogier Woltjer[2], Amy Rankin[2],
[2]Computer & Information Science, Linköping University,
Linköping, Sweden
Max Mulder[3]
[3]Control and Simulation, TU Delft,
Delft, the Netherlands

This paper reports a flight simulation study where airline flight crews were to handle unexpected situations, as part of the "Manual Operations for 4th Generation Airliners" (Man4Gen) EU research project. The analysis of their behaviour combined a cognitive systems engineering perspective with behavioural analysis methods used in aviation industry. Hollnagel's Contextual Control (COCOM) and Extended Control (ECOM) Models are applied to examine the strategies with which the flight crew responded to the simulated events. The outcome of this analysis is combined with the results of industry expert analysis of the actions that flight crew were expected to perform. ECOM illustrates the different strategies applied to handling the situation.

Modern civil aviation has become an extremely safe mode of transport, an accomplishment that can be largely attributed to the application of reliable and advanced aircraft systems. Despite the high levels of automation found in all main aircraft systems, ranging from fuel control to flight control, the flight crew remains to be responsible. In the rare circumstances that automation fails, the crew can be faced with (a chain of) unexpected events and are in those situations expected to respond appropriately. The "Manual Operations for 4th Generation Airliners" (Man4Gen) EU FP7 research project is investigating the risk assessment and decision making strategies applied by flight crews facing an unexpected situation in a modern airliner (see www.man4gen.eu). The project aims to contribute to *short-term* recommendations to the aviation community, to increase the overall resilience of the crew-aircraft system. Cognitive Systems Engineering (CSE) techniques are applied to investigate and better understand the actions, behavioural patterns and strategies of the flight crew-aircraft Joint Cognitive System (JCS). The research question investigated in this paper is: Which strategies are applied by the flight crew to recognise unexpected threats and respond? We will report the results of an experiment that was carried out in a research flight simulator, where twelve flight crews were subjected to a scenario with three unexpected events. Hollnagel's Extended Control Model (ECOM) (Hollnagel & Woods, 2005) is used to identify the difference in strategies applied by the crews, and related to performance measures currently being used in aviation industry.

Background

CSE sets out to investigate the ways in which people work within the applicable context for the work, here the flight crew operating in the flight deck. Studying work practice in an operational setting warrants the main contextual factors to be included in the analysis; factors such as the influence of organizational, as well as cognitive and situational demands (Woods & Hollnagel, 2006). By examining crew behaviour in an operational setting, the identification of interactions between the crew members, as well as with the aircraft and systems are included in the analysis. Within CSE, both people and technical systems are considered as elements collaborating as a Joint Cognitive System, which enables an analysis of how the humans and systems function together. CSE methods analyse the behaviour of this JCS to describe the patterns and characteristics of observable behaviour (Hollnagel & Woods, 2005). In the experiment described here we consider both pilots – Pilot Flying (PF) and Pilot Monitoring (PM) – along with the aircraft automation and systems, as the JCS.

At the core of the CSE perspective is the closed loop relationship between human perception and action, which can be illustrated by both ECOM and the Contextual Control Model (COCOM) sensemaking and control loop (Hollnagel & Woods, 2005). COCOM and ECOM models have been applied in the analysis of human-machine systems – in aviation and beyond (e.g., Feigh, 2010; Kontogiannis & Malakis, 2011; Rankin, Woltjer, Field & Woods (2013) described how COCOM can be applied within the crew-aircraft JCS context. The knowledge and experience of the flight crew form the basis of the interpretation of a situation the crew encounters. Their understanding of the situation is built up from interacting with the cockpit displays and interfaces, as well as the physical cues (noise, vibrations) from the aircraft. This understanding is what the crew's actions and decisions are based on. Their actions, combined with external events, yield useful feedback which modifies the understanding, forming a closed perception and control loop.

The ECOM consists of four parallel control loops, very similar to the COCOM loop described above, where cognition is described as control (Hollnagel & Woods, 2005). ECOM describes the multiple layers of performance of the crew-aircraft JCS, illustrated in Figure 1. We applied this model to examine the distribution of tasks and roles across the crew members and aircraft systems (see more elaborate method description in Field, Rankin & Woltjer, 2014). ECOM can be used to describe how both anticipatory and compensatory control is exercised by the JCS, in this paper primarily in response to an unexpected event in the experiment simulation. As the crew-aircraft JCS responds to the situation, the distribution of activities over the ECOM layers may change. An example of this is how the use of different automation levels affects the crew's actions and interaction with the aircraft.

Method

Experiment

The experiment applied an operationally relevant situation for trained line pilots, and aimed to investigate how the crew deals with an unexpected event. Of particular interest was to study the crew risk assessment and decision making. The scenario included events that crews were unlikely to have encountered during routine training. Events were designed to address the main project goals: automation failure and reversion to manual control; an event that required authoritative decision making; a challenging and ambiguous situation.

The experiment was carried out with a total of 12 crews of line pilots, both captains and first officers – a total of 24 pilots. All crew members were active line pilots or recently retired. Crews were unaware of the events in the scenario, and were instructed to treat the scenario as a normal operational flight. The NLR "GRACE" research simulator was used for the experiment; the flight deck was set-up in a Boeing 747-400 configuration.

The experimental scenario was developed by operational experts within the consortium including representatives from aircraft manufacturers, operators and training organisations. The scenario focused on the final descent and approach phase to an airfield, after a long-haul flight. Three key events in the scenario formed the unexpected events that were being studied. The first occurred during the final approach to the runway – an increase and shift in the wind, destabilising the approach path leading to a go-around. An additional loss of visibility at the decision height would force a go-around if necessary. During the go-around, the second event occurred, which was a subtle failure of the autopilot heading control that would necessitate a reversion to manual control to regain control of the aircraft heading. The third event was a birdstrike during the go-around climb-out that caused a failure of engine 1, and damage to engines 3 and 4. The damaged engines would surge and stall until thrust was reduced on those engines, at which point the aircraft could be stabilised. The crews were free to decide the appropriate response to the failures, and to decide on the course of action – for example returning to the airfield for a landing, or stabilising the aircraft and diagnosing the problems before landing.

The research simulator was set up to record data for the analysis of the flight crew's actions, decisions and behaviour, including simulator log data, audio and video recordings. At the end of the experimental scenario, the flight crew were debriefed by the project researchers. Recordings of the debriefings were transcribed, and contributed to the analysis. The crew's communication and actions were captured during the analysis by transcribing the video and audio recordings.

The data used in the COCOM/ECOM analysis presented in this paper consisted primarily of the observation log, video data (of the cockpit, flight crew and displays), and audio recordings from the flight deck and debriefing interviews. Performance of the crews as compared to expected actions and decisions was rated by three industry experts, to determine the crews that exhibited desirable and undesirable behaviour. The performance ratings were an account of the decisions made and actions taken as either carried out or not carried out.

Analysis

For the description of the degree and kind of control that the crew-aircraft JCS displayed during the simulator session, the COCOM and ECOM models were translated to an operational context for classification of the observations from the experiment. The ECOM was used to classify the behavioural patterns, and the COCOM was applied to assess the degree of control.

To operationalise the ECOM, the four layers - Targeting, Monitoring, Regulating and Tracking - have been defined through the experimental data and context of the crew-aircraft JCS, as described in Figure 1. Assigning the observations to the different layers was done through an iterative process of classifying the observations based on the theoretical descriptions of the ECOM model (Hollnagel & Woods, 2005). Dataset from two crews were used to develop the classification scheme using three independent raters. This classification scheme was documented

and extended iteratively reaching full inter-rater consensus while being applied to the remaining datasets.

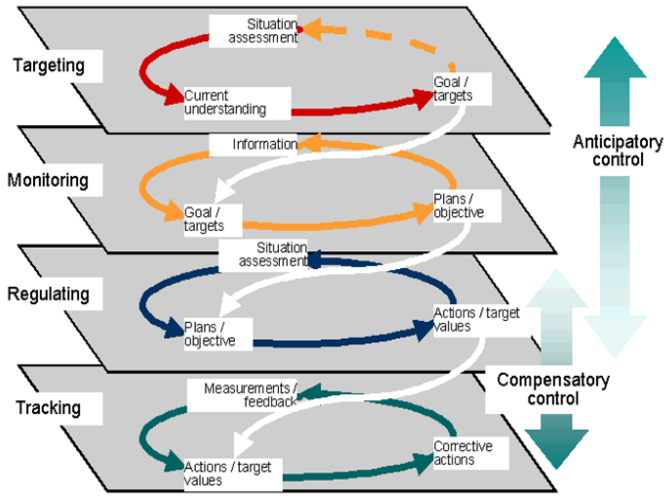
Extended Control Model	ECOM layer	Activities in flight context
	Targeting	Setting and resetting high-level goals of what may happen in the future
	Monitoring (planning)	Carry out monitoring and planning actions, where is the A/C going to be and how it is going to get there.
	Regulating	Executing the tasks as part of the plan that has been put together in the monitoring layers.
	Tracking	Actions that have a direct impact on the A/C control surfaces, direct effect on the A/C behaviour

Figure 1. The Extended Control Model, ECOM (Hollnagel & Woods, 2005), with activities described in a flight context.

Patterns of activities in relation to context were classified according to the four ECOM layers were identified, for a selection of flight phases/segments and crews: Engine management (after birdstrike); Trajectory management (second approach after Go-Around) to landing; Manual reversion (after HDG failure).

The COCOM classification scheme describes the degree of control that the crew-aircraft JCS has in a specific time period of performance. A classification of the control mode (strategic, tactical, opportunistic, scrambled) per flight phase was made using the literature definitions proposed by Hollnagel & Woods (2005), based on three parameters: (i) subjectively available time; (ii) evaluation of outcome, and (iii) selection of action. These were assessed by two of the project researchers that also performed the ECOM classification.

Results

This paper reports the trajectory management (second approach after Go-Around) flight phase analysis, and focuses on the broadest variability of performance between the crews – using 6 crews to illustrate the differences. The ECOM classification was used to identify patterns of observed performance within the ECOM layers. Then the subjectively available time, kind of evaluation of information, and what information was used for the basis of decision and action were assessed, from which a COCOM control mode was identified. The industry expert ratings of the crew performance were related to the ECOM and COCOM results in the sensemaking analysis, as a rough indication of how the crew performed against the key actions expected for specific moments in the scenario. These 6 crews - the top and bottom 3 performers (as rated by the industry experts) - illustrate a description of the ECOM control strategy and COCOM control mode for the activity of performing the second approach, i.e., flying the approach to landing after the go-around and birdstrike. Results are illustrated in Table 1, which includes the percentage of the score that the industry raters assigned to each crew out of the maximum number of

performance events related to this specific activity. The trajectory management activity is to a large extent concurrent to the activity of engine management, thus, prioritization that the crews did between the two activities was included in the analysis of control modes.

Table 1. *Analysis results for trajectory management assessment and decisions.*

Crew	ECOM strategy	COCOM control mode[1]	Performance %
6	Evaluation, double check, actions follow through across levels	Strategic/tactical	75
10	Prioritize evaluation and problem solving, actions, risk assessment and evaluation. Take time.	Strategic > tactical	75
11	React and extend planning horizon, buy time	Tactical > Strategic	75
5	Identify & assess, consider alternatives in plan and execution, prioritize flying	Tactical (opportunistic)	38
12	Actions without prior evaluation, information not discussed	Scrambled (opportunistic)	13
4	No discussion, limited evaluation, aim to land	Opportunistic (scrambled)	0

In general, higher performance (according to industry performance rating definition and scoring) seems to relate to patterns where there is interaction between the ECOM layers of Targeting, Monitoring, Regulating, and Tracking. Most activities are triggered as part of procedures or checklists at the monitoring/planning layer and then subsequently discussed between the crew at the regulating layer; decisions for actions are made, and finally implemented at the regulating and/or tracking layers. If, on the basis of feedback and evaluation, minor adjustments need to be made by the crew to the execution of the plan; this is then done at the regulating layer. If the trajectory needs to be changed to reach the same goal of the flight, these "flight plan" changes are discussed and decided at the monitoring layer, while higher level goals and prioritization (e.g., runway, alternate airport) is done at the targeting layer. Thus, if there is a regular and frequent interaction between the activities at the various layers of control, performance tends to be of a better quality. For these complex events crews are required to act simultaneously at multiple layers, determining strategies for multiple activities.

The ECOM layers analysis shows that crews with less desirable performance tend to have difficulties in the follow-through and follow-up in the interactions between the layers. For example, if monitoring/planning decisions and observations are not lifted to the targeting layer when necessary , important considerations regarding choice of runway, and consideration of alternate, and other trade-offs and prioritization of goals may be disregarded. This in turn may lead to lower-layer activities that could be better adjusted to situational circumstances if they would be evaluated and reoriented by higher-layer activities, but instead continue to execute plans that are not well-adjusted to circumstances.

Though this ECOM/COCOM analysis was applied in an experimental setting, it is possible to apply the method to understand how crews respond to events – in a training session for example. Identifying the behavioural patterns would help instructors to understand the

[1] Mode 1 / Mode 2 means characteristics of Mode 1 and 2, roughly to the same extent. Mode 1 > Mode 2 means first mostly characteristics of Mode 1, then mostly of Mode 2. Mode 1 (Mode 2) means mostly characteristics of Mode 1 with characteristics of Mode 2 to a minor extent.

performance of the crews, and identify how crews can be assisted in the ways that they develop their decision making and problem solving strategies.

Conclusions

COCOM and ECOM classification schemes were applied in an operational context for the cockpit environment of the NLR B747 experiment. This operationalization should be seen as a result of the study as this is (to our knowledge) the first application of COCOM/ECOM to a cockpit environment. In general, higher performance (according to industry performance rating definition and scoring) seems to relate to patterns where there is interaction between the ECOM layers of Targeting, Monitoring, Regulating, and Tracking. Crews with less desirable performance tend to have difficulties in the follow-through and follow-up between the interactions between the ECOM layers. The analysis contributes to an in-depth understanding of crew actions anchored in an operational context as well as an academic understanding of CSE contextual control concepts.

Acknowledgements

The authors would like to thank the flight crews that participated in the Man4Gen project, as well as the experts and other consortium partners that contributed to the experiment work at NLR. The Man4Gen research is funded as part of the FP7 2012 Aeronautics and Air Transport programme under EC contract ACP2-GA-2012-314765-Man4Gen. The views and opinions expressed in this paper are those of the authors and do not necessarily represent the position and opinions of the Man4Gen consortium and/or any of the individual partner organisations. If you have any questions regarding the Man4Gen project, please contact man4gen@nlr.nl.

References

Feigh, K. M. (2010). Incorporating multiple patterns of activity into the design of cognitive work support systems. Cognition, Technology & Work, 13(4), 259–279.

Field, J. Rankin, A. & Woltjer, R. (2014). Modelling Flight Crew Strategies in Unexpected Events: A Cognitive Systems Engineering Perspective. *Proceedings of the 31st Conference of European Association of Aviation Psychology.* 22-26 September 2014, Valletta, Malta.

Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. Boca Raton, FL: CRC Press/Taylor & Francis.

Kontogiannis, T., & Malakis, S. (2011). Strategies in controlling, coordinating and adapting performance in air traffic control: modelling "loss of control" events. Cognition, Technology & Work, 15(2), 153–169.

Rankin, A., Woltjer, R., Field, J., & Woods, D. (2013). "Staying ahead of the aircraft" and Managing Surprise in Modern Airliners. In I. Herrera, J. M. Schraagen, J. Van der Vorm, & D. Woods (Eds.), *Proceedings of the 5th Resilience Engineering Association Symposium* (pp. 209–214). Soesterberg, NL: Resilience Engineering Association.

Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. Boca Raton, FL: CRC Press/Taylor & Francis.

STATISTICAL ERRORS IN AVIATION PSYCHOLOGY: COMMONSENSE STATISTICS IN AVIATION SAFETY RESEARCH

Christopher D. Wickens
AlionScience & Colorado State University Fort Collins Colorado

I discuss problems with the use of null hypothesis significance testing, as it is particularly applied to safety research such as that in aviation psychology. Such problems are manifest in the inherent bias of traditional statistics to avoid type 1 statistical errors, and hence to discourage findings of safety improving effects as significant, when low powered experimental designs are required by necessary constraints. In contrast, I offer several approaches or remedies. Researchers should think about the decisions made by consumers of their research, based on the costs and values of those decisions; they should form alternative hypotheses, use smart planned comparisons where possible, and present data on the size of effects that do not meet conventional .05 levels of significance. Meta analyses are also encouraged.

In a hypothetical research project, investigators have examined an instructional program to train pilots to better understand the flight management system modes, and respond appropriately to unexpected surprises. A group of 20 line pilots from commuter airlines are selected to go through either conventional training or the augmented "understanding" instructional program. A transfer of training experiment is then done and after a 1 week delay pilots are confronted with an unexpected configuration of the FMS in a high fidelity simulator; the time until the initial correct diagnosis and response is recorded for each of the 10 pilots in the two groups. The authors report a mean RT of 9.5 seconds for the "understanding" group and of 14 seconds for the control group, a non-significant ($p>.05$) effect. A follow up study, with a slightly revised "understanding" curriculum is carried out later with 16 pilots (8/group), and it also provides a non-significant ($p>.05$) benefit, here of 3 seconds. It is concluded, based on the two studies showing no significant benefit, that the new curriculum is non effective. The developer of the curriculum points out to the investigators that if the samples of the two studies are pooled, with a resulting N=18/group, the mean benefit, now of approximately 4.5 seconds would have proven significant ($p<.05$). Furthermore, examining more closely the statistics that underlay the two experiments, the developer noted that the two p-values were, respectively, .07 and 0.11.

This hypothetical (but plausible) research scenario illustrates the potentially serious flaws in the manner that classical null-hypothesis significance testing (NHST) is applied in our safety-critical profession of aviation psychology. The likely conclusion by the research sponsor, and airline training groups that the technique was "ineffective", quite possibly results in a decision not to adopt it, and perhaps the resulting failure to take steps that could prevent serious FMS-related mishaps down the line.

In the following, I will outline some of the main concerns underlying the above sequence of events, and suggest remedies that might ameliorate some of these concerns. I will be drawing on some of my previous thinking about "common sense statistics" (Wickens, 1998), which itself was inspired by an earlier article on aviation safety by Don Harris (1991), as well as the more recent seminal article by Cumming (2014) on "the New Statistics". Cumming writes much more than room allows here to summarize that is relevant for our profession.

Five flaws in conventional statistical thinking.

Flaw #1. The p-value is a dichotomous , black-white cut off of significance, at p=0.05.

Fisher (1925), who developed the concept of the p value, never intended it to be used as a dichotomous criterion. The concept represents a continuum of the degree of evidence, in support of a hypothesis given the data. No different from degree of altitude on approach to a runway, or degree of temperature, there may be certain relatively important values along these continua, akin to a 25,000 foot "sterile cockpit" altitude on approach; or a 32 degree freezing point, but this certainly does not mean other changes in the variable are unimportant or to be disregarded; despite assertions (by many reviewers) that a $p>.05$ is "just non significant, and should not be talked about as if it were" (paraphrasing from several reviewers of my submitted

manuscripts). Indeed such dichotomous thinking can often lead to what I have referred to as "statistical illogic" of dichotomous thinking as in the following case: An ANOVA reveals a "significant" workload effect on performance, across three levels, low, medium and high; but then separate post-hoc comparisons reveal that the low-medium contrast is NS or "statistically equal", as is the medium-high contrast, while the low-high contrast is significantly ($p<.05$) different. So if L=M and M=H. Then in the logic of comparisons L must equal H. But the third comparison shows that this is not true: a contradiction.

Flaw #2. p = 0.05 represents a "decision rule".

It does not. The decision rule is defined by alpha, set by the experimenter. It can be at any p value chosen (convention often does select .05), and so the black-white thinking associated with the p value is more correctly associated with alpha. Whereas Fisher conceived of p as a continuous "evidence variable", it was Neyman and Pearson (1933) who developed the logic of the decision rule often associated with NHST; (Hubbard & Bayarri, 2003), to firmly either "accept" the alternative hypothesis, or "accept the null" (reject the alternative). If we can think of the traditional statistical analysis package as a form of automation, the distinction between the p value and alpha very closely parallels the distinction between stage 2 automation (information integration and inference): the p value (or its closely associated confidence interval) and stage 3 automation (decision making): the alpha level, with its associated decision to "accept" or "reject" an effect as meaningful. As we have pointed out elsewhere (Parasuraman Sheridan & Wickens, 2000: Onnasch, Wickens et al, 2013), errors of automation at stage 3 have more problematic consequences than those at stage 2. And as we see below, such automation can easily make errors.

Flaw #3 NHST is biased toward the status quo.

Table 1 presents the standard matrix underlying NHST. Across the top there is some "ground truth" effect that exists in the world (population) that we are trying to confirm. Here, this might be the truth that our experimental manipulations will make people better pilots and hence improve safety. We run the experiment, compute the statistics and derive a conclusion, based on whether our p value exceeds or is less than alpha (which is conventionally set to be.05). By accepting a criterion of .05, our decision rule is designed to "assure" that given our data, there **is** only one chance in 20 that we will conclude there is an effect, if the experiment is repeated multiple times (Cumming, 2104), when there is actually none to be found in the population; an errant conclusion resulting from the contributions of randomness to the data. This is the **type 1 error.**

Table 1.
The conventional table of statistical decisions with NHST

	State of the world: "the truth"	
Experimental results	Improve safety	No improvement
Disconfirm Ho (e.g., $p<.05$): An effect		Type 1 error (.05). Strongly discouraged
Confirm Ho ($p>.05$). "NS"	Type 2 error	

In contrast, conventional NHST is silent on the probability of concluding that there is **no** effect, when there actually **is** one, as shown in the bottom row, concluding a "non significant effect" such as that described in our story above: **a type 2 error**. This is a real number, which can be estimated from statistical power calculations (Cohen, 1988). But application of conventional statistics places far less emphasis on this, than it does on keeping the type 1 error below .05; and as a result, most experiments show a bias toward a considerably higher probability of the type 2 than the type 1 error. In essence, there is a direct analogy to our criminal justice system that cares much more to avoid an innocent person being found guilty than the converse; and hence requiring unanimous jury decisions to convict, and only one dissenting member to exonerate. The standard of evidence for guilt is set very high, just as in traditional NHST, the standard of avoiding a type 1 error is set high. This state of asymmetric concern for avoiding type 1 more than type 2 errors is an outgrowth of concern in basic sciences, that somehow type 1 errors are "worse" than type 2 errors, and a case can be made that the scientific community does not want a plethora of effects claimed, that turn out to be "untrue". But should this bias apply equally to safety research? As I argue below, it should only apply less severely, leading to the 4th "flaw".

Flaw #4 NHST does not address values in decision making.

Table 2 presents a classic decision table in expected value theory, populated by the specific characteristics of our automation training example above. It is similar in some respects to table 1, but also quite distinct. The

two possible states of the world regarding a "ground truth" are again shown across the two columns, and the two rows again represent decisions. However these are not decisions to reject or accept the null hypothesis by the researcher, but instead represent decisions, made the consumer of our safety research, to either adopt the concept suggested y the experimental results (e.g., implement the training plan) or ignore it. This is a very different form of decision than that made by the researcher to "decide" to say in print, whether the effect is "significant" or not.

Table 2.
The classic expected value decision matrix.

	State of the world	
Decision	Improves safety (P)	Does not improve safety (1-P)
It works: adopt the procedure	Mishaps saved cost of adoption	Cost of adoption
It does not work. Discard the procedure	Unnecessary mishaps created	No cost

Most importantly, we can now depict specific costs and benefits of different outcomes, particularly for the two types of decision "errors" that corresponded to the type 1 and type 2 errors in Table 1. Rather than, as in table 1, simply saying "type 1 errors are worse than type 2 errors", one can begin to put some approximate numbers on these to make a more objective judgment, as indicated by the cells of table 2.

Flaw #5 NHST does not address probabilities in decision making.

A second feature of the decision matrix in table 2, is the explicit presentation of some a-priori estimation of the prior probability (p) that the state of the world is true, as shown across the top row. These are quite different from the probability value depicted in table 1, which only represents the probabilities that the alternative hypothesis (H1) is falsely accepted, **given the data** observed in the experiment. The two probabilities have totally different meaning. Of course there is often no basis for making such prior probability estimation in Table 2, but by setting it a 50/50 for example, you might be essentially saying that "independent of, or prior to observing my experimental results, I think it is just as likely that my intervention will improve safety as not". This starts both hypotheses with even odds of confirmation

There is certainly plenty of debate in the statistical and experimental world about the role of prior probabilities in hypotheses testing, an issue defined as "Bayesian statistics" (e.g., Berger, 2000; Cumming, 2014); but a little common sense can be applied. If several prior experiments have suggested evidence that the instructional intervention "works" (i.e., in the above example, the results of the first experiment should have made it clear that it was more likely to work than not), then one can approach an additional experiment with some bias to assume that it works, or at least an equal footing. Of course, this guidance is only feasible to the extent that we can decide how big an effect is defined as "it works", and this requires some estimation of an effect size defined as "working". This procedure is the hallmark of statistical power analysis, and once an effect size is chosen (e.g., 50% of the variance accounted for), then it is possible to specifically compute the probability of a type 2 error in one's experiment, And it may then be possible to set alpha at that same level of probability, hence providing more equal odds for the two types of errors.

In summary, there are two general points to be made here: (1). The consumer of the research, who is the ultimate implementer of safety-critical procedures should be provided by the researcher with more data upon which to base her decision, than simply an "accept/reject" output of a statistical decision rule which implicitly or explicitly sends the message, in one case that "there is nothing there" (2) application of this decision rule, without adequate statistical power, provides an inherent bias against adopting safety improvement procedures or equipment.

What is to be done?

Below, I have outlined two general categories of remedies for this state of affairs; changes to how the researcher should approach experimental design and analysis, and changes to the way data are presented in written reports and articles.

Design and Analysis.

Increasing statistical power. It should be apparent that increasing statistical power, typically by running more subjects, which will reduce the estimated variance in effect size, will increase the confidence that a given effect is "true", and hence decrease the probability of a type 2 error. If this probability can be decreased to 0.05, then there is the desired symmetry between the two types of errors, and not the inherent bias against concluding safety-improving effects. Unfortunately however there are two issues that mitigate against such an increase in N in aviation human factors research (and research in other safety critical professions). First, it may be extremely difficult to obtain the participation of highly skilled professional workers in high fidelity simulation experiments and, with a restricted budget the researcher may well feel lucky to get even the 20 qualified line-pilots of our first example to find the time. Second, for reasons I have articulated at this symposium in 2009 (Wickens 2009), the very responses that may be most safety critical (and safety compromising) are those in which the pilot or controller might be most surprised (not expecting); the so called "black swan" event. This would be true of the unexpected first failure event used to estimate the success of training in our example above. Responses to a single, first failure event, of which there is by definition only one per pilot, do not afford the luxury of averaging across trials to reduce variance (and increase statistical power), and in contrast to so many other variables. Yet such first failure responses are unique in their ability to yield worst case response times and detection rates (Wickens Hooey et al., 2009). It is often these unexpected events that compromise safety.

Formulate an alternative hypothesis. An alternative hypothesis can be explicitly formulated to provide equal footing to the null hypothesis of "no effect". Statistical power calculations require this to be done. However these are typically based on deciding an effect size that is "meaningful" (e.g., 75% of the variance accounted for). But it is often more compelling to express this in meaningful performance units. In our example above, the researcher may state that under cases of possible spatial disorientation, in which timely diagnosis of an automation-induced upset is critical, any time-savings of greater than 3 seconds is *the* desired effect. Hence H1 = 3, while Ho = 0, and a savings of 2 sec would stand as more confirming of H1 than Ho. And it is always possible to "reserve judgment" for such intermediate effects. Similar "point estimates" of an alternative hypothesis might be made for the % improvement in performance that results from a particular display, or training innovation. In our field of aviation, where, because of its physical/spatial constraints, safety margins can so often be defined explicitly in terms of time and distance (separation), such alternative hypothesis point estimations are quite feasible.

Use "smart statistics" and planned contrasts. [and tolerate them if you are a reviewer].The data in figure 1 provide a case study of a typical experimental result. Two displays to aid mid-air collision avoidance are contrasted and both tested under low vs high workload conditions. Using conventional statistics, an ANOVA is performed and reveals a significant "p<.05" effect of workload, but a "NS p>.05" effect of both display type and its interaction with workload. Conventional statistics says "end of story". However what I have labeled "smart statistics" would make two points. First, the condition and effect you really care about is collision avoidance under **high workload** (potentially a worst-case, safety critical scenario). Hence what you should really do is to focus a **planned comparison** on the high workload condition. Indeed Cumming (2014) has effectively argued that going into an experiment, the investigator should **know** what s/he is looking for in terms of specific effects, and the omnibus F test ("is anything happening?") is a rather indirect way, of asking whether "something specific that you care about" is happening.

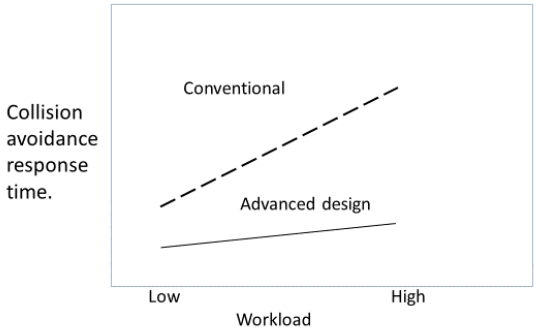


Figure 1. Hypothetical example of Results in a 2X2 experimental design

A second role of "smartness" in statistical testing, and particularly for most safety critical research, is that the comparisons should be 1 tailed, which provides more statistical power, rather than 2-tailed. What this essentially means is that you care if the effect is one that is safety-improving (i.e., typically shortens RT, improves accuracy or lowers workload). But you DON'T care whether there is no effect at all, or the effect works in the other direction. In either of the latter two cases, your proposed safety improvement is **not working**.

Presentation of experimental results.

Present the "raw data". It is worth highlighting again, the importance of presenting more, rather than less "raw data" to the readers of your article. By "raw data", we do not of course mean the individual subject measures, but we do mean graphs, 95% confidence intervals, effect size measures, and all statistical test measures (not just those of the magical "significant $p<.05$" type). The added relevance of this last guidance to meta-analyses will be described below. Here, if we think about statistics and stats packages in terms of the stages and levels of automation framework (Parasuraman et al., 2000), later stage automation is good if it is correct, but more problematic if it is in error of either the type 1 or type 2 variety. As we know, a stats package that simply tells you to accept or reject the null hypothesis is an example of late stage decision aiding automation, and, of course, can have a (typically) 5% chance of being in error. The best mitigation of this, in human-automation interaction research, is to let automation provide and convey to the reader more assistance in the earlier stage of integration and inference, and here, that specifically means providing graphed data and confidence intervals, along with the full array of inferential statistics.

Choose language carefully. Be very careful that the language you use in the text, does **not** convey the impression that effects which might be important for safety improvement but fail to reach the magic .05 levels are to be disregarded. The offenses here, ranked from worst to better for such a phrase to describe, say a .07 effect would be to say: "there was no effect", "not different", "not significantly different". Even if you **do** report the p values of such $p>.05$ effects in the results, time-limited readers of only a Discussion, or Abstract may not have taken the time to find those statistics. More plausible approaches (although here I have had to argue with editors) would be to label such effects as "marginally significant" or "approaching conventional levels of statistical significance" or even a "non-significant trend". It is equally important, when such effects are in evidence to describe in the text (and not just tables and graphs), their actual magnitudes, in terms such as the 4 second savings in response time, or the 30% gain in accuracy.

Accumulating evidence over experiments.

Earlier, we referred to "prior probabilities" for assuming that an effect might actually exist in the world, before we have seen the data from our current experiment. Of course the best source of such prior odds comes from other research on the topic, that may have used the same or similar variables to reveal the effect in question. Literature reviews can qualitatively summarize that research. But the ideal tool for this is the meta-analysis Rosenthal, 1991; Cumming, 2014; Onnasch et al., 2014). Various meta-analytic approaches can actually yield a quantitative estimate of the "collective wisdom" of that prior research, which may enable our researcher to not only express that an effect is likely to be there (or not), but also can give a point estimate of how large it is; that is, an explicit alternative hypothesis. The importance of meta-analyses has two implications: first, in a review of the literature, even an informal meta-analysis can establish the prior likelihood of finding the effect. Second, in helping others do their own meta-analyses, reporting the null effects that you do observe in your own data (or effect sizes of "NS $p>.05$" effects), you can provide data for the analyses of others that are not inherently biased toward more positive results.

Conclusions

In conclusion, we note that many of the concerns that have brought about the emergence of the .05 level of significance to avoid type 1 errors are well formulated, and we do not argue for a total abandonment of the tenets of NHST. However I argue that people should clearly understand the biasing implications of the black-white approach fostered by alpha levels, particularly when "the effect" that is examined (and is often rejected because of NHST) is one that has safety-improving implications. I hope that some of the remedies suggested above, can be adopted by the aviation psychology research community when they consider the decisions that the consumers of their work may make.

Dedicated to Tom Wickens: 1943-2012

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd Ed.). New York: Springer-Verlag

Cohen, J. (1988) *Statistical Power analysis for the behavioral sciences*. 2nd Edition. Hillsdale, N.J,: Erlbaum.

Cumming, G. (2014) The New Statistics: Why and How *Psychological Science*., *12,* 7-29.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Harris, D (1991). The importance of type 2 error in aviation safety research.In E. Farmer (Ed.) *Stress and Error in Aviation* (pp 151-157). Brookfield Vt.: Avebury.

Hubbard, R. & Bayarri, M. (2003). Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. *The American Statistician*. *57*, 171-188.

Neyman, J. & Pearson, E. (1933) On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. A.231 289-337.

Onnasch, L., Wickens, C., Li, H. & Manzey, D. (2014) Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. Human Factors. 56(3), 476–488.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30, 286-297.

Rosenthal, R. (1991). Meta-analytic procedures for social research (Rev ed.). Beverly Hills, CA: Sage.
Wickens, C.D. (1998) Commonsense Statistics. Ergonomics in Design. Oct. pp 18-22.

Wickens. C.D. (2009). The psychology of aviation surprise: an 8 year update regarding the noticing of black swans. In J, Flach & P. Tsang (eds). Proceeedngs 2009 Symposium on Aviation Psychology: Dayton Ohio: Wright State University

Wickens, C.D. Hooey, B. Gore, B.F., Sebok, A. & Koenicke, C. (2009) Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions. Human Factors. 51, 638-651.

FLIGHT OPERATIONAL QUALITY ASSURANCE (FOQA) – DO EXCEEDANCES TELL THE STORY?

Brian G. Dillman, Purdue University, West Lafayette, Indiana, USA
Dennis Wilt, Florida Institute of Technology, Melbourne, Florida, USA
Shawn Pruchnicki, Ohio State University, Columbus, OH, USA
Lukas Rudari, Purdue University, West Lafayette, Indiana, USA
Mark Ball, Purdue University, West Lafayette, Indiana, USA
Marshall Pomeroy, Ohio State University, Columbus, OH, USA

The concept of Flight Operational Quality Assurance (FOQA) programs have been widely utilized throughout the aviation industry. The premise behind the concept is to establish thresholds for flight situations based upon company operations specifications, regulatory guidance, aircraft limitations, and standard operating procedures and then monitor performance on the aggregate to determine if operations fall within acceptable boundaries. A singular exceedance may not trigger corrective actions, but if overall exceedances for the company exceed a predetermined acceptable threshold then mitigation strategies are employed to bring performance back within acceptable limits. Does this tell the whole story? Would allowing feedback for performance against idealistic targets be a better method for safety improvements? This paper will discuss the dynamics of investigative analysis of a project for stability of an approach and discuss the pros and cons of using systems that traditionally measure aggregate performance versus a system that determines degrees of performance.

FOQA Programs

The concept of a Flight Operations Quality Assurance (FOQA) program has roots in previous quantitative and qualitative aviation recording programs such as flight data recorders (FDRs), the Aviation Safety Action Program (ASAP), and the NASA Aviation Safety Reporting System. As indicated by program success at the airline level, a FOQA program should be accompanied by safety management systems (SMS) and a sound safety culture (Wiley 2007; FAA, 2006b). Airlines have realized much success from FOQA programs, and there have been recent efforts to bridge that success into the General Aviation sector, and despite the efforts of the FAA to expand FOQA, only 17% of smaller air operators have adopted it (Accardi, 2013).

FOQA is a significantly different program than all previous safety programs discussed. Unlike the ASRS and FAA Aviation Safety Action Programs (ASAPs), FOQA uses quantitative, objective data from flights to enhance trend monitoring and address operational risk issues (FAA, 2004; FSF, 1998). These operational risk issues, as discovered by FOQA data can lead to the development of more specific training programs such as Advanced Qualification Programs (AQPs). Historically only those on the flight deck during the flight know the true events of a given flight in relation to the flight data parameters collected. However, with the increased accessibility of FOQA type data across operations of all sizes, the aviation industry is still determining the best way to understand and utilize this rich data source.

The first workshop attempting to identify the benefits, utilization, and to encourage adoption worldwide of FOQA programs was by the Flight Safety Foundation (FSF) in Taiwan in 1989 (FSF, 1998). According to the Foundation (1998), their blueprint for FOQA has been the backbone for FOQA progress in the United States. However, this was only a starting point and there is potentially more work to be done in order to completely understand its full potential. The FAA took initiative to develop a formal FOQA program in 1990 by hosting a FSF workshop in Washington, DC, and in 2001 developed a rulemaking committee to further work in this area (FAA, 2003; FSF, 1998).

Before FOQA received full support from the FAA, a demonstration project was carried out to assess the costs, benefits, and safety enhancement associated with the program (FSF, 1998). During this project, the FAA provided hardware and software to four airlines which agreed to implement FOQA programs and share data with the FAA. As a result of the success of the project, the FAA determined that FOQA programs would be made voluntary as data collection and use for advanced FOQA programs were still in primitive form. The project demonstrated that the FOQA concept was a success for airlines by allowing enhanced trend monitoring and the identification of operational risks (FSF, 1998). The FAA did not attempt to create a FOQA program for non-commercial use during their three year demonstration project (FSF, 1998), although it is possible that a FOQA program for the general aviation sector would improve safety and operational performance in addition to assisting in flight training (Mitchell, Sholy, & Stolzer, 2007).

For those flight operations that plan to begin a FOQA program, a program development guideline is available in Advisory Circular 120-82, which discusses the benefits, set up, and maintenance of FOQA programs (FAA, 2004). This document also provides a template for the Implementation and Operations (I & O) plan set-up as well as key definitions that must be addressed during program establishment (FAA, 2004)

Airline officials, pilot union representatives and the FAA recognized that data protection issues were the biggest roadblock for FOQA program implementation (FSF, 1998). Initially, pilot unions were reluctant to sign FOQA agreements with airlines as they feared a lack of protection for collected FOQA data. FSF (1998) highlights three concerns airline pilot unions had with program implementation:

> "[first,] that the information may be used in enforcement/discipline actions; [second,] that such data in the possession of the federal government may be obtained by the public and the media through the provisions of FOIA; and [third] that the information may be obtained in civil litigation through the discovery process" (FSF, 1998, p. 7).

To address these concerns, 14 CFR Part 13 Section 13.401 was created. This document mandates FOQA data be stripped of any information that may identify the submitting airline before the data is passed to the FAA (FAA, 2004). The FAA ensures that "aggregate data that is provided to the FAA will be kept confidential and the identity of reporting pilots or airlines will remain anonymous as allowed by law" (FAA, 2004, p. 1). It is believed that relatively little exposure or experience with FOQA programs in any context will directly impact the perceptions of the individual within the flight program utilizing FOQA.

Traditional FOQA Data Analysis

With the data analysis focus of FOQA operations geared toward aggregate data, the natural inclination is to capture outliers from the normal operations rather than an analysis of the data to determine degrees of performance from a pre-determined objective. In a traditional FOQA program the system is set up with thresholds of measurements based upon one or more measures. For approach stability access these measures could be airspeed, vertical speed, roll rate, pitch rate, g-forces, or a combination of individual measures and the flight path angle. It is common for a FOQA program to establish "gates" along an approach flight path where flight parameters and aircraft configuration have to be within predetermined thresholds or a missed approach/go-around is warranted. If the aircraft goes beyond the boundaries of the flight path angle or exceeds the limits at an individual gate then an exceedance is recorded. The organization then follows up with a mitigation strategy to reduce the number of exceedances and continues to monitor the trends within the system. Of course, part of this understanding includes clarifying more contextually specific details that might offer a better interpretation of why these exceedances occurred. Upon looking at Figure 1a and 1b the framework of this system can be seen in a representation for approaches to an example runway. Looking at the blue line that

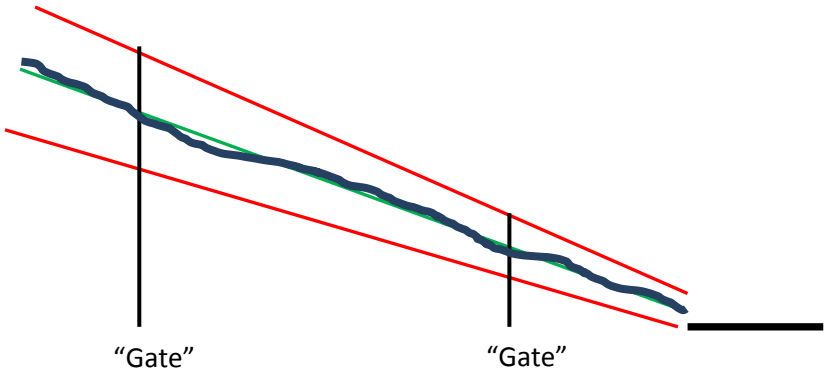


Figure 1a. FPA Analysis

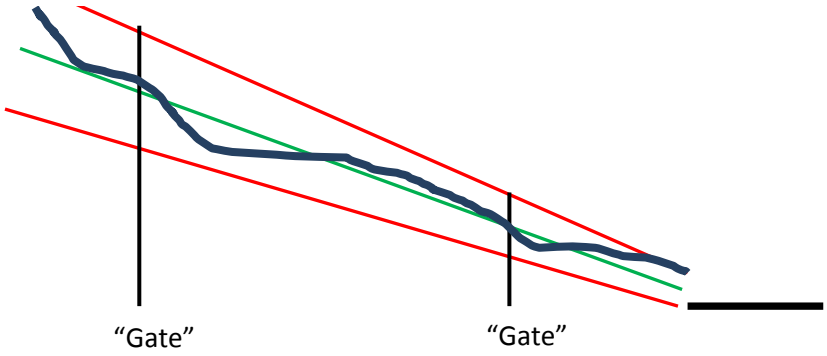


Figure 1b. FPA Analysis

represents the flight path it can be seen that Figure 1a stays relatively close to the center line and

Figure 1b varies along the flight path but never exceeds the outer boundaries. If the aircraft had met the criteria at the given "gates" then the FOQA system might not have recorded either approach as an exceedance even though the aircraft in Figure 1a could be considered more stable.

Alternative FOQA Data Analysis

Because of the aforementioned limitations to a more strictly exceedance based approach, an alternative method for FOQA approach analysis is to measure the Flight Path Angle (FPA) at 1 second intervals along the approach, calculate the absolute value of the difference between a given second and its subsequent second value, and then the sum of the variations in the FPAs for the last 30 seconds could be calculated. An approach that maintained a perfectly consistent FPA would have no difference in the FPAs at each second interval and then the sum of the variations in the FPAs for the last 30 seconds would equal zero. An approach that had a lot of variation would end up with a larger sum of the variation in the FPAs for the last 30 seconds. It is this measure that could then be used to determine the stability of an approach path. Therefore, this technique will give an overall dynamic view of an approach trend, rather than a measure of FPA boundaries. Aircraft that exhibit high variations in FPA will be considered less stable and will warrant further review just as the approaches that trigger an exceedance require further inquiry. This type of analysis also allows comparisons of all approaches, and can identify trends for operational improvements.

A system similar to the one described above was developed at the NASA Ames Research Center is the Aviation Performance Measuring System (APMS) (Chidester, 2003). According to Chidester (2003), the mission of APMS has three major thrusts; moving beyond exceedance-detection to routine analysis of all the data, providing focused analysis of higher risk phases of flight, and mining the data for atypical, potential precursors of incidents and accidents. The major movement from APMS is a shift from waiting until an aircraft operates outside of established parameters (exceedance) and recording it, seeing if there is a trend in recorded exceedances, and then identifying if it's a systemic problem in the operation or isolated to a given airport or aircraft. The system works by analyzing data and grouping operations into "normal" and "outliers". If the preponderance of operations to a given airport all look the same it is assumed that the operation is normal and therefore safe. If an operation is grouped outside of the typical performance parameter then it is flagged for follow-up by an aviation safety professional. This prevents needless oversight and focuses efforts to the operations that have a higher likelihood of needing analysis. In 2004 APMS was put to into action and was able to take more than 16,000 flights over a two year period and narrow it down to the most statistically extreme 5% of the dataset (Chidester, 2004). These flights were further analyzed and it was found that they fell into 8 different categories; high-energy arrivals, turbulence and accommodation, go-arounds, landing rollout anomalies, atypical climbs, takeoff anomalies, TCAS resolution advisories with escape maneuvers, unusual arrival paths (Chidester, 2004). This is a significant step forward in data analysis for FOQA efforts but there is still room for improvement. Of the 95% of the flights that fell into the "normal" operation category there is still variability from the target or ideal operation. Even though the parameters evaluated in the APMS system tend to fall around the mean for each parameter, the approaches conducted are still measured at "gates" or intervals at 1500, 1000, 500, and 100 feet above the runway

(Chidester, 2003). By measuring the change in performance at subsequent intervals the variability or change in performance can be determined which could be a better indicator of operational stability for certain types of procedures.

Conclusion

FOQA programs continue to evolve and are becoming more robust as the technology affords opportunities to analyze data in different ways. A significant barrier to consider moving past traditional FOQA analysis is the desire to group flight operations into a binary mode. "Stable" versus "unstable" categorizations or "normal" versus "extreme" can be replaced by measurements from targeted objectives. It will be necessary to avoid the temptation to label operations that fall within a "stable" category but are determined to have room for improvement to be considered "unstable". The focus of every professional pilot should be to want to improve their performance beyond what has already been achieved. Professionals should also want to identify when their performance deviates from the target beyond what they normally achieve. The individual performance may still be considered acceptable and within normal parameters, but these types of evaluations can assist pilots in determining reasons for why their performance is decreasing. Hindrances to performance such as fatigue, recency of experience, aircraft familiarity, and environmental conditions could all be qualified as to how they affect a particular pilot from their target objective. These types of measurements could provide robust feedback as a step toward process improvement.

References

Accardi, T. (2013). *Public Sector Pilot Perceptions of Flight Operational Quality Assurance Programs* (Doctoral dissertation, Oklahoma State University).

Chidester, T. (2003) Understanding Normal and Atypical Operations Through Analysis of Flight Data. *Proceedings of 12th International Symposium on Aviation Psychology*. Dayton, OH.

Chidester, T. (2004) Example Application of The Aviation Performance Measuring System (APMS). *Retrieved from http://flightsafety.org/files/APMS_application.pdf*.

Federal Aviation Administration. (2006b). *Introduction to safety management systems for air operators.* (DOT Advisory Circular No. 120-92). Washington, DC: U.S. Government Printing Office.

Federal Aviation Administration. (2004). *Flight operational quality assurance.* (DOT Advisory Circular No. 120-82). Washington, DC: U.S. Government Printing Office.

Federal Aviation Administration. (2003). *Flight operational quality assurance aviation rulemaking committee*. (DOT Order No. 1110.131A). Washington, DC: U.S. Government Printing Office.

Flight Safety Foundation. (1998). Aviation safety: U.S. efforts to implement flight operational quality assurance programs. *Flight Safety Digest, 17*(7-9), 1-54.

Mitchell, K. and Sholy, B. and Stolzer, A.J. (2007) "General Aviation Aircraft Flight Operations Quality Assurance: Overcoming the Obstacles". *Aerospace and Electronic Systems Magazine, IEEE*. Vol. 22, No. 6, Pgs. 9-15.

Wiley, J. (2007, June). C-FOQA: Has its time arrived? *Business & Commercial Aviation, 100*(6), 76-82.

UN-ALERTED SMOKE AND FIRE: CHECKLIST CONTENT AND INTENDED CREW RESPONSE

Barbara K. Burian
NASA Ames Research Center
Moffett Field, CA

An in-flight smoke or fire event is an emergency unlike almost any other. The early cues for un-alerted conditions, such as air conditioning smoke or fire, are often ambiguous and elusive. The checklists crews use for these conditions must help them respond quickly and effectively and must guide their decisions. Ten years ago an industry committee developed a template to guide the content of Part 121 checklists for un-alerted smoke and fire events. This template is based upon a new philosophy about how crews should use the checklists and respond to the events. To determine the degree to which current un-alerted checklists of in-flight smoke or fire comply or are consistent with the guidance outlined in the template, I collected and analysed checklists from North American air carriers.

In-flight smoke, fire, and fumes (SFF) events, particularly those that are un-alerted, such as electrical smoke or fire, are among the most critical emergencies faced on-board aircraft.[1] Timely response by the flight deck and cabin crews is essential (Federal Aviation Administration [FAA], 2014). During these events, crews must divide their attention among and accomplish a wide variety of tasks. When necessary, they must protect their ability to respond and continue to fly the aircraft by donning oxygen masks and goggles and must establish and maintain good communication and must coordinate their responses. The source of the SFF must be identified and appropriate actions taken to isolate and eliminate it, if possible. At the same time, however, crews must be recognize the possible need to divert and make an emergency landing at a suitable airport, or even ditch, should it be necessary. Thus, during smoke and fire identification and elimination activities, crews must maintain the "big picture" view of their dynamic situation and be prepared, if necessary, to shift from attempting to identify and quell the source of the fire to eliminating dense smoke and fumes and preparing to land or ditch (Burian, 2005).

Approximately 10 years ago, representatives from the aviation industry constructed a template to guide the development of checklists for un-alerted SFF events (Flight Safety Foundation [FSF], 2005). The underlying philosophy for checklist content and crew response to un-alerted SFF events embodied in this template contrasted markedly with that underpinning typical checklist content and crew response expected up to that time. Previously, Quick Reference Handbooks (QRHs) contained multiple checklists for a variety of un-alerted SFF events (e.g., air conditioning smoke, electrical smoke and fire, etc.) and crews had to first determine the type of SFF before being able to access a checklist. Additionally, it was typical that reminders or directions to complete smoke removal actions, if necessary, or divert to an airport for an emergency landing appeared only at the end of these checklists, if they appeared at all, after *all* source identification and elimination actions had been accomplished. Reports of crews being unable to determine the correct type of SFF, and thus, accomplishing the wrong checklist for their situation, and SFF incidents and accidents in which a diversion was delayed (e.g., National Transportation Safety Board [NTSB], 1998; Transportation Safety Board [TSB] of Canada, 2003) spurred the industry to re-think checklist content and the underlying philosophy of how crews should respond to these events.

Four major concepts or features are represented in the checklist template that was developed (see Table 1). The first is that all actions and information necessary for response to *all* different types of un-alerted SFF events are to be integrated into a single checklist. Thus, crews will not have to first make a determination of the type of event they are dealing with before being able to identify the correct checklist to access. The second major concept is that crews should be guided to take quick action to isolate and eliminate the most likely source(s) of SFF, as determined through historical analysis of SFF on each aircraft type, *without* first being required to determine if those likely sources are indeed the source of their SFF situation; these are referred to as "Manufacturer's initial steps" and "Remaining minimal essential manufacturer's initial steps" in the template (Steps 5 and 9, respectively; see Table 1). The third major feature is that checklists constructed according to the template guidance will be appropriate for

[1] It can be difficult to develop consistently correct and reliable alerts for some types of SFF conditions, such as the presence of toxic fumes, and electrical and air conditioning smoke and fire. As a consequence, typically no alerts for these conditions are provided through integrated aircraft crew alerting systems.

use in situations in which the source of SFF is obvious, easily accessible, and can be extinguished quickly (Steps 6-8 in Table 1) as well as for those situations in which it is not (see system specific actions, Steps 12-14 in Table 1). The final major concept incorporated in the template is that reminders or guidance to the crews regarding their overall situation management should be provided, particularly with regard to conducting a diversion (Steps 1 and 10), performing an immediate landing (the Warning after Step 10 and Step 15), and the necessity to consider accomplishing smoke/fumes removal actions (after Step 5 and Step 17). When under the stress and high workload characteristic of these critical emergencies, it can be easy to lose sight of the need to manage the overall situation; these reminders are included in the template to counteract this tendency.

Table 1.
Smoke/Fire/Fumes Checklist Template[1,2]

Steps	Action
1	Diversion may be required.
2-4	Crew protection (don masks and goggles) and establish crew communication
5	Manufacturers initial steps
Accomplish Smoke/Fumes Removal Checklist any time they become the greater threat	
6-8	Extinguish source if it is immediately obvious and can be extinguished quickly
9	Remaining minimal essential manufacturer's initial steps
10	Initiate a diversion to the nearest suitable airport while continuing the checklist
Warning: If the SFF situations becomes unmanageable, consider an immediate landing	
11	If landing is imminent, go to Step 16. If not, go to Step 12.
12-14	System specific actions (e.g., air conditioning smoke, electrical smoke and fire, etc.)
15	If SFF continues after all system specific actions are accomplished, consider landing immediately.
16	Review Operational Considerations (e.g., overweight landing, etc.)
17	Accomplish Smoke/Fumes Removal Checklist, if required.

[1] Steps in the template have been condensed and wording has been minimally altered to save space. Refer to FSF (2005) for the complete wording and expanded view of all checklist steps.
[2] More than one step or action in the actual SFF checklists that are developed may be included as part of a single step on the template.

It has now been a decade since the template was developed. This study was undertaken to identify the degree to which a sample set of airline checklists currently in use for in-flight, un-alerted SFF conform to the guidance and underlying philosophy of the industry's SFF checklist template (FSF, 2005).

Method

Participants and Materials

Seven North American air carriers (including international, major, and regional carriers) provided 11 QRHs containing the checklists analyzed in this study. The QRHs[2] were used on five aircraft types: Airbus A320 (n=3); Boeing B737NG (n=2); Boeing B777 (n=2); Bombardier Canadair Regional Jet-700 (CRJ700, n=2); and Embraer E190 (n=2). Un-alerted checklists for in-flight SFF events as well as checklists to be used for the removal of smoke and toxic fumes were analyzed. Additional checklists pertaining to passenger evacuation, ditching, and emergency landing/descent were also reviewed. All QRHs and checklists were current and in use by the participant air carriers at the time they were provided to the researcher.

Procedure

These checklists were analyzed for the degree to which their structure, content, and implicit philosophy of crew response to un-alerted, in-flight SFF were consistent with or deviated from guidance in the industry SFF checklist template (FSF, 2005). Particular attention was paid to the four major concepts or features of the template

[2] One carrier which flies the Boeing B777 does not use a printed QRH on-board the aircraft; only the B777 Electronic Checklist (ECL) is used for response to non-normal events, such as in-flight smoke, fire, and fumes. Electronic copies of the checklists contained in this air carrier's electronic Operations Manual were analyzed for this study. For ease of wording, the electronic Operations Manual will be referred to as a QRH.

described earlier. Font sizes, inclusion of memory items, checklist length and numbers of items for diverse SFF scenarios, and reference to items pertaining to other checklists that might be needed in these events (e.g., evacuation, ditching, etc.), were also analyzed but are not reported here due to space limitations.

Results

Template Concept 1: A Single Integrated Checklist

At the time the template was developed the concept of a single, integrated checklist to be used for response to all types of un-alerted SFF events was relatively novel. It was not uncommon to see separate checklists for un-alerted SFF events occurring in specific locations or involving different aircraft systems: air conditioning, electrical, cabin, galley, lavatory, avionics, engine tailpipe, cargo, and unknown source or hidden. On the aircraft types included in this study, SFF involving avionics, cargo, or occurring in lavatories are now most often alerted through flight deck caution and warning systems (i.e., Engine Indication and Crew Alerting System [EICAS], or Electronic Centralized Aircraft Monitoring [ECAM]). With the exception of engine tailpipe fire[3], of the 11 QRHs analyzed in this study, the integration of actions for response to the SFF types that remain un-alerted was seen in 10; one of the three A320 QRHs analyzed did not have a main integrated SFF checklists and included separate checklists for a) cabin smoke and fire and b) air conditioning smoke or fire. The other two A320 QRHs, as well as the QRHs for the other four aircraft types, integrated response to these un-alerted conditions along with others, such as electrical smoke or fire, into a single checklist. Furthermore, the other two A320 un-alerted SFF checklists included items for *alerted* avionics SFF. Similarly, both EMB190 un-alerted SFF checklists included items for *alerted* cargo compartment fires. Thus, based upon this small sample of QRHs and checklists currently in use, it appears that the concept of providing a single, integrated checklist for many types of un-alerted SFFs, and on occasion some types of alerted SFF, has gained some acceptance within the industry.

However, some separate un-alerted SFF checklists were identified in a few of the 10 QRHs that also contained a main integrated checklist for un-alerted SFF (e.g., Aft Avionics Rack Smoke, n=1; EFB Computer Overheat/Fire, n=2; and Tailpipe Fire, n = 5). With the exception of Tailpipe Fire (discussed below), it is not known why the air carriers or manufacturers who developed these checklists chose to keep them separate and not incorporate them into the main integrated checklist for un-alerted SFF.

Template Concept 2: Eliminate the Most Likely Sources of SFF without Analysis

One of the most novel concepts in the template suggests flight crews should isolate and eliminate the most likely sources of un-alerted SFF *without* first determining if they are in fact the cause. All 10 QRHs that included a single integrated checklist for most types of un-alerted SFF included these types of steps—referred to in the template as initial manufacturers steps (range: 1-10 items, mean = 5.6 items). According to the supplementary information provided with the template (FSF, 2005), these initial steps or actions should be "quick, simple, and reversible; will not make the situation worse or inhibit further assessment of the situation; and do not require analysis by the crew" (pg. 32). Additional manufacturer's steps which do not require crew analysis but may not meet the other criteria for the "initial steps" just outlined (see Step 9, Table 1) and which are distinctly separate from system specific actions (see Steps 12-14, Table 1) were found only in the four checklists for Boeing aircraft (B737: n= 11 items in each checklist, and B777: n=4 items in each checklist) . Thus, in the checklists analyzed there is high consensus on directing crews to extinguish likely sources of SFF without going through a lengthy process to confirm which source might actually be causing the event. Of the four manufacturers, only Boeing chose to provide additional actions, separate from system specific items, that might not be reversible, might inhibit further assessment of the situation, or in some other way do not meet the criteria spelled out for "initial manufacturer's steps" (FSF, 2005).

Action reversal. All four of the Boeing checklists and one of the EMB190 checklists included an item stating that at the captain's discretion, actions just performed (i.e., manufacturer's initial steps or elimination of an obvious and quickly extinguishable source) could be reversed if the SFF could be confirmed to have been extinguished and the smoke/fumes was dissipating. One EMB190 and two A320 un-alerted SFF checklists instruct

[3] Due to the unique nature of tailpipe fires and the inability of flight and cabin crews to directly fight and/or confirm that tailpipe fires have been extinguished, actions for addressing these fires are not integrated into checklists for dealing with other types of un-alerted SFF. Five of the 11 QRHs analyzed included separate checklists for dealing with tailpipe fires.

the pilots to reverse some steps to re-power some equipment needed during landing while on final approach. Additionally, two checklists (one for an A320 and one for an EMB190) also gave pilots the option or directed them to reset the Display Units to Auto if they were required for landing but did not state when this should occur.

Template Concept 3: Dealing with Sources that are Obvious and Quickly Extinguishable or are Not

Of the 10 integrated SFF checklists analyzed, nine contained actions to accomplish associated with sources that were obvious and accessible and could be extinguished quickly (Steps 6-8, Table 1); these actions appear after the completion of the manufacturer's initial steps (Step 5, Table 1). The QRH that did not (for a CRJ700) took a different approach than that suggested by the template. In this checklist, after completing the manufacturer's initial steps, if the source was known (e.g., electrical smoke or fire) crews were directed to complete that (system specific) section of the checklist. If the source was not known upon completion of the manufacturer's initial steps, pilots were directed to attend to diversion and landing activities and were not to complete any system specific actions at all. In the template, and the nine integrated checklists that adopted the template's approach, if the source is obvious and quickly extinguishable, crews are directed to do so but are not provided with specific actions to accomplish. They are only directed to complete system specific actions (Steps 12-14, Table 1) if the source is not immediately obvious or if attempts to extinguish it have been unsuccessful.

Furthermore, the template guides the accomplishment of actions for *additional* aircraft systems if those accomplished in the *first* system specific section are not successful in isolating and extinguishing the SFF. Hence, pilots would start with actions for the first system (typically the one identified as most often the source of SFF on that aircraft type) and continue accomplishing items through that system and subsequent system sections until the source has been eliminated or the end of the checklist has been reached. All 10 of the integrated SFF checklists analyzed contained system specific items, although in the two CRJ700 checklists, pilots were directed to accomplish *only* the items for the specific system thought to be the source of the SFF. In other words, in those checklists, pilots were not directed to accomplish items associated with any other systems, even if those accomplished associated with the suspected system had been unsuccessful in terminating the SFF. It should be noted that in six checklists, although system specific items were provided, which system they pertained to was not specified (i.e., through a header or section title) and in some checklists it appeared that all or most items for some systems (e.g., electrical smoke/fire) were included as part of the initial manufacturer's items near the beginning of the checklist. Furthermore, analysis of the checklists in one CRJ700 QRH revealed that actions for dealing with air conditioning smoke or fire were not located in the integrated SFF checklist but instead were included in the checklist for Smoke and Fumes removal.

Template Concept 4: Support for Overall Situation Management

Diversion. The template includes two items with regard to a diversion. The first is Step 1 and is intended to be a reminder to the crew or "establishes the mindset" (pg. 34) that a diversion may be necessary (FSF, 2005). The second (Step 10) actually directs that a diversion to the nearest suitable airport should be initiated while continuing with the rest of the checklist. This action is reached if the initial manufacturer's actions have proved unsuccessful and if the source is not immediately obvious or is immediately obvious but cannot be visually confirmed to have been extinguished. Thus, the diversion is directed after some steps have been taken quickly—that have proved unsuccessful—and prior to the pilots accomplishing more "analytical" actions in the system specific sections. In this study checklist analysis distinguished items worded as reminders (e.g., "Consider a diversion"), consistent with the intent of template Step 1, from items that directed that a diversion be initiated/conducted, consistent with template Step 10 (see Table 1).

The degree to which the checklists analyzed conformed with these two template steps varied greatly although all of the main un-alerted SFF checklists analyzed did address diversion, most often going beyond what is suggested by the template. Five checklists included some type of reminder that diversion may be necessary at or near the beginning of the checklist and six checklists made such reminders (three of them for the second time) in the middle of the checklist. Three checklists actually direct the initiation of a diversion (or stated "Land Immediately/ ASAP") at or near the beginning of the checklist and the other seven direct the initiation of a diversion in the middle of the checklist, similar to placement of this direction in the template (Step 10). One CRJ700 checklist actually contained nine separate places where a diversion was directed, many of these occurring at the end of sets of items to be completed in the system specific sections. Thus, with regard to the intent of the template relative to reminding or directing a diversion, the checklists analyzed in this study conformed and even went beyond by often directing a diversion much earlier during situation response than suggested by the template.

Landing is Imminent. Step 11 in the template is what is known as an "opt out gate" (Burian, 2014). An opt out gate is a Conditional/Decision Item that, if true, directs the user to abandon checklist accomplishment and to shift attention to some tasks other than accomplishment of the checklist; in this case, the user jumps to the final items on the checklist in preparation for an impending landing. Such items can be critical in helping to assure that pilot attention is not fixated on checklist accomplishment but rather is focused on the most essential tasks relative to aircraft condition and phase of flight.

Despite the importance of such an item, none of the checklists analyzed in this study incorporated this item as stated in the template. However, approximately half way through a 4½ page checklist for use in an EMB190, there were directions about what actions to take if an airport was nearby. Additionally, four checklists (all for Boeing aircraft) stated that diversion/landing should not be delayed in an attempt to complete the following (system specific) items. In contrast, the checklist for a CRJ700 instructed pilots to complete as many items as possible before completing the Descent and Before Landing checklists. In each of these six checklists, the manufacturers or air carriers attempted to address the distribution of attention given to fighting the fire relative to diversion/landing. However, this was accomplished in three different ways, only one of which (for the EMB190) was close to matching the guidance as stated in the template. This is explored in the Discussion section below.

Consider an Immediate Landing. There are two steps in the template in which the pilots are told to consider an immediate landing: if the SFF situation has become unmanageable and if all the appropriate actions to isolate and eliminate the SFF have been accomplished but were unsuccessful. Seven of the checklists include the first item although none of them worded it as a Warning Statement as it appears in the template. The three other un-alerted SFF checklists directed the pilots to accomplish a diversion or land ASAP/immediately early in the checklist, obviating the need to suggest later in the checklist that an immediate landing be considered. Only five of the integrated checklists include the suggestion to consider an immediate landing if all SFF elimination actions have failed. However, three of the other five integrated checklists instructed their pilots to land as soon as possible at or near the beginning of those checklists; hence, a later suggestion to consider an immediate landing was unnecessary.

Smoke/Fumes Removal. Only one of the ten integrated SFF checklists failed to include some reference to the completion of the Smoke/Fumes Removal Checklist, if necessary, and that checklist (for a CRJ700) appeared to include actions for smoke removal in the main integrated SFF checklist. Where reference to the possible need to complete the Smoke/Fumes Removal Checklist appeared in the main SFF checklist varied, with some checklists containing more than one reminder: 3 near the beginning of the SFF checklist, 6 in the middle before the completion of system specific actions (similar to its first location in the template), and 8 at the end of the checklist (similar to its second location in the template).

Discussion

This study reveals general compliance with the main guidance put forth in the industry-developed template for un-alerted SFF, although this was by no means a perfect match. As it is just that—guidance—some deviations are to be expected. Thus, the underlying philosophy with regard to checklist content and crew response to these events was adopted in the checklists examined with just a few notable exceptions. Ten of the 11 QRHs analyzed provided an integrated checklist to be used for a variety of types of un-alerted SFF situations. However, some un-alerted checklists that appeared on their face to be appropriate for integration (e.g., EFB Overheat) were not integrated. It is possible that the failure to include actions for these un-alerted SFF events into the main integrated checklist was an oversight. However, it may also have been intentional, thinking that the sources for these events were easily identifiable and warranted a different approach to isolation and extinguishing than that put forth by the template. Checklist developers will need to carefully weigh the pros and cons when deciding that new un-alerted SFF checklists should remain separate (and not be integrated) lest as some point in the future, pilots again find themselves with a long list of un-alerted SFF checklists that must be searched through when looking for the correct one.

Another area where some deviation from the template was observed pertained to diversion and landing guidance. Rather than just a reminder that such a diversion might be necessary at the beginning of the checklist, as suggested by the template, three of the checklists analyzed directed that a diversion or landing be initiated right away. In all of the checklists analyzed developers appear to have embraced the idea of suggesting or even directing a diversion, early on during event response, or at least before extensive source identification actions are undertaken.

Extrapolating from the small sample of checklists analyzed in this study, it appears that gone is the day when checklists guide numerous actions to identify, isolate, and eliminate the source of fire before recommending that pilots conduct an immediate diversion or landing.

Related to this, but more concerning, however, is that, at best, only one of the checklists analyzed fully adopted the template's suggested approach for dealing with an imminent landing. During checklist design it is difficult to determine where to put an item that relates to an external event (imminent landing) which might occur at any time during system response and checklist accomplishment. Rather than, at a specific point in the checklist, calling for an assessment of phase of flight and directing which items to bypass if landing is imminent, as per the template guidance, the four Boeing checklists instructed the pilots to continue to complete checklist items but to not delay the descent and landing. While allowing for greater checklist flexibility for use in a wide range of situations occurring at a variety of phases of flight, it places the onus on the pilots for keeping the big picture in mind while also focusing narrowly on checklist accomplishment and deciding where in the procedures to break off relative to landing—two demands on pilot situation awareness and cognition meant to be alleviated through the directed evaluation and checklist accomplishment suspension incorporated in the template. Of far greater concern was the direction in one of the CRJ700 checklists that as many items on the SFF checklist should be accomplished as possible before turning attention to the completion of Descent and Before Landing checklists. Such guidance could actually have the effect of delaying the descent and landing in an effort to complete the SFF checklist; something that is completely opposite from that intended by the template guidance.

Acknowledgments

This work was conducted under the NASA Aeronautics Research Mission Directorate. Deep appreciation is extended to the air carriers who participated in the study through the provision of checklists and quick reference handbooks and to Key Dismukes, Janeen Kochan, and Mary Connors who reviewed earlier versions of this paper.

References

Burian, B. K. (2005). Do you smell smoke? Issues in the design and content of checklists for smoke, fire, and fumes. *Proceedings: International Society of Air Safety Investigators 2005 Conference*. Fort Worth, TX: ISASI.

Burian, B. K. (2014). Factors affecting the use of emergency and abnormal checklists: Implications for current and NextGen operations. NASA Technical Memorandum. NASA/TM—2014-218382.

Federal Aviation Administration (2014). In-Flight Fires. *Advisory Circular 120-80A*. http://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-80A.pdf. Accessed 12/27/14.

Flight Safety Foundation. (June, 2005). Flight crew procedures streamlined for smoke/fire/fumes. Flight Safety Digest. http://flightsafety.org/fsd/fsd_june05.pdf. Accessed 10/12/14.

National Transportation Safety Board (1998). *Aircraft Accident Report – In-flight Fire/Emergency Landing, Federal Express Flight 1406, Douglas DC-10-10, N68055, Newburgh, New York, September 5, 1996.* Report Number NTSB AAR-98/03. Washington, DC: NTSB.

Transportation Safety Board of Canada (2003). *Aviation Investigation Report A98H0003, In-flight Fire Leading to Collision with Water, Swissair Transport Limited McDonnell Douglas MD-11 HB-IWF, Peggy's Cove, Nova Scotia 5nm SW, 2 September 1998.* Gatineau, Quebec, Canada: TSB of Canada.

RELIABILITY OF INSTRUCTOR PILOTS' NON-TECHNICAL SKILLS RATINGS

Patrick Gontar
Institute of Ergonomics, Technische Universität München
Munich, Germany
Hans-Juergen Hoermann
Institute of Aerospace Medicine, German Aerospace Center (DLR)
Hamburg, Germany

This paper presents the results of different methods to assess reliability when instructor pilots rate pilots regarding their non-technical skills (NOTECHS). In preparation for a major inter-rater reliability study, this pretest analyzes the rating behavior of two instructor pilots during a full-flight simulator mission. Besides inter-rater reliability and test-retest reliability, the pilots' self-rating (n =12) and the instructors' point of view is analyzed. Results indicate a wide spread from poor to excellent reliabilities as a function of the different rating dimensions. Regarding inter-rater reliability, it is found that non-technical skills are rated more reliably under high workload conditions than under low workload conditions, and social aspects of non-technical skills are rated more reliably than cognitive aspects. Test-retest reliability is found to be .6 on average, whereas self-rating / instructor rating reliability is .5 on average. Based on these findings, implications for the major inter-rater reliability study will be derived and incorporated.

The importance of effective Crew Resource Management (CRM) has been known since the late 1970s, when NASA held their workshop on "Resource Management on the Flightdeck", and came to the conclusion that a majority of accidents are directly linked to interpersonal skills (Helmreich, Merritt, & Wilhelm, 1999; Dietrich, 2004; Gontar, Hoermann, Deischl, & Haslbeck, 2014). Consequently, adequate training methods and corresponding evaluation metrics were developed (O'Connor, Hoermann, Flin, Lodge, & Goeters, 2002). Although huge efforts were undertaken to train the raters, inter-rater reliability (IRR) is still an issue to be discussed. For example, Flin and Martin (2001), Law and Sherman (1995), Law and Wilhelm (1995), and Seamster, Edens, and Holt (1995) found different influencing factors that result in reduced inter-rater reliability in the aviation context. Sevdalis et al. (2008) and Yule et al. (2008) showed similar reduced reliabilities within the medical domain. In current airline practice, the trainer has to operate the simulator, simulate the air traffic controller, and assess the pilots during the mission – all at the same time. These circumstances make it worth analyzing the current evaluation practice in an airline to develop general recommendations to improve reliability of CRM ratings during training.

Background

The study presented here serves as a pretest in preparation for a study that aims to investigate the IRR of the most experienced instructor pilots (n = 45) when rating pilots' CRM skills within a major German airline. The goal of this pretest is to validate a flight scenario regarding general feasibility and its appropriateness for CRM ratings by instructor pilots. In addition to the instructor pilots' input, the self-assessment of the participating pilots will be taken into account as well. The findings will provide a rough estimation of the different reliabilities so the test design for the main study can be developed.

Research Questions

Based on the previous literature and the mentioned motivation for this study, the research questions (RQs) aiming for inter-rater reliability and test-retest reliability can be formulated as follows:

RQ 1 – Rating while operating: How reliable can two instructor pilots rate airline pilots' non-technical performance while operating a full flight simulator?

RQ 2.1 – Retest rating based on video recordings: How reliable can a rater asses pilots' performance based on a video recording?

RQ 2.2 – Self rating vs. instructor ratings: Which rating of the instructor (simulator or video-based) better reflects pilots' self-perception?

Method

Operationalization

In order to answer these questions, a full flight simulator scenario seems appropriate in order to have the same realistic environment as during normal simulator training missions. Further requirements are: captain and first officer as participants, no confederate pilot, realistic air traffic controller communication and noise (Schubert & Haslbeck, 2014), realistic unforeseen scenario (Casner, Geven, & Williams, 2013) with appropriate malfunctions to measure the influence of workload. Furthermore, the participants shall not be recruited on a volunteer basis but randomly selected to exclude any self-selection bias (Rosenthal & Rosnow, 2008).

To rate the pilots' CRM skills, it is important that the raters are already familiar with the evaluation tool. In this case, we use the company-adopted evaluation form, which uses the NOTECHS method (O'Connor et al., 2002). It was adapted to the airline's philosophy (Burger, Neb, & Hoermann, 2003) and is known to the two instructor pilots as well as to all participating pilots (important in terms of self-evaluation). The four dimensions measured on a five-point scale are defined as: *Communication, Leadership & Teamwork, Work Organization,* and *Situation Awareness & Decision Making*. In order to evaluate pilots' procedural and more technical skills, the Line Operations Safety Audit (LOSA) *Descent / Approach / Land* sheet (Klinect, Murray, Merritt, & Helmreich, 2003) is appropriate and measures *Planning, Execution, Review & Modify,* and *Overall Behavioral Markers* on a four-point scale. In contrast to the internal company evaluation form, the instructor pilots did not work with this LOSA sheet before.

When it comes to reliability measurements, one will find a lot of different metrics that can be computed. In the domain of evaluating non-technical skills ratings, intraclass correlation coefficients (ICCs) are commonly used as a measurement of reliability (Shrout & Fleiss, 1979). To assess systematic differences in the mean values, the ICC model can be adjusted to take those differences into account, called *absolute agreement.* Since both raters will rate every participant, a two-way random model can be applied (Wirtz & Caspar, 2002). The values of ICC can range between 0 and 1, where 1 represents fully explained variance; Landis and Koch (1977) postulated that values between .41 and .60 are moderate, values greater than .61 are substantial and values above .8 are almost perfect for kappa statistics. Fleiss, Levin, and Paik (2003, p. 604) stated that values "greater than 0.75 or so may be taken to represent excellent agreement". Current practice sets the cut-off value of reliability values for CRM ratings to .7 (Yule et al., 2008), as does this paper. To enhance reliability, it is possible to calculate the mean of two or more raters. With the help of the Spearman-Brown prophecy formula, the reliability of the average of *m* different raters can be calculated when the reliability of a single rater is known (Lienert & Raatz, 1994).

Test Design

Based on the research questions stated above, the test design can be visualized as shown in Figure 1. Both raters (rater 1 and rater 2), in this case the flight instructors, rate *n* subjects (pilots) during a full-flight simulator mission, while simultaneously operating the simulator and acting as air traffic controllers. In addition, all subjects rate themselves regarding their CRM performance (Gontar & Hoermann, 2014).

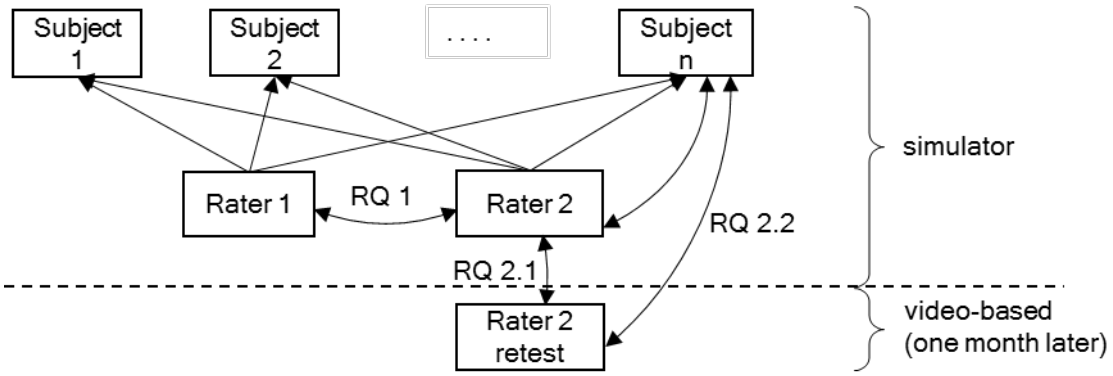


Figure 1. Test-design visualization.

The reliability between the two raters can be seen as inter-rater reliability (RQ 1). Four weeks after all simulator missions were completed, one of the raters (rater 2) assessed the performance of n = 8 pilots based on video and audio recordings again (retest). The comparison between the two ratings of rater 2 (simulation vs. video-based) is considered as test-retest reliability (RQ 2.1). The accordance between the self-ratings on the one side and the two time-delimited ratings of rater 2 are interpreted as the degree of accordance between self and external perception (RQ 2.2).

Experiment

In order to assess instructors' performance when assessing pilots' non-technical skills, it is important that a broad performance variance is shown by the participants. Within a full-flight simulator experiment, this means to induce a rather high workload with the help of an adequate scenario and malfunctions. The experimental scenario has to be sufficiently difficult, so that a proportion of pilots will not succeed and the corresponding wide range in performance can be observed.

Scenario. Before the experiment began, the pilots were informed about the aircraft's current state, including fuel on board, remaining flight time, navigation issues (e.g. approach details, maps), position and altitude via email two weeks in advance. After arriving at the simulator facilities, the pilot flying conducted his approach briefing. When the pilots entered the simulator, the aircraft was established on a visual approach under good weather conditions; fuel on board would suffice for 60 minutes at that time. Upon lowering the gear for the final approach, a malfunction was evoked which represented the leakage of the hydraulic system so that the nose gear was not able to fully extend and remained unlocked and unable to retract (malfunction 1). With this failure, the crew was forced to perform a go-around and work through the mandatory checklists and procedures. Due to the doubled aerodynamic drag, the fuel shortage meanwhile led to a mayday situation. With about 20 minutes of remaining flight time, the crews were now on their second approach. As a consequence of the damaged hydraulic system, the flaps and slats did not only extend slowly, but also jammed in their current position (malfunction 2); at that time, the high workload condition began. Again, the crew had to abort the approach and handle the procedures. At that point it was expected that about one half of the crews had to abort their trouble shooting process and force a landing with the current flight configuration.

Participants & Experimental Conduction. For this part of the experiment, 12 randomly selected pilots (6 Airbus A320, 6 Airbus A340) and therewith 6 crews flew the scenario one by one, while two instructor pilots operated the full flight simulator and acted as air traffic controllers in parallel. The pilots (Captains / First Officers) of the A320 fleet were M = 47/29, SD = 1.7/2.7 years old and had a total amount of M = 14,277/2,900, SD = 525/848 flight hours. The pilots of the A340 fleet were M = 52/37, SD = 1.7/3.3 years old and had experience of M = 17,667/9,354, SD = 1,699/2,248 flight hours; all the pilots hold valid ATP licenses with appropriate type ratings. The two flight instructors were both recently retired, M = 60 years old and had M = 21,000 hours of flight experience and served M = 20 years as instructor pilots within the same airline as the participating pilots.

The experimental conduction took place during two nights at a training facility, where three crews flew the scenario in a full flight simulator (*JAR STD 1A Level D*) each night. After the scenario was completed, the pilots left the simulator and both instructors independently rated the pilots' performance using the CRM evaluation form and the LOSA sheet mentioned above; the instructors were not allowed to talk to each other and the two participating pilots were separated into two rooms to conduct their ratings. One month later, one of the two instructors received the edited video and audio recordings of the scenario for retesting. At that point, the rater did not have any copies of his initial ratings. Since he rated another 72 pilots during the remaining experiment, it can be assumed that he did not remember particular ratings, but in the case of outliers, it is assumed he would probably recognize the pilot's behavioral patterns.

Results & Discussion

The results are presented in the order of the research questions stated; a short discussion of the particular aspects directly follows. Reliabilities are analyzed using intraclass correlation coefficients based on a two-way random model *ICC(2)* under the requirement of absolute agreement.

Results regarding the first research question (RQ 1), which refers to a classical inter-rater reliability problem, show the dependency of inter-rater reliability on different rating dimensions (compare Figure 2). When

looking at the CRM skills, only *Communication* and *Leadership & Teamwork*, which can be defined as social competencies, reach the cut-off level of .7, whereas, in contrast, the cognitive skills (*Work Organization* and *Situation Awareness & Decision Making*) do not. This could be explained by the fact that those social aspects can be observed directly and no further interpretation is necessary. *Work Organization* and *Situation Awareness* may require of the instructors more assumptions about observable behaviors, which could differ. These results are in accordance with Sevdalis et al. (2008), where *Communication* and *Teamwork* as well as *Leadership* achieve the highest reliabilities (.63 and .66) compared to the other categories. Results from the LOSA analyses show that only the *Planning Behavioral Markers* under the high workload condition reach the required reliability of .7; all other dimension are rated with a lower reliability. The reliabilities of the LOSA rating under the low workload conditions are significantly smaller than under the high workload conditions. It can be assumed that the spread of performance under high workload is more developed and therefore easier to rate. Both workload conditions reflect medium to high reliability regarding the *Planning Behavioral Markers* and very low reliabilities regarding *Review & Modify Behavioral Markers*.

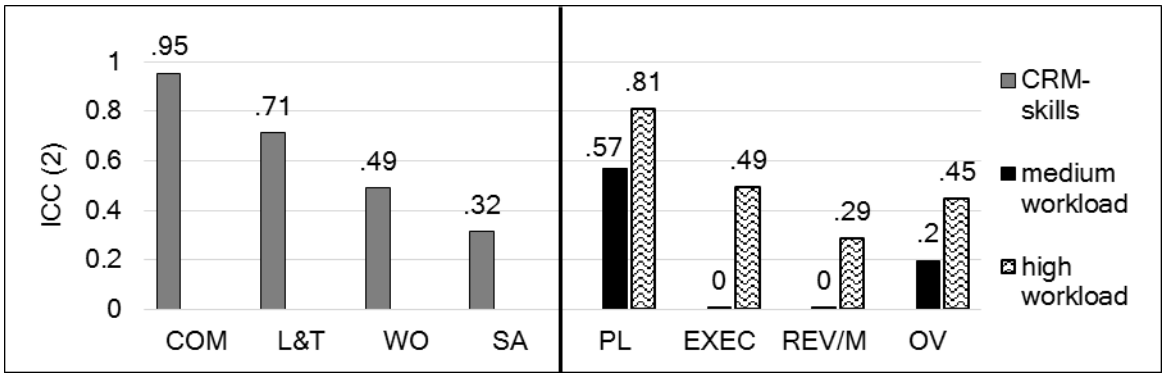


Figure 2. Inter-rater reliability between rater 1 and rater 2 as a function of different rating dimensions. COM = Communication, L&T = Leadership & Teamwork, WO = Work Organization, SA = Situation Awareness & Decision Making, PL = Planning, EXEC = Execution, REV/M = Review and Modify, OV = Overall.

In order to reach an acceptable level of reliability (.7), the Spearman-Brown prophecy formula (see Lienert & Raatz, 1994) was used to calculate the minimum number of raters required for a reliable rating of pilots during simulator missions. For this calculation, the single dimensions of the respective categories were averaged using *Fisher z' transformation* (Fisher, 1925). It is confirmed that for average non-technical skills ratings, one rater is sufficient. In comparison, for averaged LOSA ratings under the low workload condition, nine raters would be needed; under high workload conditions, two raters are sufficient. Applying this data to a required .9 reliability level, CRM ratings would need eight pilots. These findings are in accordance with Brannick, Prince, and Salas (2002), who postulated a need for nine raters for their comparable data on the .9 reliability level. Regarding research question 2.1, which aims for the test-retest reliability, results indicate, in contrast to the previous mentioned results, that this kind of reliability is higher for the LOSA categories than for the non-technical skills (compare Figure 3).

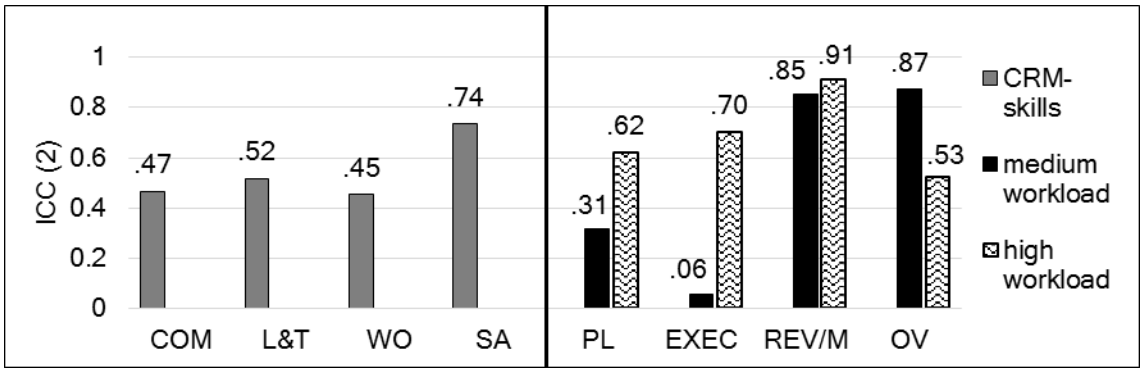


Figure 3. Test-retest reliability within rater 2.

Especially the test-retest reliability for the high workload condition leads to very high reliabilities in comparison to the inter-rater reliability. This means that the rating is consistent for one rater (high test-retest reliability), but nevertheless strongly differs between the raters (medium reliability). The personal interpretation of the instructor seems to induce more variance for the rating than the actual performance of the subject does.

In terms of agreement regarding pilots' self-estimation compared to an external point of view, RQ 2.2 delivers the following results (compare Figure 4):

1) The agreement highly depends on the dimension that is rated; aspects of *Situational Awareness & Decision Making* are rated with good reliability.
2) In three out of four dimensions (L&T, WO, and SA), the test rating (simulator) fits better with the self-evaluation than the retest rating (video-based). Only the category *Communication* is rated with slightly higher agreement during the retest rating.

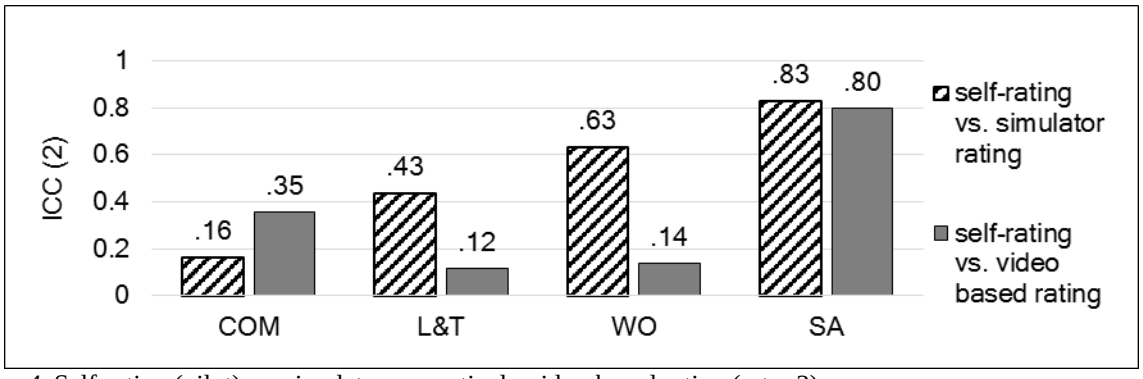


Figure 4. Self-rating (pilot) vs. simulator respectively video-based rating (rater 2).

Conclusion

When interpreting the results, it has to be kept in mind that the LOSA rating forms were new to the instructor pilots. In general, the results showed that reliability highly depends on the dimension that is rated and even a retest does not lead to higher congruency between pilots' self-estimation and instructors' ratings. This could mean that in video-based debriefing situations, where the instructor and the pilot have time to reflect the mission more often, the subjective perception can differ more between pilot and instructor than directly after the mission being completed. Furthermore, it seems that it can make sense to incorporate more instructor pilots in one assessment when it comes to specific rating dimensions.

Acknowledgements

The authors acknowledge the support of Cpt. Manfred Binder, Cpt. Peter Croeniger and Tanja Kammann, B.Sc. during data collection and handling. This work was funded by the German Federal Ministry of Economics and Technology via the Project Management Agency for Aeronautics Research within the Federal Aeronautical Research Program (LuFo IV-2).

References

Brannick, M. T., Prince, C., & Salas, E. (2002). The Reliability of Instructor Evaluations of Crew Performance: Good News and Not So Good News. *International Journal of Aviation Psychology, 12*(3), 241–261.

Burger, K.-H., Neb, H. & Hoermann, H.-J. (2003). Lufthansa's new basic performance of flight crew concept - A competence based marker system for defining pilots performance profile. *Proceedings of The 12th International Symposium on Aviation Psychology, 1*, 172–175.

Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for Abnormal events. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 55*(3), 477–485.

Dietrich, R. (2004). Determinants of effective communication. In T. M. Childress & R. Dietrich (Eds.), *Group interaction in high risk environments*. Aldershot: Ashgate.

Fisher, R. A. (1925). *Statistical Methods For Research Workers*. Edinburgh: Oliver and Boyd.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). *Wiley series in probability and statistics*. Hoboken, N.J.: J. Wiley.

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *International Journal of Aviation Psychology, 11*, 95–118.

Gontar, P., & Hoermann, H.-J. (2014). Flight Crew Performance and CRM Ratings Based on Three Different Perceptions. In A. Droog (Ed.), *Aviation Psychology: facilitating change(s): Proceedings of the 31st EAAP Conference* (pp. 310–316).

Gontar, P., Hoermann, H.-J., Deischl, J., & Haslbeck, A. (2014). How Pilots Assess Their Non-Technical Performance - A Flight Simulator Study. In N. A. Stanton, S. J. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Human Aspects of Transportation* . AHFE International.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology*, *9*(1), 19–32.

Klinect, J. R., Murray, P., Merritt, A. C., & Helmreich, R. L. (2003). Line Operations Safety Audit (LOSA) - Definition and operating characteristics. In *Proceedings of the 12th International Symposium on Aviation Psychology* . Dayton, OH.

Landis, R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.

Law, J., & Sherman, P. (1995). Do raters agree? Assessing inter-rater agreement in the evaluation of aircrew resource management skills. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 608–612). Columbus, OH: Ohio State University.

Law, J., & Wilhelm, J. (1995). Ratings of CRM skill markers in domestic and international operations. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 669–675). Columbus, OH: Ohio State University.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th ed.). Weinheim: Beltz, Psychologie Verl.-Union.

O'Connor, P., Hoermann, H.-J., Flin, R., Lodge, M., & Goeters, K.-M. (2002). Developing a Method for Evaluating Crew Resource Management Skills: A European Perspective. The International Journal of Aviation Psychology, 12(3), 263–285.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed). Boston: McGraw-Hill.

Schubert, E., & Haslbeck, A. (2014). Gestaltungskriterien für Szenarien in Flugsimulatoren zur Untersuchung von Verhalten und Leistung von Verkehrspiloten. In Deutsche Gesellschaft für Luft- und Raumfahrt - Lilienthal-Oberth e.V. (Ed.), *DGLR-Bericht, 2014-01. Der Mensch zwischen Automatisierung, Kompetenz und Verantwortung* (pp. 125–137). Bonn.

Seamster, T., Hamman, W., & Edens, E. (1995). Specification of observable behaviors within LOE/LOFT event sets. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 663–668). Columbus, OH: Ohio State University.

Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., & Vincent, C. A. (2008). Reliability of a revised NOTECHS scale for use in surgical teams. *The American Journal of Surgery, 196*(2), 184–190. doi:10.1016/j.amjsurg.2007.08.070

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, *86*(2), 420–428.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System. *World Journal of Surgery*, *32*(4), 548–556.

BEHAVIORAL TRAPS IN CREW-RELATED AVIATION ACCIDENTS

Jonathan Velázquez
Inter-American University of Puerto Rico
Bayamón, Puerto Rico
Allen Peck
U.S. Air Force
Montgomery, Alabama
Timothy Sestak
Embry-Riddle Aeronautical University
Prescott, Arizona

The majority of aviation accidents are still attributed to human error, with flight crew actions accounting for the majority of these mishaps. The Federal Aviation Administration (FAA) has identified 12 behavioral traps that can ensnare even experienced pilots. This study examined the FAA-defined behavioral traps and the regularity with which they occurred in flight crew related accidents. The top three traps were *Neglect of Flight Planning, Preflight Inspection, and Checklists*; *Loss of Positional or Situational Awareness;* and *Getting Behind the Aircraft,* which were found in 72%, 61%, and 48% of aviation accidents, respectively. The results showed the contributing factors of training inadequacies/lack of Crew Resource Management, night, and low ceiling and/or visibility compounded the effects of the unsafe attitudes. These conditions were found in 48%, 46%, and 42% of accidents, respectively.

Approximately three out of four aviation accidents result from human error (FAA, 2009). The FAA uses studies in human behavior in an effort to reduce human error in aviation accidents. Flying consists of decision making activities, some of which are routine, others more complex. Effective aeronautical decision making (ADM) is essential to flight safety. The first two steps of ADM are "(1) identifying personal attitudes hazardous to safe flight and (2) learning behavior modification techniques" (FAA, 2009, p. 5-3).

Unsafe pilot behaviors have been part of FAA literature since the foundations of ADM. The concept of *hazardous attitudes* refers to pilot personality factors that may affect decision making and judgment. These attitudes include macho, anti-authority, impulsivity, resignation, and invulnerability. *Behavioral traps* (refer to Table 1) are operational pitfalls to which aviators may fall prey as a result of bad decision making, often leading to negative consequences.

Literature Review

Helmreich and Foushee (1993) found that flight crew actions were the cause in more than 70% of accidents between 1959 and 1989. Wetmore and Lu (2006) studied fatal general aviation (GA) accidents and found that hazardous attitudes have a devastating effect on risk management, decision making, and the utilization of all resources, three of the most important skills in Crew Resource Management (CRM).

Table 1.
Overview of Behavioral Traps as defined by the FAA (2008, p. 9-12)

Behavioral Trap	Definition
Peer Pressure	Poor decision making may be based upon an emotional response to peers, rather than evaluating a situation objectively.
Mind Set	A pilot displays mind set through an inability to recognize and cope with changes in a given situation.
Get-There-Itis	This disposition impairs pilot judgment through a fixation on the original goal (destination), including a disregard for any alternative action.
Duck-Under Syndrome	A pilot may be tempted to make it into an airport by descending below minimums during an approach. A pilot may believe that there is a built-in margin of error in every approach procedure, or a pilot may not want to admit that the landing cannot be completed […].
Scud Running	This occurs when a pilot tries to maintain visual contact with the terrain at low altitudes while instrument conditions exist.
Continuing Visual Flight Rules (VFR) into Instrument Conditions	Spatial disorientation or collision with ground/obstacles may occur when a pilot continues VFR into instrument conditions. This can be even more dangerous if the pilot is not instrument rated or current.
Getting Behind the Aircraft	This pitfall can be caused by allowing events or the situation to control pilot actions. A constant state of surprise at what happens next may be exhibited when the pilot is getting behind the aircraft.
Loss of Positional or Situational Awareness	In extreme cases, getting behind the aircraft results in a loss of positional or situational awareness. The pilot may not know the aircraft's geographical location or may be unable to recognize deteriorating circumstances.
Operating without Adequate Fuel Reserves	Ignoring minimum fuel reserve requirements is usually the result of overconfidence, lack of flight planning, or disregard of regulations.
Descent Below the Minimum En Route Altitude	The duck-under syndrome, as mentioned above, can also occur during the en route portion of an Instrument Flight Rules (IFR) flight.
Flying Outside the Envelope	The assumed high-performance capability of a particular aircraft may cause a mistaken belief that it can meet the demands imposed by a pilot's overestimated flying skills.
Neglect of Flight Planning, Preflight Inspections, and Checklists	A pilot may rely on short- and long-term memory, regular flying skills, and familiar routes instead of established procedures and published checklists. This can be particularly true of experienced pilots.

The understanding of individual pilot attitudes and their role in CRM still requires further research (Salas, Shuffler, & Diaz, 2010). The study of unsafe pilot attitudes has extended over three decades (Casner, 2010; Hunter, 2005; Lester & Bombaci, 1984; Murray, 1999). However, much has been limited to GA and to the hazardous attitudes. This study examined pilot behavioral traps in the multi-crew environment and aimed to see with what regularity behavioral traps were extant in crew-related aviation accidents. The specific research questions were:

1. Which behavioral traps are present, and with what frequency do these occur, in flight crew related accidents?

2. What relationships exist between the pilot behavioral traps and the contributing factors to aviation accidents?

Methodology

The study used archival methods to explore the behavioral traps contributing to flight crew accidents. The primary data source was the Flight Safety Foundation's (FSF) accident report archives. Research focused on FSF's *Accident Prevention* periodical, which cataloged 218 accidents from 1988 to 2006. From a total of 218 reports, 110 were determined to have flight crew-related causes. Using the Krejcie and Morgan (1970) formula, an appropriate sample of reports to review consisted of 83 accidents attributed partly or wholly to flight crew error. A description of the dataset and the database itself can be obtained at the FSF website (http://flightsafety.org).

The research team analyzed the accident reports to determine the presence of a primary behavioral trap, then wherever applicable, any secondary behavioral traps that may have been contributory. The researchers also identified contributing situational factors, such as weather, training/CRM, maintenance, etc., that may have exacerbated the effect of the behavioral traps. The researchers employed *a priori* codes, specifically, the FAA-defined behavioral traps. Once the coding process was completed, the research team explored any relationships among them with the contributing factors. All the relevant information from the accident reports was entered into NVivo (v. 10), a computer-aided qualitative data analysis software. The use of such software allowed for a second stage of coding where themes began to emerge (e.g., contributing factors) in conjunction with the behavioral traps themselves.

Results

Descriptive Statistics

Table 2 shows the frequency with which behavioral traps were present in the 83 accident reports as either a primary or secondary behavior. A primary behavior is a flight crewmember action or inaction which is most closely related to the investigative agency's accident probable cause. Accidents are usually the result of a series of events that each add operational risk. Thus, secondary behavioral traps are actions contributing to the accident but not directly associated to the investigative agency's probable or primary cause statement. The three most prevalent traps were *Neglect of Flight Planning, Preflight Inspection and Checklist; Loss of Positional or Situational Awareness;* and *Getting Behind the Aircraft.*

While night was tracked separately, darkness can be considered a contributory factor and was included in Figure 1 showing the frequency of occurrence of the contributing factors. Results showed that *training inadequacies/lack of CRM*, *night*, or *low ceiling and/or visibility* compounded the effects of the unsafe attitudes; these conditions were found in 48%, 46%, and 42% of accidents, respectively. The *other* category included miscellaneous conditions such as medical issues, optical illusions, etc.

Table 2.
Frequency Count of Behavioral Traps in FSF Accident Reports.

Behavioral Trap	Primary	Secondary
00-Neglect of Flight Planning, Preflight Inspections, Checklists	32	30
01-Peer Pressure	2	1
02-MindSet	5	18
03-Get-There-It is	2	9
04-Duck-Under Syndrome	5	3
05-Scud Running	0	1
06-Continuing VFR into IMC	2	2
07-Getting Behind the Aircraft	12	28
08-Loss of Positional or Situational Awareness	19	34
09-Operating Without Adequate Fuel Reserves	1	0
10-Descent below the MEA	1	1
11-Flying outside the Envelope	2	7

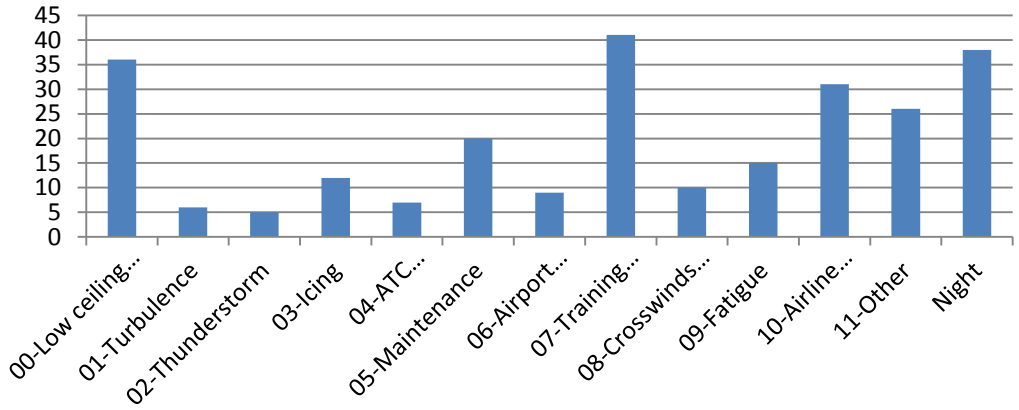


Figure 1. Contributory factors across all cases.

Relational Analysis Results

The research team became interested in exploring the relationships between the most prevalent behavioral traps and the contributing factors most present during the aviation accidents. Cluster analyses are good visualization tools based on the frequency with which words or coding are shared in the coded text. Figure 2 explored these relationships between the primary behavioral traps and low ceiling and/or visibility while Figure 3 explored the association between the primary behavioral traps and the crews' training inadequacies (lack of CRM).

The Figure 2 dendogram indicates how sources of information have coding similarities, which in turn could suggest relationships between two concepts. The proximity to, and color of codes within the diagram, suggest associations among the concepts. Low ceiling and/or low visibility is near, and shares the same color, to the behavioral trap known as *Loss of Positional or*

Situational Awareness. These connections are not difficult to comprehend since having restrictions to visibility could logically contribute to loss of positional awareness.

CRM is the ultimate expression of teamwork between flight crewmembers. Good CRM practices are predicated on following checklists, standard procedures, conducting good preflight, and engaging in proper flight planning to prepare for unexpected events. Not surprisingly, Figure 3 illustrated a relationship between training inadequacies (lack of CRM) and the behavioral trap known as *Neglect of Flight Planning, Preflight Inspection, and Checklists.*

Figure 2. Cluster analysis between low ceiling and/or visibility and the behavioral traps.

Figure 3. Cluster analysis between training inadequacies (lack of CRM) and the behavioral traps.

Discussion, Conclusions, and Recommendations

Behavioral traps were not present in a uniform distribution in the accident reports analyzed. In fact, three of them, *Neglect of Flight Planning, Preflight Inspection, and Checklist*, *Getting Behind the Aircraft*, and *Loss of Positional or Situational Awareness* accounted for 63 (over 75%) of the primary behavioral traps, and as secondary behavioral traps, they each appeared in over one-third of the cases. In all but three accidents considered, one or more contributing factors were present. One could infer that behavioral traps are exacerbated by adverse environmental factors.

The researchers found relationships between restrictions to vision (e.g., night conditions, low visibility/ceilings) with the behavioral trap known as *Loss of Positional or Situational Awareness.* The fact that restrictions to vision is still a factor in many accidents may prompt the FAA to research and develop training or public awareness on how to improve overall situational awareness during conditions such as these and study technological enhancements. The link found between training inadequacies and the trap known as *Neglect of Flight Planning, Preflight Inspection, and Checklists* suggests that some pilots are not employing effective teamwork practices, rules, and standard procedures.

From the standpoint of accident prevention, training and education focused on the top behavioral traps would likely prove to have the highest payoff. Knowledge of how these behavioral traps manifest themselves in crews can re-focus portions of CRM teaching to include cognitive biases training and/or hazardous behavior identification and modification techniques. Currently, the FAA lacks standardization for CRM training and guidelines concerning attitude management. The present study of behavioral traps could provide an excellent starting point.

References

Casner, S. M. (2010). General aviation. In E. Salas & D. Maurino (Eds.), *Human factors in aviation* (2nd ed.). (pp. 595-628). Burlington, MA: Academic Press – Elsevier.

Federal Aviation Administration. (2008). *Aviation instructor's handbook.* Washington, D.C.: Government Printing Office.

Federal Aviation Administration. (2009). *Risk management handbook*. Washington, D.C.: Government Printing Office.

Helmreich, R. L. & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 3-45). San Diego, CA: Academic Press.

Hunter, D. R. (2005). Measurement of hazardous attitudes among pilots. *The International Journal of Aviation Psychology*, *15*(1), 23-43.

Krejcie, R. V. & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychology Measurement, 30,* 607-610.

Lester, L. F., & Bombaci, D. H. (1984). The relationship between personality and irrational judgment in civil pilots. *Human Factors, 26,* 565-572.

Murray, S. R. (1999). FACE: Fear of loss of face and the five hazardous attitudes concept. *The International Journal of Aviation Psychology*, *9*(4), 403-411.

Salas, E., Shuffler, M. L., & Diaz, D. (2010). Team dynamics at 35,000 feet. In E. Salas & D. Maurino (Eds.), *Human factors in aviation* (2nd ed.). (pp. 249-291). Burlington, MA: Academic Press – Elsevier.

Wetmore, M. & Lu, C-t. (2006). The effects of hazardous attitudes on crew resource management skills, *International Journal of Applied Aviation Studies, 6*(1), 165-182.

FAA TRAINING ASSESSMENT OF ON-THE-JOB TRAINING

Darendia McCauley
FAA Civil Aerospace Medical Institute
Oklahoma City, Oklahoma USA

The field training administered to air traffic controllers is provided by instructors who are also controllers but have undergone training to become on-the-job training instructors (OJTIs). The training of these field OJTIs is under revision based on a training needs assessment conducted in 2011. Controllers who were not successful in training at their first facility can be reassigned to another facility, where they receive field training specific to that facility. The controllers who did not succeed in training initially and requested reassignment may have a useful perspective on the training they received. This perspective may aid in revising the training provided to OJTIs, in addition to and in conjunction with the earlier training needs assessment.

Field training occurs through on-the-job familiarization (OJF) and on-the-job training (OJT). OJF consists of classroom and scenario training provided to familiarize students with the specific airspace, procedures, and processes unique to a given facility. OJT can occur in a local lab environment, but it predominantly occurs on position while working live traffic. OJT is administered to all air traffic control (ATC) trainees new to the facility, regardless of whether it is their first facility assignment, they already certified at another facility, or they requested reassignment (called ERR, for Employee-Requested Reassignment) from another facility before completing their training. The OJTI must possess the knowledge to teach, coach, and demonstrate techniques for safely and efficiently controlling air traffic using a specific set of procedures. The local airspace and conditions surrounding it are taught both through OJF and OJT. OJT instruction is based on ATC procedures and must also provide guidance on control judgment (when to intervene on position). The OJTI role is critical to the success of training.

FAA completed a field needs assessment of the training (Lacroix, Shelly, Lake, & Brodie, 2011b) required by OJTIs. This assessment documented several areas of OJTI training that needed to be updated or enhanced. In addition, an independent review panel (Barr, Brady, Koleszar, New, & Pound, 2011) had specific recommendations for updates and revision for both initial and recurring OJTI training. There was some overlap in these recommendations. However, the needs assessment targeted specific training areas necessary for training success in need of improvement, while the independent review panel identified more organizational or policy changes needed. The recommendations that came from the training needs assessment identified the need for more OJT experience with simulators, improvement in OJTI training skills and tools, and better communication skills, especially in safety critical areas like developing air traffic situations and remediating personality issues between the OJTI and the trainee. There were also specific concerns identified related to student preparation for the OJT experience and overall training provided for OJTIs.

Trainees who requested reassignment before completing field training at their first facilities (ERR) return to the FAA Academy as a part of their training for a new field assignment. They

are required to take Academy training appropriate to the type of facility to which they have been assigned if they had not previously received it. For example, students who had failed training at large radar facilities (either En Route or TRACON) are usually reassigned to a smaller tower or tower/TRACON facility. Those students return to the Academy for initial tower training if they did not take it previously. As a part of all Academy training, controller developmentals also receive training about fatigue. For those who return for tower training after having previously attended Academy training for radar facilities, the fatigue lesson was modified to accommodate their prior exposure.

Controllers who failed training at their first facility were unsuccessful in their initial participation in OJT. Their perceptions of OJT and the OJTIs providing it may provide useful information that can be used to revise the way OJTIs are trained. It also provides an opportunity to compare their insights with already identified training needs.

The FAA's Office of Safety and Technical Training identified three OJTI training projects. The first project, Supplemental OJTI training, was developed for existing OJTIs in FY14, and will be launched in FY15. The second project is a revision of OJTI initial training, which is currently underway. OJTI initial training is usually conducted at field locations for prospective OJTIs. That training project involves a rebuilding of initial OJTI training based on several recommendation sources, including the previously conducted training needs assessment. The third OJTI training project is a CADRE course for training instructors to deliver the OJTI initial course to candidate OJTIs at their facilities. The CADRE course will be developed shortly following the initial course and using the same concepts. The purpose of this research is to use information from trainees who were unsuccessful in field training to validate OJTI training needs identified in the prior needs analysis.

Method

Participants

Eighteen employee-requested reassignment (ERR) students volunteered to answer questions about the fatigue lesson as well as some questions about their OJT experience and their OJTIs. The focus of this paper is on the information from the second part of the survey, which focused on ERR perceptions of OJTIs and OJT issues.

Tools and Procedures

We asked students if OJTIs provided any input concerning scheduling and dealing with fatigue issues. In addition, we asked students about access to and experience with simulators, as well as how OJTIs responded to other OJT issues previously identified as needing inclusion or revision to OJTI instruction. We obtained these questions from a training needs assessment for the OJTI course conducted in 2011.

We asked students to indicate from strongly disagree (1) to strongly agree (6) about statements relative to the OJT experience. The questions were extracted from the earlier training needs

assessment for OJTI course revision. There were no demographic indicators collected in order to ensure anonymity.

Results

Survey Responses

Student survey responses to OJT questions are indicated in Table 1. While there was some variability, responses were generally reflected toward disagreement with statements. On average, survey respondents indicated that the needs previously identified by the training needs analysis were also needs perceived by those who had recently had unsuccessful field training experiences.

Table 1.

ERR Student Responses to OJT Questions.

Based on your recent field training experience, do you agree or disagree with the following statements? Strongly Disagree (1) to Strongly Agree (6)	Mean	SD
Access to and experience with simulators was adequate	3	1.7
OJTIs used past experience to enhance instruction	3.1	1.5
OJTIs provided adequate emphasis on training	3.2	1.4
OJTIs adequately communicated developing situations	3.4	1.7
OJTIs were an adequate match for me	2.4	1.6
OJTI instructors were adequately trained	2.8	1.6
Adequate preparation for the OJT experience was provided	3.2	1.6

Limitations

The sample size was small (n=18). It should also be noted that as trainees for whom training was not successful, ERR students have a unique perspective of OJT and the OJTIs they have experienced. Comparable data were not collected from CPCs who successfully completed training and those who did not request reassignment to a different facility. Therefore, the outcomes presented here came from a very small and a very specific group of students.

Discussion

Support for OJTI Training Revisions

On average, ERR students indicated that the training received did not provide some of the experiences necessary for successful training. This supports the needs for OJTI training revision previously identified by FAA OJTI workgroups and specific recommendations from the OJTI training needs assessment identifying those experiences necessary for successful OJT. Student responses indicated a need for an increased focus on simulation experiences. This would include using simulation equipment for terminal or radar exercises, where students and OJTIs can work together to better develop needed skills without the pressure and risk of controlling live traffic.

OJTIs can also be encouraged to provide their own past ATC experiences to provide a personalized reality and enhance instruction. Communication skills are critical for an OJTI. The OJTI must be able to respond appropriately and quickly in critical situations to maintain safety in developing situations to make the trainee aware of the situation and why intervention was needed. Effective communication must be able to cross personalities and learning styles. There is also a need for providing improved OJT expectations to the trainees. This could result in improved OJF as well as improved OJT. The need for improvement in OJTI preparation and training is evident.

Acknowledgments

Research reported in this paper was conducted under the Air Traffic Program Directive / Level of Effort Agreement between the Human Factors Research and Engineering Group (ANG-C1), FAA Headquarters, and the Aerospace Human Factors Research Division (AAM-500) of the Civil Aerospace Medical Institute.

References

Barr, M., Brady, T., Koleszar, G., New, M., & Pounds, J. (2011). FAA Independent Review Panel on the Selection, Assignment, and Training of Air Traffic Control Specialists, Washington, DC.

Lacroix, D., Shelly, K., Lake, J., & Brodie, L. (2011 b). On the Job Training Instructor (OJTI)Task and Skills Analysis Report. (FAA OJTI Workgroup Internal Publication). FAA Civil Aerospace Medical Institute, Oklahoma City, OK.

EXPLORING THE MATHEMATICAL PREDICTABILITY OF THE ADVANCED AIRCRAFT TRAINING CLIMATE

Preven Naidoo, PhD
University of Pretoria
South Africa

Effective pilot training on advanced aircraft is vital in ensuring flight safety, and positive perceptions of the training climate can contribute to the success of the training. Hypothetically, characteristics of the trainee can predict the training climate. Thus far, predictive models have provided little information about the mental viscosity or psychological comfort of the processes of pilot training. The purpose of this study was to develop a mathematical model to predict the psychological comfort of the organisational environment for advanced aircraft pilot transition training using a dichotomous categorical criterion. A predictive model of the phenomena was contemplated from a non-parametric regression process. Original data were captured using a previously validated instrument, namely the Advanced Aircraft Training Climate Questionnaire (AATC-Q) from a cohort of 229 respondents. The final regression model containing four independent variables correctly predicted 63.8% of overall cases. The developed model may assist in implementing training interventions.

Pilot selection, recruitment and training are some of the more important measurable antecedents available, to detrmine the suitability of a pilot to operate an advanced aircraft (Machin & Fogarty, 2003; Pasztor, 2009). Anecdotally, for example, the final report into a well-publicised air crash involving an Air France Airbus A330 in 2009, suggested that human action, stemming from specific training issues, was a significant contributor. This paper attempts to statistically analyse some of the possible behavioural variables which may impact advanced aircraft pilot recruitment and training.

The 'advanced aircraft training climate' construct

Two separate climatic constructs are generally differentiated in the literature, namely a 'psychological climate' and an 'organisational climate' (Denison, 1996, p. 619). The psychological climate refers to making sense cognitively of the organisational environment. An organisational climate refers to the subjective summated (average) sense that individuals make of interpersonal constructs. An extension to this is their understanding of policies, procedures and structure in an organisation. The organisational culture can be defined as a set of group assumptions created after learning from a number of internal and external difficulties or problems. Climate researchers are therefore 'generally less concerned with [social] evolution but more concerned with the impact that organisational systems have on groups and individuals' (Denison, 1996, p. 621). The advanced aircraft training climate is (for the purposes of this paper) defined as all factors in the person, learning and organisation that influence [the] transfer of knowledge to the job function. According to Fishbein and Ajzen (2001), the study of behaviour is defined as an analysis of the functional relationship between events in the environment and human action, or inaction. The theory supports a link between an individual's attitude and subsequent behaviour (Fishbein & Ajzen, 2001). Thus, it follows that a training perception may directly impact trainees' learning attitudes. Prior research in this topic tend to adopt a linear regression method in order to formulate relationships between perceptions, attitude, beliefs and intention with behaviour. These variables are then interconnected within relationships with specific mediating demographic variables. Analyses of these linear linkages provided the foundational hypothesising in this study to predict both flight performance success and inter-crew behaviour. Due to space constraints this is not elaborated here.

Briefly, airline pilot training in general, consists of a theoretical or learning part, flight simulation training and route (or practical flying) training. The airline uses route training for the final assessment of a

candidate's ability to safely operate an advanced aircraft. Based on this background, specific demographic variables were selected and analysed when regressing a predictive mathematical model of the problem.

Research design

Participants

The target population consisted of around 1200 South African airline pilots who were experienced in advanced commercial aircraft. A final cohort of 229 participants was used for the analyses. A convenience sample was derived from the six large, medium and small airline organisations based in South Africa. Although 10.9% of the total sample held a South African pilot's licence, they were not affiliated to any particular carrier. Nonetheless, in general the sample frame was well represented. The majority of the participants (48.7%) were affiliated with the state carrier. In terms of the general flight experience levels of the group, the sample was fairly well distributed, and most of the respondents had more than 5 000 hours flying experience (Mean=9753.29; SD=6116.719). However, the dispersion of the participants in terms of flight experience was large – the majority of the sample had between 3 000 and 16 000 flight hours. The high standard deviation of this descriptor simply reflects the heterogeneity of pilots found in the South African airline industry (Vermeulen, 2009). This is also a testament to the high levels of industry experience in the South African aviation market. A fair proportion of the respondents (41.5%) were experienced in some of the most advanced commercial airline aircraft currently in operation globally, namely Airbus A319/A320/A330/A340 (41.5%) and Boeing 747-400/737-400/800 (24.9%). These demographics further articulate the skewed proportions regarding aircraft type and manufacturer category within this airline environment.

Measuring instrument

The Advanced Aircraft Training Climate Questionnaire (AATC-Q) was used as a data collection instrument (Naidoo, Schaap & Vermeulen, 2014). The instrument is a valid and reliable (α > 0.70) measure of South African airline pilots' perceptions of the training climate associated with the advanced aircraft mentioned earlier. The main scales of the instrument measured three latent factors. Factor 1 (Organisational Professionalism) consists of statements that measure both the macro domain (the airline) and the intermediate (instructor-trainee) domain. The component expresses the theoretical construct in terms of the efficiency, effectiveness and professionalism of both the company and its flight instructors. Factor 2 (Intrinsic Motivation) consists of items representing the micro level of analysis (the individual or person). The subscale predominantly reflects individual trainees' ability and eagerness to learn. Factor 3 (Individual Control of Training Outcomes) expresses the micro level of analysis (the person) relating to an individual trainee's own perceived level of control in terms of stress levels and learning decision-making. The essence of this third behavioural subscale relates to the levels of perceived personal control experienced by trainee pilots during training, their belief in their ability to effect the outcome of a training session, their capacity to maintain appropriate levels of stress (eustress or anxiety) in order to perform well, and ultimately their grasp of the amount of information required to cope with their training (intelligent decision-making).

Results

A backward stepwise logistic regression, was the process of choice, because '[r]egression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables' (Hosmer & Lemeshow, 2000, p. 1). The probability of a binary outcome on a discrete variable was modeled from the most likely relationship between specific demographic covariates. The logistic equation began with all the selected independent variables first entered and then deleted after evaluation (Table 1). A dichotomous dependent

variable was constructed, based on the level of favourability perceived by the respondents. As in discriminant analysis, a dummy variable (1 or 0) was allocated to the dichotomy of perceiving a favourable or unfavourable training climate, assuming that a positive perception of the favourability of climate exceeds 5.0 (on the AATC-Q behavioural scale), which is an average perception on the aforementioned three-factor model. This provided a dichotomous measure (labeled 'Favourability') of a categorical outcome that indicated the level of airline pilots' comfort (psychological viscosity) in the advanced aircraft training climate, in terms of their overall perception.

Table 1.
Values for demographic predictors used in the logistic model (independent variables)

Categorical variable	Value	Meaning
Interaction effect between flight experience*perceived level of computer literacy	1	Low Experience*Low Computer Literacy
	2	Low Experience*High Computer Literacy
	3	High Experience*Low Computer Literacy
	4	High Experience*High Computer Literacy
Age	0	40 years old or younger
	1	41 years old or older
Actual flight experience in advanced aircraft	1	Low experience (< 2000 hours)
	2	High experience (> 2001 hours)
Preference for route training	1	Never enjoy
	2	Sometimes enjoy
	3	Always enjoy
Preference for simulator training	1	Never enjoy
	2	Sometimes enjoy
	3	Always enjoy
Size of carrier employed at	1	Large
	2	Medium
	3	Small
Pilot unionisation	1	Unionised
	2	Non-unionised
Perceived level of computer literacy	1	Low
	2	High
Position at company	0	Co-pilot
	1	Captain

Final regression model

The data were regressed on an S-curve, which begins exponentially and thereafter tapers off. The plotted logistic regression model (not reproduced here) describes an initial exponential change in the probability of a favourable climate and thereafter, at some critical point, a slowing down or tapering off of probability. The logit formula, or curve, logit (p) = ln (p/1-p), also referred to as the log odds, described the final mathematical probability model. It was also noted that there were some distinct, yet acceptable disadvantages associated with using this particular logistic regression technique. For instance, the method required the researcher to use a high number of data points to produce meaningful and stable results. To determine how powerful the developed regression equation was in predicting the proportion of variance in the criterion variable associated with the predictor variable, a pseudo R^2 was computed, based on the methods of Cox and Snell's R^2, Nagelkerke's R^2, and McFadden's (adjusted) R^2 (Table 2). Therefore a pseudo R^2 was computed to evaluate the goodness-of-fit of the logistic model. The value of R^2 computed in this paper ranged from 0 to 1 (because squaring the correlation between the predicted values and the

actual values of the regression model would produce a positive value). This value is referred to as the pseudo R^2. Generally, a high pseudo R^2 value indicates that there is a high magnitude of correlation between the predicted values and the actual values. A cut-off value of $R^2\Delta = 0.02$ was used because the pseudo R^2 is not technically a goodness-of-fit index, and cannot explain the proportion of the variance *per se*. Nagelkerke's $R^2\Delta$ was computed by first calculating the value of the pseudo R^2 at the initial step and thereafter finding the difference at each subsequent step (Table 2). The Wald test was used to test the statistical significance of each of the coefficients in the regression model. A Z-score (Z = coefficient [B]/SE) was also calculated. The hypothesis of inclusion or exclusion of the coefficients was thus based on the subsequent chi-square fit. Because of the relatively small sample size in this study, a decision was made to use the likelihood-ratio test, comparing the maximised value of the likelihood function for the full model (L_1) to that of the likelihood function for the simpler or null model (L_0) associated with a chi-square goodness-of-fit. In addition, the Hosmer-Lemeshow goodness-of-fit test was applied to determine whether or not the model prediction differed significantly from the observed number of subjects in each group.

Discussion of results

A backward stepwise regression analysis was completed after five steps (Table 2). The final model containing four predictors subsequently emerged. Analyses revealed that the variables were, [1] an interaction effect between a pilot's level of flight experience in advanced aircraft and their perceived level of computer literacy (a unit change in this variable correctly predicted climate favourability by 65%); [2] practical flight experience in advanced aircraft; [3] preference regarding training in the flight simulator (a unit change in the pilot's enjoyment of simulator training improves the predictability of a positive climate by 69%); and [4] preference regarding route training in the actual aircraft. Additionally, five other exploratory predictor variables (Table 1) were not significant and were removed iteratively. It should be noted at this stage that McFadden's (1996) model contrasts the present results by finding that the independent variables; age, flying experience (total flying hours) and employer (major airline or non-major) as significant in pilots' perceptions of training in general. However, the difference in results between the two models may stem from the fact that the present model assesses the flight deck behaviour in terms of the training climate, whilst the McFadden model examined flight deck behaviour from an indepth analysis of aviation incidents. The overall percentage of cases for which the dependent variable was correctly predicted by the present study's mathematical model was 63.8%. The model is regarded as robust becuase it correctly predicts perceptions of a favourable climate 100% of the time (high positive predictive validity), however, disappointingly does not successfully predict respondents' perceptions of an unfavourable training climate. This result may be due to the design of the data collection instrument, which consisted of only positively worded statements, or items. A computation of Nagelkerke's $R^2\Delta$ suggested that the effect size of the model at each subsequent step was less than 0.02 and should therefore be regarded as not practically significant in terms of this criterion. Nonetheless, the findings in this study provide sufficient evidence to suggest that the final model is a highly efficient perception predictor. The efficiency of the resulting model is endorsed by the non-significance in the result of the Hosmer and Lemeshow test chi-square statistic in the final step ($\chi 2$ [7, N=229] = 2.365, p = 0.937). Because the dependent variable in this regression model is dichotomous or categorical in nature, only approximations of R^2 is possible. Hence Nagelkerke's pseudo R^2 was selected to gauge an alternative effect. Both Cox and Snell's R^2, together with Nagelkerke's R^2 values, were used in the study to conclude that the final model could reasonably account for approximately 12% to 17% of the variability in whether trainee pilots' perceived the training climate as either favourable or unfavourable. The researcher discovered moderate changes occurring in the -2 log-likelihood values between the constant only model, and the first and last step, which was a good indication that the modeling in the final step had an improved predictive power. A nominal regression of the final four predictor variables then produced a comparison in which the -2 log-likelihood values of the intercept only (95.338) and final model (65.456) indicated that the change in the amount of predictive power provided in the final solution was statistically significant [$\chi 2$ (4)

= 29.883, $p < 0.0001$]. McFadden's p^2 [1-log likelihood (final)/log likelihood (constant)] = 0.313 was computed as an indication of a measure of the strength of association between the predictor variables and the model. McFadden's p^2 is expected to be 'lower' than the traditional R^2 as a measure of effect size, and values between 0.20 and 0.40 are considered highly satisfactory. In terms of McFadden's p^2 (as opposed to Nagelkerke's $R^2\Delta$), the study concluded that the size of the final logistic model is large and of practical significance. Finally, the aforementioned logistic regression analyses show that the probability that a respondent would perceive the advanced aircraft training climate as favourable can be modelled using the following two logistic regression equations:

Equation 1

Logit = ln (p/1-p) = -2.603 + 0.63 * (interaction effect) – 1.064 * (advanced aircraft experience) + 0.485 * (route training) + 0.806 * (simulator training)

Equation 2

Prob (Favourable perception) = $(e^{-2.603 + 0.63\ X_1 - 1.064\ X_2 + 0.485X_3 + 0.806\ X_4})/(1+ e^{-2.603 + 0.63\ X_1 - 1.064\ X_2 + 0.485X_3 + 0.806\ X_4})$

Table 2

Final logistic regression prediction model

Predictors in the equation (X_J)	B	S.E.	Wald Chi-Square ($B^2/S.E.^2$)	Df	Sig.	Odds Ratio (E^b)	95% C.I. For Odds Ratio	
							Lower	Upper
Interaction effect	0.630	0.310	4.126	1	0.042	1.878	1.022	3.448
Advanced aircraft experience	-1.064	0.613	3.011	1	0.083	0.345	0.104	1.148
Enjoy route training	0.485	0.289	2.814	1	0.093	1.624	0.922	2.861
Enjoy simulator training	0.806	0.267	9.138	1	0.003	2.238	1.327	3.773
Constant	-2.603	0.912	8.142	1	0.004	0.074		

Practical implications and conclusion

The present study showed that predicting the psychological viscosity (ease and comfort of learning within the training climate) could be quantified within two mathematical formulae. This provides a more practical understanding of airline training success or failure. Airline recruitment specialists would therefore find that knowledge of these predictors might be of value when determining the success rates of potential new-hire pilots. The logistic equations show that perceived computer literacy (pilots' attraction or averseness to technology) plays a significant predictive role in the model only when it is combined with flight experience. Counter-intuitively, high flight experience did not necessarily imply a high probability of flight training success. The model predicts that when high flight time was coupled to technological averseness, the result was low psychological viscosity at the training level (that is, low climate favourability). In addition, pilots' preference for simulator training was by far the most

statistically dominant predictor. The study provides evidence to suggest that pilots tend to associate overall advanced aircraft flight training with their experiences in the flight simulator. Advanced aircraft incident and accidents have attracted, and will continue to attract international attention and scrutiny of pilot training. Selection, recruitment and training therefore play a critical role in flight safety and public opinion of organisations. However, determining precisely whether the organisation itself is actually capable of producing the competency in its pilots plays an increasingly important role in what eventually occurs on the flight deck. An implication predicted by the derived model in this paper, is that airlines should ensure that simulator-training devices are used to its full potential and are of a world-class standard. Advanced aircraft pilots should be given more opportunities to practice non-jeapordy exercises in flight simulator training devices. Additionally, the model also suggests that organisations that focus solely on a pilot's flight time (hours of experience), whilst neglecting perceived levels of computer literacy, create a myopic view of individual ability, and can hamper the overall training effort. Conversely, selecting pilots for advanced aircraft training with an experience level below a minimum threshold can also have an adverse impact on the training climate, due to the S-curve nature of this regression model. The logistic model produced here can provide predictions of what the ideal candidate experience and perception levels should be, implying greater training success for the individual, and effective and safer overall flight deck beahviour for the company.

Acknowledgements

This paper is partially based on the author's earlier doctoral dissertation, supervised by Professor Leo Vermeulen and Professor Pieter Schaap at the University of Pretoria, South Africa. The views in this research report do not reflect the views of the organisation's who participated in the study.

References

Caretta, T.R., & Ree, M.J. (1995). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379–388.

Denison, D.R. (1996). What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. *Academy of Management Review, 21*(3), 619–654.

Fishbein, M. & Ajzen, I. (2001). *Belief, attitude, intention and behaviour: an introduction to theory and research*. Reading, MA: Addison-Wesley.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression*. (2nd edn.). New York, NY: Wiley.

Machin, M.A., & Fogarty, G.J. (2003). Perceptions of training-related factors and personal variables as predictors of transfer implementation intentions. *Journal of Business and Psychology, 18*(1), 51–71.

Naidoo, P., Schaap, P., & Vermeulen, L. P. (2014). The development of a measure to assess perceptions of the advanced aircraft training climate, *The International Journal of Aviation Psychology, 24:3,* 228-245. doi: 10.1080/10508414.2014.918441

Parasuraman, R., & Byrne, E.A. (2002). Automation and human performance in aviation. In P.S. Tsang, & M.A. Vidulich (Eds.), *Principles and practices of aviation psychology* (pp. 311–356). Washington, DC: Erlbaum.

AIRFRAME PARACHUTE KNOWLEDGE AND DEPLOYMENT SCENARIOS: A COLLEGIATE PERSPECTIVE

Scott R. Winter[1], Robert C. Geske[2], Stephen Rice[1], Richard O. Fanjoy[2], and Lauren Sperlak[2]
Florida Institute of Technology[1]
Melbourne, FL USA
Purdue University[2]
West Lafayette, IN USA

As airframe parachutes in general aviation aircraft become more popular, training is essential in fostering a willingness to use the system in the appropriate situation. Aviation decision-making literature suggests that individuals make choices based on experience and pattern matching, such as emergency situations and airframe parachute deployment scenarios. This led the researchers to investigate the knowledge and perspectives of collegiate pilots who train in aircraft equipped with airframe parachutes. Participants completed a surveyfocused on airframe parachute knowledge and scenario-based examples. Training experts were used to validate the parachute deployment scenarios used in the instrument. Responses indicate that pilots find aircraft with a parachute safer than those without, but several participants reported inconsistencies with training for the use of the parachute. Findings suggest that when new safety technology is implemented into aircraft, training is necessary to ensure the technology is understood and implemented to its fullest effect and level of safety.

Airframe parachutes have become more common in the general aviation and light sport aircraft markets (McMahon, 2008). BRS Aviation, an airframe parachute manufacturer, reports that over 30,000 parachutes have been installed on general aviation aircraft and over 294 total lives have been saved from these safety devices (2012). These safety devices provide pilots or passengers the option of deploying an airframe parachute system, in certain emergency situations, that will lower the entire aircraft safely to the ground. However, in order for this safety device to be effective, pilots must be trained and willing to use it in an appropriate setting. The purpose of this study was to examine the knowledge and perspectives of a sample of collegiate pilots who use an aircraft equipped with an airframe parachute for primary training at the subject university. The research investigated student and instructor perceptions of usage scenarios and training for the airframe parachute system safety device equipped on the subject university's training aircraft. Errors occurring within the deployment decision-making process were reviewed, as well as the recent literature on perceptions and training to use an airframe parachute.

Review of Literature

Aviation Decision-Making in Critical Contexts

Decision-making is defined by O'Hare (2003) as "the act of choosing between alternatives under conditions of uncertainty" (p. 203). A key term within this definition is uncertainty. Since the effects may not be clear, the decision-maker must assess the situation based on incoming cues to determine the most appropriate outcome. Decisions may be further broken down into static and dynamic decisions (Craig, 2000). Static decisions are those which may not have a time component associated with them, while dynamic decisions occur within a more fluid environment where time pressure is associated with the decision.

Decision-making within the aviation field is frequently discussed within the context of naturalistic decision-making (NDM). When viewing decision-making through the lens of NDM, it is recognized that in real-world situations, there is likely to be some uncertainty with the outcome of a decision (O'Hare, 2003). NDM is defined as "an attempt to understand how people make decisions in real-world contexts that are meaningful and familiar to them" (Lipshitz, Klein, Orasanu, & Salas, 2001, p. 332). Unlike classical decision-making (CDM), in NDM it was found that persons did not identify and compare all options before making a decision, but rather, used previous experience and pattern matching to rapidly identify the appropriate outcome(s) (Klein, 2008; Orasanu, 1995). A key component of this is the term previous experience. As will be discussed in the current study, this is one reason why training for emergency situations such as an airframe parachute deployment is essential because in a real-life emergency, it may be easier to recognize the scenario than if this training had never been completed. Klein's Recognition-Primed Decision (RPD) model discusses how people use the matching of patterns and previous experience when it is necessary to make decisions (Klein, 2008). If the person recognizes the cue inputs from a previous situation, they will be likely to attempt the outcome that proved successful in the last experience.

Current Perspectives and Training on Airframe Parachute Decision-Making

An earlier study completed by McMahon (2008) investigated pilot perceptions of airframe parachute systems. An electronically administered survey gathered information on perceptions of airframe parachutes along with four scenarios for participants to consider if they would use the system. Over 1,000 participants completed the survey, however, these participants may or may not have ever flown or been trained on airframe parachutes or their usage. Participants indicated that 77% felt an aircraft with a parachute system was safer than one without. When isolated by flight instructors, those instructors with less than 2 years of experience were more likely to deploy the parachute system in the four scenarios, but this group also believed that the parachute system would lower them gently to the ground, an inaccurate perception. Young flight instructors, 94%, indicated that training on a parachute system was not necessary. McMahon (2008) concluded that flight schools using parachute equipped aircraft should invest in training so pilot expectations are more in line with the realities of using the system, along with a pre-flight briefing in how and when to use these safety systems. In the current study, the population consists of pilots who are trained in flying aircraft equipped with an airframe parachute system.

Research completed by Blickensderfer, Strally, and Doherty (2012) reviewed the impacts of scenario-based training on decision-making and airframe parachute deployment. Scenario-based training (SBT) has historically been used within aviation to train pilots for real-world situations, and it recently has had a renewed focus as the result of the FAA Industry Training Standards (FITS) program (Summers Halleran & Wiggins, 2010). Blickensderfer, Strally, and Doherty (2012) utilized undergraduate and graduate students, who held private pilot certificates with instrument ratings, for their study. None of the participants had experience flying an aircraft equipped with an airframe parachute. In their findings, it was reported that participants that received the SBT received significantly higher measures in overall pilot performance and higher self-efficacy than those in the control group (Blickensderfer, Strally, & Doherty, 2012). In terms of the overall decision to use the parachute the groups were not significantly different. The experimental group did, however, perform better on additional measures regarding the airframe parachute such as timing and altitude to deploy, along with a review of landing options (Blickensderfer, Strally, & Doherty, 2012). The researchers recognized that a number of participants in the pre-test flight admitted to forgetting about the possibility of using the airframe parachute, which suggests the need for training.

A major impetus for the current study was an earlier work by Winter, Fanjoy, Lu, Carney, and Greenan (2013) who found very few participants recognized a CAPS deployment situation when one was encountered on a flight in an aircraft training device. In that study, participants completed a scripted flight simulation in which they were flying on an instrument flight plan in instrument conditions with a ceiling of 400 feet above ground level (AGL). An expert panel validated the flight script before the experiment began, and it was determined that a CAPS deployment was the most appropriate outcome. However, in conducting the experiment, only nine out of 22 participants deployed the airframe parachute and of the nine who did deploy, only three followed the correct procedure as outlined in the aircraft's operating handbook (Winter, et al., 2013).

Methods

Participants

The participants consisted of members of the flight-training department at the subject university. Eligible participants were either students or flight instructors who completed training in the Cirrus aircraft. In total, 252 participants were deemed to be eligible to complete the survey: 49 flight instructors and 203 students. Over the three weeks the survey was available, 77 participants responded to the request, resulting in a response rate of 30%. However, only 63 participants completed the survey accurately, and therefore, only these responses were used in the data analysis. Of the 63 participants analyzed, 22 were flight instructors, a 45% response rate for this subgroup. The student subgroup had 41 responses or 20%.

Instrument

A survey was deemed the most appropriate and efficient instrument to collect data to answer the research question (Gall, Gall, & Borg, 2007). The authors developed questions, and the instrument was divided into two sections: general CAPS knowledge and scenario-based examples. Experts from Cirrus Aircraft, specifically the manager and director of flight training served as the content experts and validated the questions. A Cronbach's Alpha of 0.763 and 0.816 were recorded for the general CAPS knowledge and scenario-based examples sections, respectively, indicating a high level of reliability (Hinton, Brownlow, McMurry, & Cozens, 2005).

Procedure

The instrument was created on a web-based tool available at the subject university, which assisted in maintaining participant anonymity throughout the study, and the subject university's Institutional Review Board (IRB) provided an approval for the study. After the instrument was developed and checked for content validity, it was sent to the 252 eligible participants via e-mail addresses that were granted to the researchers from flight training records. The initial e-mail explained the purpose, instructions, and eligibility requirements to participate in the study and included a link to complete the survey. The survey window remained open for approximately 3 weeks near the end of the Spring 2013 semester and participants received three e-mail reminders requesting participation.

Results and Discussion

Demographic Information

Participants had a rather large variance in some of the demographic categories, especially within the flight instructor subgroup. For the students, the mean age was just over 19 years old and students reported an average of 150 and 101 total flight hours and total Cirrus flight hours, respectively. Flight instructors had a mean age of around 24 years old, with a median of 21 years of age. Large variances existed in the flight experience categories for this group due to some participants having in excess of 12,000 total flight hours. While the mean total flight time for flight instructors was 2,672 hours, the median was 700. Similarly, for total dual given flight time, the average was 928, with a median of 230 hours. Experience for flight instructors in the Cirrus aircraft had less variance. The average total Cirrus flight experience for flight instructors was 326 hours and total dual given in Cirrus aircraft was 251 hours. On average, instructors held a flight instructor certificate for approximately 7 years. A summary of these demographic findings is provided in Table 1. Participants represented a wide range of certificates held. Approximately two thirds reported holding at least a private pilot certificate with 46% holding instrument ratings, and 29% held Certified Flight Instructor certificates.

Table 1.
Demographic Characteristics of Students and Flight Instructors

	Students				Flight Instructors			
	n	M	*SD*	Median	n	M	*SD*	Median
Age	31	19.71	1.53	19.00	16	24.56	6.51	21.50
Total Flt Hrs.	41	150.48	83.75	150.00	22	2672.77	4803.25	700.00
Total Cirrus Hrs.	40	101.18	46.52	100.00	22	326.09	255.86	285.00
Total Dual	NA	NA	NA	NA	21	928.29	1697.44	230.00
Total Dual Cirrus	NA	NA	NA	NA	22	251.00	250.87	150.00

General CAPS Attitude/Knowledge

The purpose of this survey section was to gather participant's attitudes and knowledge on the general principles behind CAPS. In the current study, the majority of students (95%) and flight instructors (96%), responded that they would use the parachute in an appropriate deployment scenario. However, two students disagreed and one flight instructor strongly disagreed with this statement. When asked if operating an aircraft with an airframe parachute was safer than one without, a higher percentage of students (85%) strongly agreed or agreed with this statement than flight instructors (68%). This finding closely relates to the results found by McMahon (2008) where 77% of respondents indicated a parachute system increased the aircraft's safety.

Flight activity and risk assessment. Questions were also asked regarding incorporation of the parachute into normal flight activity and risk assessment. There were inconsistent responses from participants when asked if a CAPS briefing was included as part of the takeoff briefing. Student and flight instructor responses were very close on this question with 53% of students and 50% of flight instructors strongly agreeing or agreeing with this statement. Thirty-seven percent of students either strongly disagreed or disagreed with this statement, while 10% remained neutral compared to 32% of instructors who strongly disagreed or disagree while 18% remained neutral. When asked to compare risk taking in aircraft equipped with an airframe parachute an interesting dichotomy appeared. The majority of participants indicated that *they* would not assume more risk when flying a parachute equipped aircraft with both students and instructors, 83% and 82% respectively, strongly disagreeing or disagreeing with this statement. However, when asked is they thought *others* would assume more risk only 29% of students and 37% of instructors strongly disagreed or disagreed with this statement. This finding suggests that participants may

believe the parachute system would cause others to take more risks, yet they do not believe this bias would impact them.

CAPS training. When asked if they had received adequate training on when and when not to use the airframe parachute, 76% of students strongly agreed or agreed slightly more compared to 63% of the flight instructors. Both groups felt, however, that training is necessary when flying an airframe equipped aircraft as 100% of flight instructors strongly agreed or agreed with this statement and 91% of the students. This is very different from the results found by McMahon (2008) in which 94% of her grouping of new instructors (those with less than 2 years experience) felt training was not necessary. It is possible that these participants did not have a complete understanding of the parachute system or perhaps they had never flown an aircraft with a parachute, which was not a requirement to participate in the study by McMahon. The current study was also interested in determining if participants had completed training in the advanced aircraft training devices (AATD) that were available at the university to complete parachute deployment scenarios. The findings from this question are mixed yet seem to indicate that many participants did not complete training in the AATD. Of the students, 69% answered strongly disagree or disagree and 63% of the flight instructors. More instructors, 37% agreed or strongly agreed that they had completed AATD training, but only 15% of students shared this feeling; a few students, 17%, neither agreed nor disagreed with the statement.

Students were mixed in their response to whether they would be fearful of damaging the aircraft from a CAPS deployment with 44% strongly disagreeing or disagreeing while 42% strongly agreeing or agreeing. Some students, 15%, remained neutral. More flight instructors, 77%, strongly disagreed or disagreed with this sentiment, 19% strongly agreed or agreed, and 5% remained neutral. For the student group, this question produced a significant negative correlation between the support they felt they would receive from the flight training program if they used the parachute system, $r = -0.311$, $p = 0.024$, which suggests that the more fear students had of damaging the aircraft, the less they felt the flight training program would support their decision to deploy the parachute. Flight instructors experienced an opposite, yet insignificant relationship, $r = 0.219$, $p = 0.094$. Flight instructors were much less fearful of damaging the aircraft from a CAPS deployment than were students. Additionally, 73% of flight instructors felt their decision to use the parachute would be supported by the flight training program compared to 63% of students.

Scenario-Based Examples

A series of 11 scenario-based questions were asked of participants to gauge their willingness to use the airframe parachute system in real-world situations, using a strongly disagree to strongly agree scale. In all scenarios, participants were instructed to identify those situations in which they would *use* the airframe parachute. They were also instructed to assume they were flying a Cirrus SR20 G3 aircraft and out of gliding distance to an airport. The use of an airframe parachute will always involve some subjectivity on the decision-making of the pilot, however, based on these recommendations from Cirrus aircraft and the aircraft operator, there should become clear situations that favor the use of CAPS and clear situations when traditional pilot techniques, such as a forced landing may be more appropriate. Guidance on the correct response was provided from the Cirrus SR20 Pilot's Operating Handbook (POH), and through experts at Cirrus Aircraft. Cirrus Aircraft (2008) provides guidance to pilots on which scenarios are appropriate to use CAPS and these scenarios include: a mid-air collision, structural failure, loss of control, forced landing in inhospitable terrain and pilot incapacitation. Additionally, a CAPS deployment is also the required response to recover from a spin or in the event of an aircraft ditching (forced landing in water). Of the 11 scenarios, eight were designed to favor a CAPS deployment, while three were considered to not be CAPS usage scenarios.

Time of day. Three of the scenarios involved emergencies that sought to identify how time of day may influence the decision to deploy the airframe parachute: two involving engine failures and one complete electrical failure. Based on responses to these scenarios, whether day or night conditions prevailed influenced participant's decision regarding CAPS. When presented with an engine failure over Indiana during daytime conditions, no students or flight instructors agreed or strongly agreed that a CAPS deployment should be completed. Similarly, in a scenario that involved a complete electrical failure in daytime VFR conditions, 100% of flight instructors and 91% of the students strongly disagreed or disagreed with a CAPS deployment. One student agreed, and three students remained neutral. The response offered by participants in these scenarios appears to be in accordance with the Cirrus SR20 G3 POH which states, "if a forced landing is required because of engine failure, fuel exhaustion, excessive structural ice, or any other condition CAPS activation is only warranted if a landing cannot be made that ensures little or no risk to the aircraft occupants" (p. 10-5). However, when presented with a scenario the involved an engine

failure at night over Indiana, participants became more diverse in their responses. While 59% of flight instructors agreed or strongly agreed with a CAPS deployment in this scenario, only 39% of students felt the same. A possible confounding issue with this scenario may be the perceived visual references available to the pilot, which may influence a pilot's decision to use CAPS or attempt a forced landing. Cirrus subject matter experts strongly recommended a CAPS deployment in this scenario.

Location. Two scenarios presented emergency situations over different types of geographical conditions. When asked about an engine failure that occurred over inhospitable terrain, 100% of the instructors and 98% of the students strongly agreed or agreed that a CAPS deployment was the most appropriate outcome. One student neither agreed nor disagreed. When participants were presented with a scenario that involved ditching the aircraft in water, only 73% of the flight instructors and students strongly agreed or agreed that a CAPS deployment was appropriate. These questions provided an interesting finding because both scenarios clearly state, in the aircraft's POH, that a CAPS deployment should be completed.

Weather. Four scenarios presented various weather conditions, with three of them representing an engine failure in IMC, but with varying above ground level (AGL) ceilings: 1,500 AGL, 1,000 AGL, and 500 AGL. The purpose of these question was to attempt and determine how ceiling height would influence the decision-making of the pilots and to understand when most felt the parachute became a better option over a forced landing. It appears that when the ceiling was at 1,500 AGL, participants were divided in their likelihood of deploying the parachute. As the ceiling lowers, more participants indicated they were more likely to deploy the airframe parachute. When descending and the aircraft is under control, Cirrus Aircraft recommends pilot's make the determination to use CAPS no later than 2,000 AGL (2013).

Participants also replied to a scenario that involved a flight in IMC and accumulating ice where they were no longer able to maintain altitude. Students and flight instructors were divided in thirds with approximately one third strongly disagreeing or disagreeing, one third neither agreeing nor disagreeing, and one third agreeing or strongly agreeing. It was unclear from the results why the participants were so divided in their response to this scenario; however, the ceiling provided in this question was 1,500 AGL, which participants may have felt it allowed them time to continue and address the situation. Despite no longer being able to maintain altitude, participants may have felt the aircraft was still controllable. Subject matter experts from Cirrus Aircraft strongly recommended a CAPS deployment to this scenario.

Aircraft control. Two scenarios dealt with aircraft control. When asked regarding loss of aircraft control, the majority of students and flight instructors, 76% and 82% respectively, strongly agreed or agreed that CAPS should be used. However, 12% of the students and 5% of the instructors either strongly disagreed or disagreed with a CAPS deployment in this scenario. While this only represented 5 students and 1 flight instructor, it was concerning to see that when control of the aircraft was lost, a few participants appeared to still be resistant to using CAPS.

Participants were also presented with a scenario that involved an aircraft spin at 1,500 AGL. The Cirrus SR20 G3 POH (2008) states, "the aircraft is not approved for spins, and has not been tested or certified for spin recovery characteristics. The only approved and demonstrated method of spin recovery is activation of the Cirrus Airframe Parachute System" (p. 3-29). While 63% of students and 55% of the flight instructors strongly agreed or agreed with this statement, the other participants were neutral, disagreed or strongly disagreed. In the comments section of the survey one student provided the following comment: "I strongly disagree with using it right away after entering a spin. Altitude permitting, I would make a couple of attempts to recover before pulling the CAPS, regardless of manufacturer or university policy." While this finding is of concern, further research is needed to gain a better understanding to determine if this discovery is representative of the population and as to why some participants appeared reluctant to deploy the CAPS in this scenario despite the manufacturer and POH guidance.

Conclusions, Recommendations, and Limitations

After analysis, the researchers found that there appears to be training discrepancies among participants. These discrepancies appear in the scenario-based examples and the general CAPS knowledge/attitude. Within the CAPS knowledge/Attitudes sections of the survey, the majority of participants indicated that they felt safer in an aircraft equipped with a CAPS system, however participants believed that a CAPS system caused other pilots to assume greater risk. How the participants are reducing their own risk is unknown. Furthermore, there appears to be issues of standardization among participants' standard briefings, specifically the take-off briefing, where participants are not considering briefing the operations of the CAPS system and appropriate use. This issue is further

compounded with participants indicating the need for proper training, yet few participants stating they received AATD deployment training. Further research is needed to determine the effect that non-AATD deployment training has versus AATD deployment training.

When the researchers analyzed the scenario-based examples, several additional examples of training discrepancies were noted. Some of the larger issues arise when there is a disparity between Cirrus Aircraft experts and the participants. Most notably the unwillingness, as indicated by the participants, to deploy the parachute when Cirrus Aircraft experts would strongly recommend a CAPS deployment. When analyzing the responses the main concern is why there is a disparity among highly trained participants in a standardized training program and their willingness to use and understanding of the CAPS system. Since participants indicated the need for training, the researchers would conclude that any training should include a well-established and understood policy on the use and consequences of a CAPS deployment.

With any research there are limitation to that study. The researchers understand that a limited sample from a single training program may affect the generalizability to the whole of training programs. Furthermore, the researchers were unable to match responses of students to their instructors. Therefore, the researchers were unable to determine if the training disparity is isolated to specific instructors teaching their student non-standard procedures. Finally, the researchers understand that a low response rate will influence the analysis of the data.

Acknowledgments

The authors would like to extend thanks to Cirrus Aircraft for their support of this study. Specifically to Robert Haig, Director of Flight Operations, and Travis Klumb, Flight Training Manager, for their help in serving as content experts and validating the survey questions.

References

Blickensderfer, E. L., Strally, S., & Doherty, S. (2012). The effects of scenario-based training on pilots' use of an emergency whole-plane parachute. *The International Journal of Aviation Psychology , 22* (2), 184-202.

BRS Aviation. (2012, May 08). *BRS Aviation: Home*. Retrieved May 08, 2012, from BRS Aviation: http://brsparachutes.com/brs_aviation_home.aspx

Cirrus Aircraft. (2013). *CAPS: Guide to the Cirrus Airframe Parachute System.* Duluth: Cirrus Aircraft.

Cirrus Aircraft. (2008). *Pilot's Operating Handbook and FAA Approved Flight Manual for the Cirrus Design SR20.* Duluth: Cirrus Aircraft.

Cohen, M. S., Freeman, J. T., & Wolf, S. (1996). Metacognition in time-stressed decision making: Recognizing, critiquing, and correcting. *Human Factors , 38* (2), 206-219.

Craig, P. A. (2000). *Pilot in command.* New York: McGraw-Hill.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational Research.* Boston: Pearson.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1997). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. In W. M. Goldstein, & R. M. Hogarth, *Research on Judgment and Decision Making* (pp. 144-180). Cambridge: Cambridge University Press.

Hinton, P. R., Brownlow, C., McMurry, I., & Cozens, B. (2005). *SPSS Explained.* New York: Routledge.

Klein, G. (2008). Naturalistic decision making. *Human Factors , 50* (3), 456-460.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Focus article: Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making , 14*, 331-352.

McMahon, A. (2008). Pilot perceptions on using a ballistic parachute system. *International Journal of Applied Aviation Studies , 8* (1), 156-175.

O'Hare, D. (2003). Aeronautical decision making: Metaphors, models, and methods. In P. S. Tsang, & M. A. Vidulich, *Principles and practice of aviation psychology* (pp. 201-237). Mahwah: Lawrence Erlbaum.

Orasanu, J. (1995). Training for aviation decision making: The naturalistic decision making perspective. *Human Factors and Ergonomics Society Annual Meeting* (pp. 1258-1262). San Diego: Human Factors and Ergonomics Society.

Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols and other distinction in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics , SMC-13* (3), 257-266.

Summers Halleran, M., & Wiggins, M. E. (2010). Changing general aviation flight training by implementing FAA industry training standards. *International Journal of Applied Aviation Studies , 10* (1), 117-130.

Vidulich, M. A., Wickens, C. D., Tsang, P. S., & Flach, J. M. (2010). Information processing in aviation. In E. Salas, & D. Maurino, *Human factors in aviation* (pp. 175-215). San Diego: Elsevier.

DETECTING STRUCTURE IN ACTIVITY SEQUENCES: EXPLORING THE HOT HAND PHENOMENON

Taleri Hammack
Wright State Joint Cognitive Systems Laboratory
Dayton, Ohio
John Flach
Wright State Joint Cognitive Systems Laboratory
Dayton, Ohio
Joseph Houpt
Wright State Cognitive Modeling Group
Dayton, Ohio

Can humans discriminate whether strings of events (e.g., shooting success in basketball) were generated by a random or constrained process (e.g., hot and cold streaks)? Conventional wisdom suggests that humans are not good at this discrimination. For example, Kahneman (2011) writes that "the hot hand is entirely in the eye of the beholders, who are consistently too quick to perceive order and causality in randomness. The hot hand is a massive and widespread cognitive illusion" (p. 117). Following from Cooper, Hammack, Lemasters, and Flach (2014), a series of Monte Carlo simulations and empirical experiments examined the abilities of both humans and statistical tests (Wald-Wolfowitz Runs Test and 1/f) to detect specific constraints that are representative of plausible factors that might influence the performance of athletes (e.g., learning, non-stationary task constraints).

Cooper, Hammack, Lemasters, and Flach's (2014) research showed that various types of constraints on binary sequences (illustrated in Figure 1) could be reliably detected by both humans and statistical tests. This study examined both statistical tests and human performance on a success dependent learning constraint that was calibrated to reflect shooting percentages representative of shooting in NBA games.

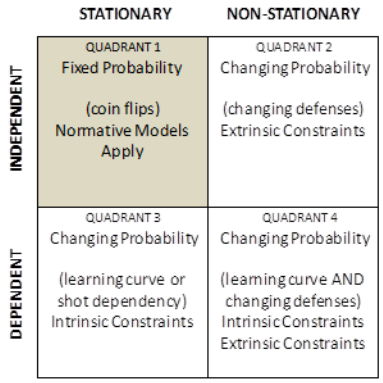


Figure 1. Types of constraints on binary sequences.

Note. This diagram illustrates four types of processes as a function of whether the generating rules are independent and stationary.

Table 1.
Monte Carlo simulation results.

		Runs Test				Frequency Analysis Beta Slope					
	N = 1024		N = 512	N = 800	N = 1024	N = 512			N = 1024		
	Mean	SD	|Z|	|Z|	|Z|	Mean	SD	*t*	Mean	SD	*t*
Quadrant 1											
Bernoulli Processes											
p(hit) = 0.3	0.306	0.013	0.04	0.18	0.44	0.02	0.06	0.87	0.02	0.04	1.50
p(hit) = 0.5	0.500	0.018	0.26	0.06	0.02	0.03	0.05	1.94	0.01	0.06	0.26
p(hit) = 0.8	0.806	0.009	0.62	0.57	0.05	-0.05	0.06	-2.58 *	-0.01	0.07	-0.63
Quadrant 2											
p(hit) = 0.2 and p(hit) = 0.6											
10% chance of alternation	0.419	0.018	3.04 *	3.87 *	4.18 *	-0.23	0.07	-10.75 *	-0.25	0.07	-11.93 *
25% chance of alternation	0.408	0.014	1.81	2.18 *	3.08 *	-0.15	0.09	-5.15 *	-0.18	0.07	-8.54 *
50% chance of alternation	0.401	0.014	0.01	0.21	0.02	0.00	0.06	-0.11	0.01	0.05	0.62
Quadrant 3											
Shot Dependencies											
Last 1 shot dependency	0.498	0.020	8.89 *	11.10 *	12.67 *	-0.55	0.12	-14.73 *	-0.56	0.08	-23.54 *
Last 5 shot dependency	0.422	0.034	3.66 *	4.40 *	5.11 *	-0.26	0.05	-15.46 *	-0.26	0.07	-11.39 *
Simple Learning Curves											
k = 0.001	0.292	0.011	0.84	2.54 *	3.79 *	-0.05	0.07	-2.30 *	-0.08	0.04	-6.94 *
k = 0.003	0.556	0.015	3.10 *	4.68 *	5.70 *	-0.14	0.08	-5.71 *	-0.10	0.05	-6.72 *
k = 0.005	0.642	0.008	4.04 *	5.33 *	5.74 *	-0.14	0.05	-9.33 *	-0.08	0.04	-6.41 *
Quadrant 4											
Simple Learning Curve and p(hit) +/-10%											
k = 0.002 +/-10%	0.391	0.022	1.98 *	2.64 *	3.18 *	-0.16	0.05	-10.23 *	-0.13	0.05	-8.28 *

Note. * p < .05 for two-tailed z-test for runs; * p < .05 for one-tailed t-test for slope (beta). N = 512 and N = 800 indicate that the beginning 512 and 800 (respectively) data points from the 1024 data point sequence were used. Includes the overall mean performance (percent success), the results of the Wald-Wolfowitz Runs Tests across sample sizes, and the mean slopes (betas) and one-tailed t-test results from the spectral analysis across sample sizes.

Table 2.
Monte Carlo simulation results for the performance dependent learning constraint.

				Runs Test					Frequency Analysis Beta Slope			Frequency Analysis Beta Slope		
		N = 1024		N = B.512N = B.800N = M.800N = E.800 N = 1024					N = 512			N = 1024		
		Mean	SD	|Z|	|Z|	|Z|	|Z|	|Z|	Mean	SD	*t*	Mean	SD	*t*
Quadrant 3														
Performance Dependent Learning Curves														
p(hit) = 0.33 and p(hit) = 0.61	k = 0.001	0.380	0.019	0.19	0.16	0.48	0.51	0.40	-0.01	0.08	-0.35	-0.03	0.06	-1.37
	k = 0.003	0.459	0.022	0.33	0.64	0.48	0.30	0.66	-0.03	0.07	-1.15	-0.03	0.06	-1.60
	k = 0.005	**0.499**	**0.020**	**0.33**	**0.48**	**0.24**	**0.48**	**0.75**	**-0.05**	**0.06**	**-2.37 ***	**-0.02**	**0.05**	**-1.64**
p(hit) = 0.20 and p(hit) = 0.80	k = 0.001	0.273	0.013	0.17	0.16	0.08	0.05	0.35	-0.01	0.08	-0.40	-0.02	0.07	-0.80
	k = 0.003	0.425	0.016	0.41	1.17	0.92	1.07	2.10 *	-0.03	0.08	-1.35	-0.03	0.06	-1.86
	k = 0.005	0.529	0.022	1.31	3.11 *	2.56 *	1.81	4.09 *	-0.11	0.06	-5.63 *	-0.09	0.05	-5.53 *

Note. * p < .05 for two-tailed z-test for runs; * p < .05 for one-tailed t-test for slope (beta). B.512 and B.800 indicates the beginning 512 and 800 data points, M.800 indicates the middle 800 data points, and E.800 indicates the end 800 data points from the 1024 data point sequence. Empirical Experiment used trials generated with initial p(hit) = 0.33, asymptotic p(hit) = 0.61, and k = .005. Includes the overall mean performance (percent success), the results of the Wald-Wolfowitz

Runs Tests across sample sizes, and the mean slopes (betas) and one-tailed t-test results from the spectral analysis across sample sizes.

Method for Human Judgment Task

Participants participated in 3 blocks of thirty trials each. In each trial participants were presented with sequences representing binary strings of basketball shots (successes and misses) and were asked to discriminate between two possible generators. In the experiment participants were asked to discriminate between a sequence generated by either a Bernoulli process (a 'steady' shooter with a constant p(hit) = 0.44) or an alternative process governed by a performance dependent learning constraint with a learning rate = .005, a starting p(success) = 0.33, and an asymptote at p(success) = 0.61. Eight hundred shots were available to participants on each trial, however, the number of shots that could be viewed simultaneously decreased from 16 on Block 1 to 1 on Block 3.

Results and Discussion

Table 3.
Empirical experiment results.

	Hit Rate		False Alarm Rate		Adjusted d'		Bias Value	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Window = 16	0.64	0.03	0.33	0.05	0.82	0.17	0.04	0.09
Window = 4	0.67	0.03	0.31	0.05	0.96	0.16	0.03	0.08
Window = 1	0.61	0.04	0.34	0.03	0.70	0.13	0.06	0.07

Note. Mean percentage rate of hits and false alarms, adjusted d', and bias value as a function of window size. Mean and Standard Deviation (SD) were taken across participants.

For the Monte Carlo simulations, constrained by performance dependent learning, the Wald-Wolfowitz Runts Tests and spectral analysis (1/f) showed that none of the sequences used for the empirical experiment were detected as being significantly different from what would be expected from a Bernoulli generated process (with the exception of the spectral analysis using the weakest sample size (N = 512)). Nonetheless, there was information available in the sequences to discriminate between the two generating models as indicated by a Bayes Factor comparisons between the models fit to the generated sequences. However, when the same simulation data was used in a discrimination task done by humans, the results indicated that they *were* able to discriminate the Bernoulli generated sequences from the alternatively constrained performance dependent learning sequences significantly better than chance.

Conventional wisdom classifies the hot hand as an 'illusion' and adds it to the collection of other biases (e.g., gambler's fallacy, availability, representativeness, etc.). However, there is an alternative perspective on human reasoning that has roots in early functional/pragmatic approaches to human cognition. For example, Peirce (1877/1997) offered the construct of *abduction* as an alternative to classical logic. Abduction is an approach to rationality that is grounded in the practical success of beliefs, rather than in the syntax of arguments. More recently, an ecological rationality has been advocated by Gigerenzer (e.g., Todd and Gigerenzer,

2012). From the perspective of Ecological Rationality, heuristics are considered to be analogous to Runeson's construct of *smart instrument.* That is, the use of heuristics reflects an attunement to structure (invariants) in natural ecologies. Thus, it may be an example of an abductive form of rationality that leverages constraints in the problem ecology in ways that support successful adaptations. Thus the belief in the hot hand seems like an effective adaptive strategy rather than neglect of probability theory.

References

Cooper, J., Hammack, T., Lemasters, L. & Flach, J. (2014). Detecting Structure in Activity Sequences: Exploring the Hot Hand Phenomenon. Poster presented at the *International Society for Ecological Psychology North American Meeting,* Miami University, Oxford, OH. (June 5-7).

Kahneman, D. (2011). *Thinking fast and slow.* New York: Farrar, Straus and Giroux.

Peirce, C.S. (1877/1997). The fixation of belief. In L. Menand (ed.). *Pragmatism.* (p. 7 – 25). New York: Vintage Books.

Runeson, S. (1977). On the possibility of "smart" perceptual mechanisms. *Scandinavian Journal of Psychology,* Vol. 18. 172–179.

Todd, P.M. & Gigerenzer, G. (2012). *Ecological rationality.* New York: Oxford University Press.

IDENTIFYING MENTAL MODELS OF SEARCH IN A SIMULATED FLIGHT TASK USING A PATHMAPPING APPROACH

Brandon S. Perelman
Michigan Technological University
Houghton, MI
Shane T. Mueller
Michigan Technological University
Houghton, MI

Aerial assets are often used for missions such as intelligence, surveillance, target acquisition and reconnaissance. The pilot's search decisions reflect a mental model for the search space, including characteristics such as target prioritization, distance-reward evaluations, and path optimization criteria. To investigate differences in these mental models, we examined 23 participants' paths flown in a synthetic task environment in which they piloted a simulated aircraft to search for targets representing missing persons. Determining similarity among flight paths is a challenge. To accomplish this, we used a new tool (Pathmapping, a package in the R statistical computing language; Mueller, Perelman, & Veinott, 2015) to determine area-based path similarities among the test subjects' flight paths, and mixture modeling to analyze those similarities. The results indicate that an area-based measure of path similarity can be used to infer mental models from flight paths produced during a simulated search task.

Search for targets using aerial assets is common across many domains. For example, military pilots and unmanned aerial vehicle (UAV) operators must search for targets during the course of intelligence, surveillance, target acquisition, and reconnaissance operations. In the civilian sector, search and rescue personnel must search multiple locations for missing persons, often with the aid of volunteer pilots. These search operations are conducted by routing search paths through probability maps that indicate locations where the operators expect to find their targets.

Routing a flight path through a probability space is functionally similar to the Euclidean Traveling Salesman Problem (TSP) task, a NP-hard combinatorial optimization problem in which subjects must plot the shortest tour through a Euclidean problem space (e.g., MacGregor & Ormerod, 1996). Applied (e.g., Evers, Dollevoet, Barros, & Mansuur, 2011) and naturalistic (Perelman & Mueller, 2013; Ragni & Wiener, 2012; Tenbrink & Seifert, 2011) versions of this task must often incorporate optimization criteria beyond shortest overall path length which constitute constraints in the problem space, and are sometimes referred to as a Discounted-Reward Traveling Salesman problem (Blum et al., 2007). In addition to the optimization criteria, a particular problem may contain additional constraints. For example, military pilots may want to avoid certain areas due to enemy anti-air assets.

Mental Model Theory (Johnson-Laird & Byrne, 1991) proposes that, in spatial tasks, operators transform all of the constraints into a mental model of the problem space. Empirical explorations of Mental Model Theory (and similar theories, e.g. Preferred Mental Model Theory; Rauh et al., 2005) in naturalistic search tasks are sparse, but a novel experiment by Ragni and Wiener (2012) investigated constraint-based reasoning in a TSP-like problem. Ragni and Wiener presented participants with four types of problems designed to test effects of congruency with participants' mental models of the problem space and the shortest-path solution (which was not always the optimal solution in this constrained paradigm). Methodologically, the authors' problems were relatively simple, consisting of only several nodes, and so determining the number of optimal solutions for each problem was computationally tractable, allowing the authors to use percent correct as an outcome variable. However, it might be useful to make inferences about mental models in more complex environments using a continuous measure of path similarity.

Inferring psychological phenomena from path information is analytically difficult. Existing area-based solutions for comparing two paths (e.g., Asundi & Wensen, 1998; Yanagisawa, Akahani, & Satoh, 2003) are not robust to characteristics such as intersections between the two paths, loops between them, incomplete routes, and self-intersections that may be fairly common in flight paths, especially when pilots orbit a specific area to examine it further. In the present study, we use the Algorithm for finding the Least-Cost Areal Mapping between Paths (ALCAMP; Mueller, Perelman, & Veinott, 2015) to determine correspondence and similarity between participant-generated flight paths to infer their mental models for the search spaces in a naturalistic TSP.

Preferred Mental Model Theory (Rauh et al., 2005) suggests that participants would use the instructions to constrain their optimization criteria used to solve the problem. These mental models would be reflected in the routes they planned through the problem spaces. Based on the routes participants draw, it should be possible to make inferences about their mental models using the pathmapping approach.

In this paper, we will explore and test methods for using model-based clustering to identify clusters of similar paths within unlabeled exemplars. The goal of this approach is to identify the underlying goals and mental models of a search path from a bottom-up perspective. Success of this method has potential applications in aviation training, monitoring, and analysis.

Method

Participants & Apparatus

We recruited 23 participants from the Michigan Technological University undergraduate participant pool who completed 18 naturalistic TSP problems according to two different sets of instructions. In one set of instructions (TSP), participants were told that they were planning the route for a food delivery UAV, and that they should draw a route that minimizes path length in order to minimize fuel usage. In the second set of instructions (Search), participants were told that they would be repurposing the UAV to search for a missing child, and to plan a route that allowed them to minimize the time it would take to find the missing child. Problems were designed so that optimal solutions to these two different problems were qualitatively different. Instruction presentation order was counterbalanced across all subjects, and problem presentation was randomized within instruction conditions. Of the 18 routing problems, one was used as a tutorial, and 11 tested preference for probability regions of different densities, and were designed to test hypothesis that are not pertinent to the present study. Of the six remaining routing problems, two were very similar to the tutorial and thus generated near-perfect performance. For these reasons, four of the problems were selected for analysis in the present study.

Procedure

Participants completed the naturalistic TSP coded in the Psychology Experiment Building Language (PEBL; Mueller & Piper, 2014) which roughly approximated flight path planning through a probability map (see Figure 1). The experimental task differed from traditional TSP in that (1) there was no requirement for the participant to return to the starting location, and (2) the starting location was fixed for each trial. Starting locations were indicated by a green dot. Participants plotted a route through the problem space by clicking on each node in sequence, which resulted in a red line segment being drawn between the previously visited and current nodes. Each trial ended when participants had plotted a route through all of the nodes. As with real navigation tasks, participants could not undo their current route and re-plan dynamically.

Results

Preliminary Analysis

The experiment data from the four selected problems resulted in 194 flight paths (23 participants x 4 problems x 2 instructions). First, within each problem, we generated pair-wise divergence measures between all participant-generated paths using the R pathmapping package (Mueller et al., 2015), an implementation of the ALCAMP algorithm in the R Statistical Computing Language (R Core Team, 2013). The package is available for download from https://sites.google.com/a/mtu.edu/mapping/tasks, and via the Comprehensive R Archive Network (CRAN). For a comprehensive description of the ALCAMP algorithm, see Mueller et al. (2015). The divergence measure can typically be interpreted as a distance, such that a divergence of 0 indicates the paths are identical and the deviation is symmetric ($D(a,b)=D(b,a)$). However, we have not assessed whether the measure satisfies the triangle inequality (either psychologically or logically; see Tversky, 1977), so that it is likely to be possible that three paths can be found such that the $D(a,b) + D(b,c) < D(a,c)$.

Because of the potential non-metric aspects of this deviation measure, we wanted to project the deviations into a metric space for more direct analysis. To do this, we used Kruskal's Non-metric Multidimensional Scaling (via the isoMDS function of the MASS package; Venables & Ripley, 2002). We examined solutions of several

different dimensionalities, but settled on two-dimensional solutions for ease of visualization, as higher-dimensional solutions typically produced similar results. Finally, we then performed a model-based clustering using a custom mixture-of-gaussians driver implemented via the flexmix package for R (Leisch, 2004; Gruen & Leisch, 2007; 2008). Stepwise flexible mixture modeling iteratively fits points to Gaussian clusters using the expectation-maximization (E-M) algorithm, and does so using a range of cluster numbers specified by the user. The ideal model (i.e., number of clusters) for each space was determined to be the number which produced the lowest Bayesian Information Criterion value. E-M is a deterministic process, but it is not guaranteed to find the global optimum. Consequently, we computed each solution from 500 randomly-chosen starting configurations and chose solutions with maximum likelihoods from each run.

We developed a custom mixture-of-gaussians model for this application. Because the axes in the metric solution produced by MDS are not meaningful, we estimated bivariate gaussian distributions with equal variance along both dimensions, and no covariance. Variance for each cluster was assessed independently, so that a tight cluster of solutions might be represented by a gaussian having very small variance (and thus high likelihood). As such, it is typical for solutions to produce several small gaussian clusters and one larger catch-all that captures outliers and other atypical solutions. Finally, there are often occasions in which a set of identical solutions is produced on different trials. These solutions are located at the same MDS coordinates, and if there are many such trials, E-M will tend to remove even close outliers from this cluster and retain only the identical set. This produces a cluster having variance 0, which leads to an undefined likelihood. Consequently, we restricted the minimum standard deviation of any group to be scaled to be 2% of the maximum range of the data in either direction.

To visualize the results of the above analysis, Figure 2 plots the paths belonging to each cluster separately over the points in each problem space. For Problem 2, the shortest-path solution is the one produced by all participants in group 5, which comprised 21/28 of the solutions for that instruction. The solutions to the search instructions produced three distinct sub-groups, each of which handled the bottom ten points differently. Problem 4 produced three groups: one large group (3) mapping onto a shortest-path solution (which was produced by most participants in the shortest-path instruction and about 1/2 of the participants in the search instructions) and two alternatives, both of which skipped some early locations in order to search higher-density legs earlier. Problem 5 produced three tightly-clustered groups: Group 1 was followed primarily on trials using shortest-path instructions, Group 2 and 5 which both were followed on trials using the search instructions, and one catch-all group. Finally, Problem 7 produced one group (3) that solved the two topmost clusters first (followed under the search instructions), two groups (1 and 2) that solved the problem along the linear sequence (one left and one right), followed primarily under the shortest-path instructions, and one catch-all group (4).

This analysis illustrates that in response to different instructions, participants change their search strategies in order to be sensitive to different constraints. However, it also shows that the execution of these problems is not always sensitive to instructions, and that even within a given instruction there are typically several distinct solutions. We infer that these solution map onto different mental models of the search process.

Discussion

The method described here provides an application in using the pathmapping techniques described by Mueller et al. (2015) in discovering underlying mental models of search. This demonstrates that the basic pathmapping technique can be used to assess path similarity and identify sets of similar paths. We suggest several applications of this method within the aviation community.

UAV Tracking and Analysis. Commercial adoption of UAVs has been limited because of FAA's cautious stance on permitting use by amateurs and outside of line-of-sight. One of the problems is that, unlike commercial aviation, where there is strong oversight, predictable and restricted flight routes, and relatively few air vehicles to monitor, commercial UAV applications will decrease oversight and predictability while increasing the number of vehicles by an order of magnitude. Methods such as the one we adopted here may prove useful for analyzing proposed flight routes to against a database of past flights, to better assess the likelihood of other air vehicles being in use in the planned area.

Pilot training. US Navy pilots must be capable of operating aircraft under harsh and uncertain conditions, executing landings on moving aircraft carriers and flying in formation under limited visibility conditions. In a 2011

Edge interview, Gary Klein relates a story in which a Navy pilot used to flying F4s is unable to adjust his mental model to allow him to safely land the newer A6. The pilot's mental model was reflected in an angle misjudgment arising from each plane's seating configuration (i.e., tandem vs. side-by-side, respectively). By measuring divergence between the optimal vs. actual landing or flight trajectories, the present approach would allow trainers to diagnose pilots' errors in landing and formation flying.

Search and Rescue Planning. In search and rescue operations, search operators generate probability maps that incorporate characteristics of the terrain, weather, and the missing person. Modeling missing person behavior is an art unto itself, and these efforts often include factors such as physical fitness, wilderness experience, and clues left in the environment. Systematic deviations from the optimal path out of a wilderness area may give search operators insight into the missing person's psychological state, the state of his equipment, or state of health. Based up testimony from prior recovered missing persons, and their reported trajectories, search operators will have a better picture of where a lost person may travel, given a particular mental model of the environment.

Anomaly Detection. Real-time trajectory-based anomaly detection algorithms have been previously applied to detect illicit activity among taxi cab drivers (Chen et al., 2012). While the intent in that domain was to catch fraudulent taxi cab drivers, in the aviation domain anomalous activity might represent failure in instruments or communications devices, or more sinister activity such as a hijacking. Computing a plane's divergence from its planned flight route would permit anomaly detection in real-time as the ALCAMP algorithm is robust to differences in path length (i.e., incomplete routes).

References

Asundi, A., & Wensen, Z. (1998). Fast phase-unwrapping algorithm based on a gray-scale mask and flood fill. *Applied Optics, 37,* 5416–5420.

Blum, A., Chawla, S., Karger, D. R., Lane, T., Meyerson, A., & Minkoff, M. (2007). Approximation algorithms for orienteering and discounted-reward TSP. *SIAM Journal on Computing, 37,* 653-670.

Brockman, J. (editor) & Klein, G. (interviewee). (2011). Insight: A conversation with Gary Klein. Retrieved from Edge.org website http://edge.org/conversation/insight.

Chen, C., Zhang, D., Castro, P. S., Li, N., Sun, L., & Li, S. (2012). Real-time detection of anomalous taxi trajectories from GPS traces. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services* (pp. 63-74). Springer: Berlin, Heidelberg.

Evers, L., Dollevoet, T., Barros, A. I., & Monsuur, H. (2012). Robust UAV mission planning. *Annals of Operations Research, 222,* 293-315.

Gruen B. & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis, 51,* 5247-5252. doi:10.1016/j.csda.2006.08.014

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11*, 1-18.

MacGregor, J. N. & Ormerod, T. C. (1996). Human performance on the Traveling Salesman Problem. *Perception & Psychophysics, 58,* 527-539.

Mueller, S. T. (2013). PEBL: The psychology experiment building language (Version 0.13) [Computer experiment programming language]. Retrieved from http://pebl.sourceforge.net.

Mueller, S. T., Perelman, B. S., & Veinott, E. S. (2015). An optimization approach for mapping and measuring divergence and correspondence between paths. *Behavior Research Methods.*

Perelman, B. S. & Mueller, S. T. (2013). Examining memory for search using a simulated aerial search and rescue task. In *Proceedings of the 17th International Symposium on Aviation Psychology*, OH.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Ragni, M. & Wiener, J. M. (2012). Constraints, Inferences, and the Shortest Path: Which paths do we prefer? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Sapporo, Japan.

Rauh, R., Hagen, C., Knauff, M., Kuss, T., Schlieder, C., & Strube, G. (2005). Preferred and alternative mental models in spatial reasoning. *Spatial Cognition and Computation, 5,* 239-269.

Tenbrink, T. & Seifert, I. (2011). Conceptual layers and strategies in tour planning. *Cognitive Processes, 12,* 109-125.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Yanagisawa, Y., Akahani, J.-i., & Satoh, T. (2003). *Shape-based similarity query for trajectory of mobile objects*. In Mobile data management (pp. 63–77). Springer.

Tables and Figures

Table 1.
Mean pairwise areas computer between optimal and participant-generated paths, given TSP and Search instructions. (TSP – Search) indicates M, (SD) divergence (pixels / 1000) between the optimal route, given TSP instructions, and the participant-generated routes, given Search instructions. Results of one-sample t-tests, difference scores for the area comparisons with each optimal trajectory, shown below each problem's descriptive statistics.

Problem	TSP Instructions			Search Instructions		
	TSP Optimal	**Search Optimal**	**Proportion Correct**	**TSP Optimal**	**Search Optimal**	**Proportion Correct**
V with Clusters	96.12 (86.04)	126.94 (62.95)	15/28	100.89 (54.93)	58.71 (59.31)	20/28
	$t(27) = 0.18$, $p = .857$			$t(27) = -3.05$, $p = .005$		
Z	0 (0)	39.07 (59.78)	28/28	46.41 (3.33)	40.18 (33.10)	15/28
	$t(27) = 73.67$, $p < .001$			$t(27) = 0.12$, $p = .905$		
Loop	77.84 (137.33)	155.32 (134.79)	21/28	91.40 (115.60)	142.17 (132.50)	24/28
	$t(27) = 3.30$, $p = .003$			$t(27) = -4.89$, $p < .001$		
Z with Clusters	14.87 (27.33)	178.70 (73.74)	27/28	213.92 (28.52)	91.79 (87.59)	17/28
	$t(27) = 19.71$, $p < .001$			$t(27) = -3.07$, $p = .005$		

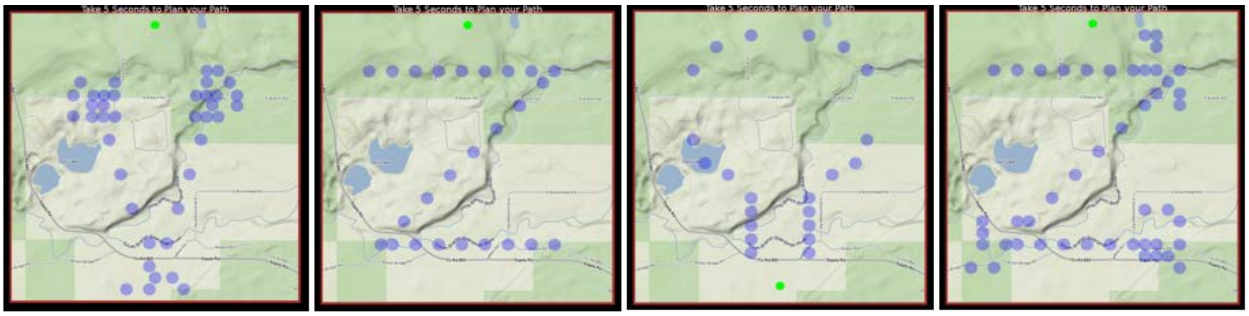


Figure 1. Selected problems from the study. Nodes are shown in blue, while the starting location is shown in green.

The distributions were mathematically validated for two solutions – one optimizing for path length (i.e., the shortest path solution) and the other for estimated time to find (i.e., the shortest average distance between nodes). Path length optimization solutions for these distributions exclude crossovers, whereas optimizing for estimated time to find permits solutions with crossovers while prioritizing clusters of nodes early in the flight path.

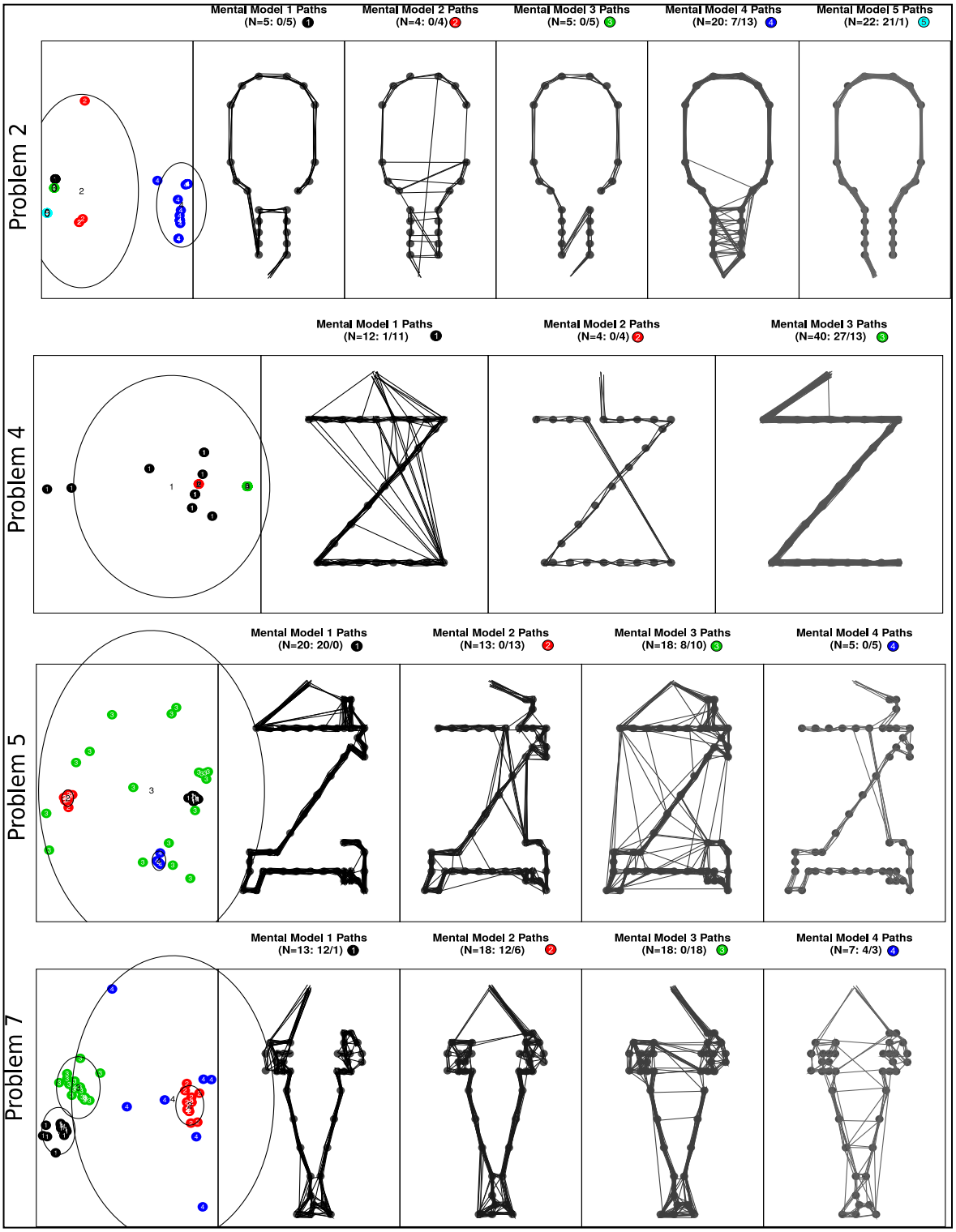


Figure 2. Mixture modeling results for four candidate problems. Leftmost panel in each row shows the isoMDS solution, and remaining panels show clusters of solutions (plotted with jitter). Title indicates the number of solutions in cluster, along with the breakdown between the two instructions. Details of each solution are discussed in the text.

MULTI-GAIN CONTROL: BALANCING DEMANDS FOR SPEED AND PRECISION

Lucas Lemasters
John Flach
Wright State University
Dayton, Ohio

Woodworth's Two-Component model (1899) partitioned speeded limb movements into two distinct phases: (1) a central ballistic open-loop mechanism and (2) a closed-loop feedback component. The present study investigated the implementation of multi-gain control configurations that utilized separate gain values optimized for each movement phase. A target acquisition task using Fitts' Law (1954) was performed within a virtual environment using multiple control devices with three gain settings: (1) mono-gain, (2) dual-gain, and (3) continuous gain. It was found that dual-gain and continuous gain configurations yielded lower movement times and information-processing rates than the mono-gain configurations. The secondary gain values presented in the dual-gain and continuous gain configurations were reported to mitigate oscillations around smaller targets that were responsible for additive settling time. Therefore, implementation of multi-gain control logic could help improve performance when navigating through large spaces and acquiring small targets.

Woodworth (1899) pioneered early research in manual control by examining speed, accuracy, and movement characteristics in continuous voluntary movements (Flach & Jagacinski, 2003). He was able to measure spatial accuracy, consistency of movements, and spatiotemporal characteristics of trajectories using a reciprocal pointing task (Elliot, Chua, & Helsen, 2001). Utilizing these metrics, he observed that for initial aiming attempts, the first portion of the limb movement was generally a rapid and uniform approach to the target. However, as distance to the target decreased, movement became slow, broke off into small sporadic adjustments in position, and finally stabilized on the target.

From these observations, Woodworth hypothesized a two-component model of goal-directed aiming where the control of speeded limb movements consisted of two distinct phases: (1) a central open-loop mechanism followed by (2) a closed-loop feedback-based component (Elliot, Chua, & Helsen, 2001). In Phase 1, an initial ballistic response maneuvers the limb into the vicinity of the target area. Once in the region, the limb comes under feedback-based control (Phase 2) where visual information regarding limb and target position is used to make adjustments in movement trajectories (Elliot, Chua, & Helsen, 2001).

For two-component based control systems, the standard has been to pick a gain that compromises between speed and precision (Kantowitz & Sorkin, 1983). In the present study, by examining each component—open-loop and closed-loop—we can independently optimize their gains based on stability constraints. A high gain is appropriate for the open-loop ballistic phase as it will get to the target vicinity faster. A lower gain is needed for the closed-loop mechanism to emphasize precision and make fine adjustments. Independent gain values for each movement phase should afford a more accurate, time efficient control system.

Preliminary testing supported this notion. In a condition using an Xbox controller with a single gain value (mono-gain), it was found that the value was set too high for the secondary closed-loop control phase. Participants could not get the cursor to stop oscillating and settle in on the smaller targets. They lacked the necessary precision to complete the task in a timely manner. Movement phases could no longer be efficiently controlled separately. However, once a dual-gain configuration was implemented, vast improvements in performance were observed as well as diminished oscillations. Thus, we found a way of getting around this compromise between speed and precision. Instead of having a single "optimal" gain, we introduced two different gain values for each movement phase that could be used at the discretion of the user.

The purpose of this experiment was to expand on preliminary findings in terms of multi-gain control. The performance of three continuous movement devices, each with three gain configurations, was examined. The three gain configurations used were (1) mono-gain, (2) dual gain, and (3) continuous gain. It was hypothesized that the multi-gain configurations (dual and continuous) would yield better performance than the mono-gain configuration. The transference of multi-gain control logic to each device was also examined. Subjects used each device with every gain configuration to complete a series of target acquisition tasks.

Method

Participants

Five participants (aged 23-36) working for Wright Patterson Air Force Base in Dayton, Ohio were used as participants. The participants were from the Human Effectiveness Directorate and worked in joint partnership with Wright State University on this project. Participants were not compensated and willingly participated.

Apparatus

Three devices were used: an Xbox 360 controller, Samsung Slate Tablet, and THRUSTMASTER Hotas Warthog Joystick and Throttle. The gain values available to participants ranged from 10 to 40 based on their configuration. The initial gain value for every device was set at 40. The mono-gain configuration was fixed at the initial gain. For the dual gain configuration, only the initial and lowest gain values were available and could be toggled back and forth using the device mechanisms. For the continuous gain configuration, the whole range of values (10-40) were available and could be scanned as a function of controller displacement (i.e., slowly depressing or releasing the Xbox trigger).

For the Xbox controller, the left thumbstick maneuvered the cursor around the screen. The left trigger served as the gain adjustor. Depressing the trigger gave access to the secondary gains. The Samsung Slate tablet had a first-order control scheme. A center crosshair was implemented at the intersection of four quadrants on the tablet display. Participants were required to drag their finger outward in any direction from the crosshair in order to move the cursor position. The three gain configurations were set up as follows: (1) Mono-gain—fixed at 40; (2) dual-gain—a button on the tablet display initiated the lower gain (10) when pressed and held; (3) continuous gain—a sliding scale adjusted values within the set parameters (10-40).

For the joystick and throttle, one hand was used to manipulate gain values on the throttle while the other hand was used to maneuver the cursor via the joystick.. For the dual gain, pushing the throttle to the most forward position initiated the lower gain while the opposite executed the highest gain. For the continuous gain, as the throttle was pushed forward, gain lowered as a function of displacement. And conversely, as the throttle was brought back, gain increased.

Procedure

The study took place inside a virtual environment. The environment was completely immersive with six walls and overhead projection panels. For this study, only 180 degrees of the environment was used. A gridded virtual landscape was displayed in front of the participant. Cursor position was indicated by a red dot. An equivalent number of targets of different widths and distances were systematically scattered around the space at varying angles of azimuth and elevation. Targets appeared at random within the environment.

Participants were seated in front of the virtual display. The control mappings and gain configurations for the device at hand were explained. After being briefed about the device, participants were given instructional steps related to the task: (1) visually search and locate the target; (2) activate the stationary home button at the bottom of the display using the cursor; (3) immediately drag the cursor onto the target to acquire; (4) repeat.

Design

Ten practice acquisitions were administered in order to get the participants accommodated to the task, device, and gain configuration; 300 recorded trials followed. Participants ran through nine conditions (3 devices x 3 gain configurations) twice for a total of 18 sessions. Each session took place at intervals of at least fifteen minutes to several days apart. Movement times (ms) were measured as the time from home button activation until target selection. Min, max, and mean movement times for each session were recorded for further analysis.

Fitts' Law was used to model human performance and compute information-processing rates. The angle of displacement from the home button to the target served as a measure of amplitude. Width was a measure of visual angle produced by each target. Information-processing rates (ms/bit) were computed for each block by plotting indexes of difficulty as a function of movement time for each trial. Gain manipulation histories were also recorded.

Results

Four 3 x 3 x 2 within-subjects repeated measures ANOVA's were conducted. The first factor was device, the second was gain configuration, and the third was block. An ANOVA was done for each of the following: (1) mean mimimum times, (2) mean maximum times, (3) mean movement times, and (4) mean information-processing rates. Mean values were calculated across participants by factor.

For minimum movement times, there were no significant main effects for device, gain configuration, or block. For maximum movement times, there were significant main effects for gain configuration ($F(1.014, 4.057) = 10.373$, $P < .05$) and block ($F(1, 4) = 33.099$, $P < .05$). The mono-gain configuration yielded the highest maximum movement times ($M = 15920.328$). The dual-gain yielded the second highest ($M = 7426.042$), and the continuous gain yielded the lowest ($M = 6712.310$). Block 2 ($M = 9104.156$) yielded better performance than Block 1 ($M = 10934.964$). There was a significant interaction between device and gain configuration ($F(4, 16) = 3.526$, $P < .05$).

For mean movement times, there were significant main effects for device ($F(2, 8) = 25.953$, $P < .05$) and gain configuration ($F(1.062, 2.247) = 13.197$, $P < .05$). The Xbox controller ($M = 2726.232$) and joystick ($M = 2965.792$) yielded significantly lower mean movement times than the tablet ($M = 3661.478$). For gain configuration, dual-gain ($M = 3003.938$) was significantly better than the mono-gain ($M = 3504.531$). Continuous gain ($M = 2845.033$) was significantly better than both. *Figure 1* shows average minimum, mean, and max movement times across participants for each block, gain configuration, and device.

For information-processing rates, there were significant main effects for gain configuration ($F(1.039, 4.156) = 9.042$, $P < .05$) and block ($F(1, 4) = 28.238$, $P < .05$). There was also a significant interaction between device and gain configuration ($F(1.625, 6.498) = 4.498$, $P < .05$). The dual-gain configuration ($M = 412.473$) yielded lower rates than the mono-gain configuration ($M = 565.792$). The continuous gain configuration ($M = 367.985$) yielded the lowest rates. Block 2 ($M = 431.540$) yielded significantly better performance than Block 1 ($M = 465.960$). *Figure 2* depicts mean information-processing rates across participants for each factor.

Discussion

Findings supported our hypotheses. Multi-gain configurations yielded lower mean and max movement times as well as lower information-processing rates. The continuous gain configuration performed the best. The dual-gain configuration performed second best and the mono-gain configuration performed the worst. Because of the sharp decrease in maximum movement times observed in the multi-gain configurations, we persist in that large max times in the mono-gain conditions were a result of excessive oscillation around the smaller targets. This problem was mitigated by the dual and continuous gain implementation. Secondary lower gain values added a level of precision in the closed-loop component that diminished oscillations and reduced maximum movement times. This reduction also led to lower mean movement times.

Further analysis of how participants used the dual and continuous gain configurations is needed to differentiate control strategies. Participants could have used a bang-bang control strategy in both cases, thus negating the perceived advantage of the continuous configuration. One limitation of our study was that the design was not completely counterbalanced due to time limitations and previously existing data. Thus, some of the variance in performance could be due to practice effects.

The transference of multi-gain control logic across devices was variable. In the Xbox and joystick conditions, we saw sharp improvements in performance going from mono-gain to dual-

gain which did not appear in the tablet. In debriefing the subjects, we found that they were utilizing a different control strategy for the tablet. A "tapping" strategy was used to make small adjustments once in the vicinity of the target. That is, the participant would tap their finger on the display in order to move the cursor in small increments. This method of input allowed them to be more precise when acquiring smaller targets and negated the multi-gain configurations, likely explaining the observed interactions between device and configuration.

In conclusion, multi-gain configurations catering to the movement phases initially described by Woodworth yielded faster, more accurate performance than standard mono-gain setups. However, this multi-gain control logic seems to be limited by device characteristics such as method of input. Further analysis on gain manipulation histories is needed to reveal differences in control strategies.

Acknowledgements

Special thanks to Wright Patterson Air Force Base Human Effectiveness Directorate for assisting me in conducting this study and participating as subjects.

Tables and Figures

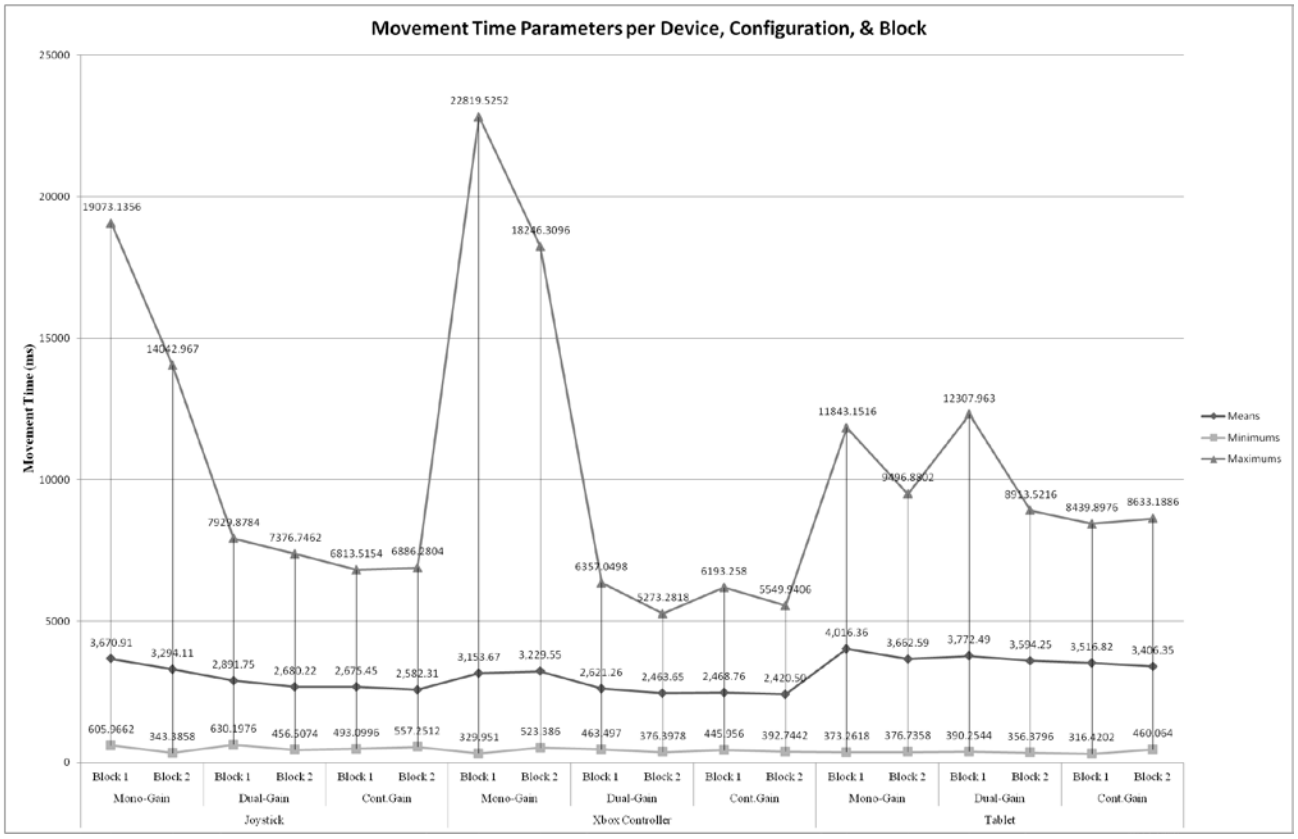


Figure 1. Mean movement time parameters (ms) across participants for device, configuration, and block.

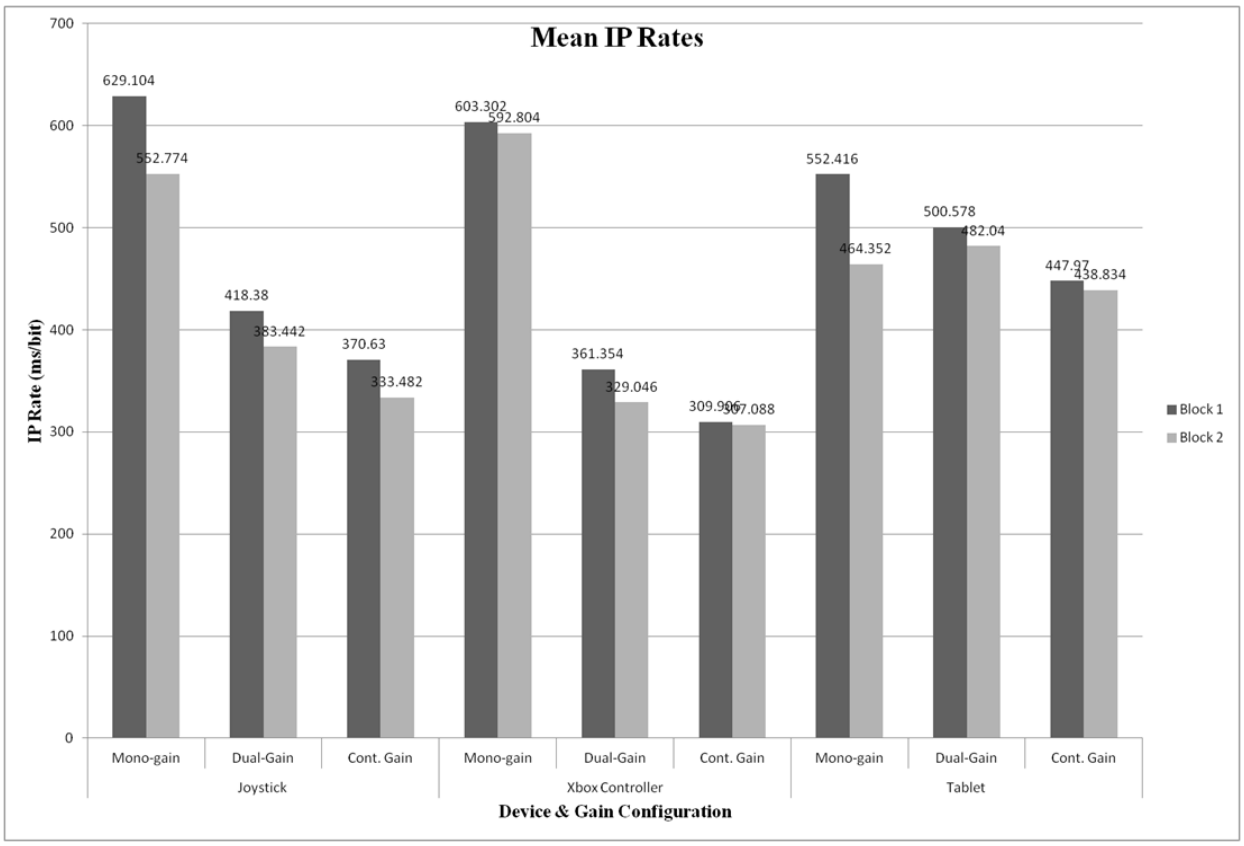


Figure 2. Mean information-processing rates (ms/bit) across participants for device, configuration, and block.

References

Elliott, D., Helsen, W., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological Bulletin, 127*(3), 342-357.

Fitts, P. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: General, 47*(6), 262-269.

Jagacinski, R., & Flach, J. (2003). Information Theory and Fitts' Law. In *Control theory for humans quantitative approaches to modeling performance* (pp. 17-26). Mahwah, New Jersey: L. Erlbaum Associates.

Kantowitz, B., & Elvers, G. (1988). Fitts' Law with an Isometric Controller. *Journal of Motor Behavior, 20*(1), 53-66.

Woodworth, R. (1899). Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements, 3*(3), I-114.

HAPTIC GUIDANCE, INTERACTION BETWEEN THE GUIDANCE MODEL AND TUNING

M. M. (René) van Paassen[1], Rolf P. Boink[1], David A. Abbink[2], Mark Mulder[2], Max Mulder[1]
[1] Aerospace Engineering – Delft University of Technology
[2] Mechanical, Maritime and Materials Engineering – Delft University of Technology
2600 HS Delft, The Netherlands

A haptic interface, also called haptic display, is a system that informs and aids a human operator by forces on the control device (stick, steering wheel or other). These interfaces are being explored for many fields, e.g., for UAV control, (tele-)robotics, automotive control and flying. The force feedback helps in control tasks and increases the operator's awareness. Proper design of such interfaces promotes "shared control", where an autonomous agent and the human operator can jointly exercise control on a dynamic system. The human's flexibility and adaptivity of his neuromuscular system offers ways to override the haptic support, should this be necessary. Haptic interfaces require design decisions on three issues: ***(a)*** The appropriate guidance laws should be developed, thus the behavior of the automated agent must be defined. This guidance should be inherently safe and useful, and it should be compatible with human control strategies, ***(b)*** The guidance should be translated to haptic input on the control device. Here additional force and modification of the control device's apparent properties (mass, damping, spring coefficients) can be used, and ***(c)*** The scaling between the guidance and the haptic input should be tuned to the proper level. From the above, it appears possible to break down the design process into individual steps. However, in a recent research project in which individualized guidance laws were investigated, we discovered an interaction between the guidance laws and the perceived haptic feedback strength, where variation in the guidance laws produced an apparent change in haptic authority by the automation. This paper discusses this experiment – car driving with lateral support – and analyses the causes of the interaction. The results include recommendations for removing this interaction.

Introduction

Recently, an increased interest is signalled for haptic interfaces (or haptic displays) for vehicles. These interfaces use an operator's sense of feeling or touch to display information about the environment or about the device that is being operated. NISSAN for example markets a haptic gas pedal, that can provide force feedback to the driver about obstacles or vehicles detected in front of one's car. In aviation, research has been performed on UAV control and in-aircraft haptic feedback (Lam, Mulder, van Paassen, Mulder, & van der Helm, 2009; de Stigter, Mulder, & van Paassen, 2007; Goodrich, Schutte, & Williams, 2011). When the forces created by the haptic display can influence the input to the controlled system, a *shared control* situation is created. Both the human operator and the system's automation, through the haptic interface, exert an influence on the control input.

The advantages of haptic shared control over conventional assistance by automation are that the actions of the automation are easily observed by the human operator, and, since the display is through the control device, the often overloaded visual channel is not further taxed. However, a shared control situation is in principle still a situation where a human operator is using automation to perform a task. Issues identified in similar situations, such as supervisory control, still apply. Thus reliability of the automation, complacency, (over-)reliance, transparency, and level of automation are relevant issues (Abbink, Mulder, & Boer, 2012). In addition to that, new aspects in shared control are ***(a)*** the continuously variable balance between the human operator's and the automated controller's contributions, and ***(b)*** the fact that the control input is now the sum of the individual inputs of human and automation.

An implication from the first aspect is that the authority of automation versus the authority of the human operator must be made explicit in the design of the haptic device. The choice for device parameters such as stiffness, and the tuning of the device's force feedback, affect this balance. Combined with the fact that the human's neuromuscular system can also adapt, this means that this tuning is not easily arrived at by trial and error. Human adaptability means that a large range of tuning setting produce acceptable behavior for nominal conditions, and an argued choice needs conscious selection of a setting based on neuromuscular system characteristics and task requirements. This issue was explored for a system with haptic feedback for a UAV (Abbink, Cleij, Mulder, & van Paassen, 2012; Sunil, Smisek, van Paassen, & Mulder, 2014).

The consequence of the summing of control forces through the haptic interface means that shared control situations need to be analysed with respect to their control properties. One of the issues is that the control actions by the human and automation must be complimentary, and not counteracting each other. We explored this in a car driving experiment on curve negotiation with haptic support. It is well known that drivers do not follow the center of the road in corners, but slightly "cut" the corners resulting in better driving comfort, and individual differences exist. In our research, we fit a guidance model to drivers' natural preferences, and evaluated the difference between individualised guidance (IG), in which the guidance model was fit to a specific subject's behavior, and a "one size fits all" (OSFA) variant. The surprising result was that many subjects rejected the individualised guidance in favor of the OSFA variant (Boink, van Paassen, Mulder, & Abbink, 2014).

This paper discusses the experiments and investigates the causes for its findings. Then it lists an overview of the design considerations for haptic shared control that were discovered after analysing our results. In addition to the abovementioned experiment, available descriptions in literature of shared-control set-ups are considered and analysed in the light of these design considerations.

Haptic shared control

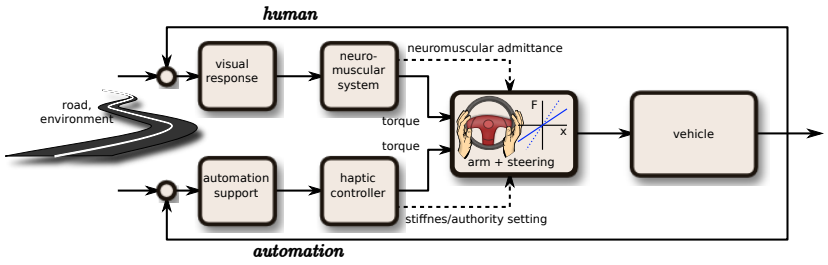


Figure 1. Schematic representation of haptic shared control. Both the automation and human user formulate an input, which is implemented by the combined torque on hands and steering wheel.

In a vehicle with haptic shared control, both the human operator and an automated agent influence the control input; the human through exerting a torque on the steering wheel, and the automation through an additional torque on the wheel from, e.g., an electric motor. In addition, the wheel may have its own dynamic characteristics, typically mass and damping, and the torques on the front wheels (self-aligning torque) are passed through the linkage, resulting in an apparent stiffness of the steering wheel(K_s). Figure 1 depicts this situation. This is effectively the same as the set-up described in Fig. 2 in (Griffiths & Gillespie, 2004), which has a slightly different format for the block diagram, since it expressly shows how the self-aligning torque in a simulation is implemented by the electric motor.

When the steering wheel – or another control device – is held by the human operator, the human's muscular force and the torque from the haptic feedback system act on the combined dynamic properties of that coupled system. A human can generally influence the dynamics of his/her limb, by changing the setting of the neuromuscular system, effectively increasing or decreasing limb stiffness. If properly equipped, the haptic device's stiffness (and possibly damping and mass as well) can be modified in an analogous manner. Such modifications serve to shift the weight of the human contribution to the system input versus the haptic automation's contribution (Abbink & Mulder, 2010); this modulation is indicated by the dashed arrows in Fig. 1.

An important component in the haptic shared control is the generation of the guidance. Two situations are generally distinguished. The shared control may have the purpose of avoiding collisions with obstacles, in that case the haptic display shows *virtual fixtures*, virtual obstacles and boundaries simulated through repulsive forces. In the case of car driving, when only one lane is considered – or a mechanism is provided to detect the desire for a lane change, and the automatic controller can switch lanes – the guidance can be continuous, and the goal of the automation can be defined as keeping the vehicle on an "optimal" track. Rather than virtual fixtures that the vehicle can "hit", a continuous virtual fixture is implemented that pulls the vehicle to a specific target.

For an effective haptic interface, this target should coincide with the driving behavior that a human driver would find acceptable. In curves, assuming a position of the car on the center of the road does not reflect how human drivers will negotiate a curve. In our experiment (Boink et al., 2014), we identified the manner in which drivers

negotiated a curve, and fit this with a simple model that calculates the steering wheel angle given the difference between the nominal track and the lateral position of the car at some look-ahead time t_{LH}:

$$\delta_{wt}(t) = K_\delta E_{t_{LH}}(t) \tag{1}$$

Here $E_{t_{LH}}(t)$ is the lateral error of a predicted position of the car created by integrating a model with the car's current velocity and rotational rate over a prediction time t_{LH}. The K_δ and T_{LH} parameters were identified for each subject individually, and Eq. 1 was used to create the nominal path for the haptic control. In addition a version of the controller was tested which used parameters in the center of the parameter space observed for all participants ("One Size Fits All", indicated with the red circle in Fig. 2).

To convert the nominal path into a guidance force, a scaling gain needs to be determined. Here, this gain is based on the stiffness of the steering wheel (K_s), on the assumption that torque from the haptic feedback system should generate the proper steering wheel angle when the user does not hold the steering wheel:

$$F_{wt}(t) = K_s \left(K_\delta E_{t_{LH}}(t)\right) \tag{2}$$

Experiment

To test our hypothesis that inddividual guidance (IG) would be preferred over OSFA, an experiment with 24 subjects was performed in a fixed-base driving simulator. The simulator was equipped with a Nissan steering wheel actuated by a Moog-FCS ECol-8000 S actuator. In a first session, participants drove a track with alternating left and right curves over 45 degrees, with a 250 m radius. No haptic feedback was provided in this session. A visual scene was projected on the walls in front and to the side of the simulator, providing a field-of-view of almost 180 degrees. The velocity of the simulated car was fixed to 80 km/h. An individualised model (IG) as in Eq. 1 was fit to the data of this run. On the basis of the IG model fits, also an OSFA fit was determined. In a second session, subjects performed runs with haptic guidance, with either the IG or the OSFA tuning. Figure 2 shows the spread of the tuning parameters and the OSFA tuning. After each pair of runs, subjects were asked to indicate their preference for either the first or second run.

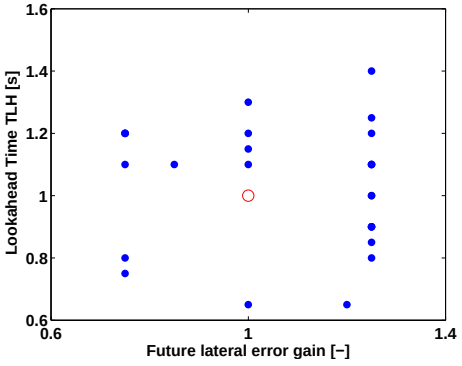


Figure 2. Individual fits of look-ahead time and lateral error gain ($K_S K_\delta$) used in the experiment (Boink et al., 2014).

Results and Discussion

Figure 3 gives an example of a single curve driven by a "medium-gain" subject, in the OSFA condition. In addition, the curve driven by the subject in the absence of haptic support, and the result of letting the haptic support system "drive alone" is given. A number of surprising results can be noted ***(a)*** The lateral error (note that this is the lateral error at the look-ahead point) is fairly small when the human is driving – with or without haptic support –, indicating a successfully driven curve; ***(b)*** When the haptic automation drives alone (hands off steering wheel), the errors are fairly large, indicating that the control law in Eq. 2 is actually not effective; ***(c)*** Finally, the force from the guidance actually seems to oppose the human torque over a large stretch of the curve.This latter result was found with multiple subjects, and often with subjects with curve negotiation behavior that resulted in model fits with high gains and large look-ahead times.

The experiment expressly adressed one of the design decisions in creating a haptic support system, namely the question of *what should be the reference trajectory for the haptic support system*. Rather than taking the lane's center, which would result in unnatural driving behavior, a simple control law is fitted to observed control behavior.

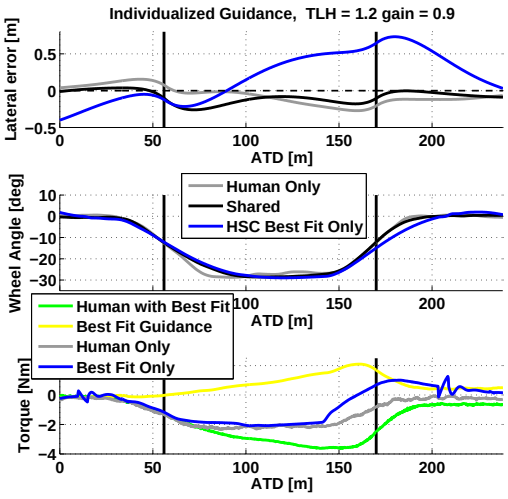


Figure 3. Example run from a subject with and without individualised haptic shared control, illustrating in this case initially no, and later negative contribution of the guidance to the steering wheel torque.

Compared to an older experiment (Abbink, Cleij, et al., 2012), in which the reference was created by averaging a number of previous runs, the present approach is more general.

One of the other surprising results from this experiment was that more subjects preferred the OSFA tuning of the controller over the adaptive tuning. Upon further inspection, it proved that this correlated with the gain of the individual tuning; subjects which had a lower gain, preferred the individualised tuning settings, and subjects with higher gain preferred the OSFA tuning. To investigate possible causes for this, a further analysis of the controller and the haptic feedback it provides was done.

Given that the log data indicates that the haptic guidance often counteracts the human control input, it could be expected that a weaker haptic guidance is to be preferred over a stronger one. This was also found by Mars et al., 2014, although that research in addition forced unnatural curve negotiations, since the haptic feedback was based on the lane center. However, why would the haptic guidance not contribute to the control goal or even counteract the human control? And also, why would haptic guidance alone not create a proper control of the vehicle?

To analyse the effect of the control law, a small-angle approximation is used for the future lateral position error $E_{t_{LH}}$:

$$\tilde{E}_{t_{LH}} = Vt_{LH}\left(\Psi(t) - \Psi_r(t)\right) + \left(y(t) - y_r(t)\right) \tag{3}$$

To determine the effect on haptic feedback, consider a run in which the subject exactly replicates the steering commands, as measured in the runs without haptic support and as captured in the model in Eq. 1. In that case, the lateral error at the look-ahead point ($E_{t_{LH}}$) is minimal; the only source of deviation between the reference model and the user's run would be the remaining variation in the user's driving that could not be captured by the model. According to Eq. 2 the haptic feedback force would be nearly zero in this case. Inspection of a similar architecture in literature (Griffiths & Gillespie, 2004) suggests that the same occurs in that set-up; with successful control by the human, and a zero control error, there is no torque contribution from the haptic support for curve negotiation. This relates to a second design decision that needs to be made, *how much should the haptic support system contribute to the control effort in nominal (no-deviation from target) cases?*. The haptic support system tested in our experiment relied on error between the determined nominal path and the driven path. In this case, subjects who seek support from the haptic system need to allow deviations from the nominal path before getting this support. The hands-free runs in Fig. 3 are an illustration of this point. If the haptic support system needs to supply a contribution to the steering input, it needs separate information from both the target signal and the current error signal; enabling the calculation of separate haptic support torques for following the track (feed-forward of the target signal, block *LoHS* in 4) and for correcting deviations from the track (feedback of remaining execution errors, block *SoHF*).

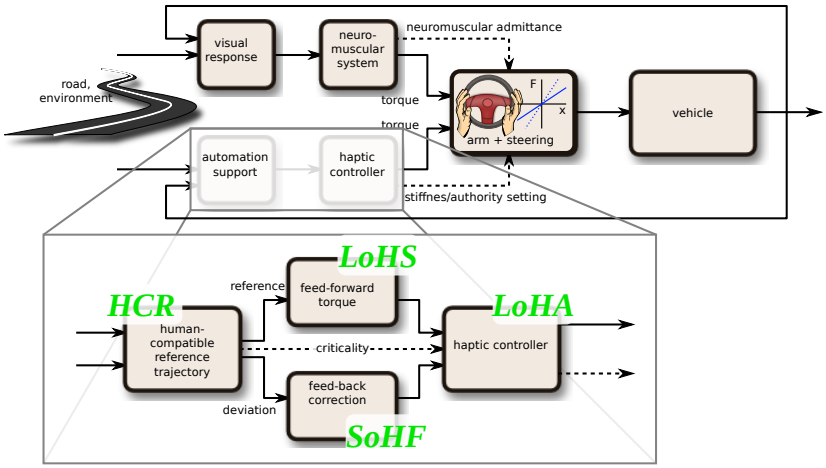


Figure 4. Schematic representation of haptic shared control, detailing the four design choices identified in this paper.

Now consider a lateral error in the car position $y_r(t) - y(t)$. For each unit in lateral deviation, a feedback force of $K_s K_\delta \left(y_r(t) - y(t)\right)$ [Nm] will be generated. A lateral deviation in car heading has a similar effect, now with a gain of $K_s K_\delta V t_{LH} \left(\Psi_r(t) - \Psi(t)\right)$. However, in this case both the gain K_δ and look ahead time t_{LH} are parameters from the individual lateral guidance model fit to the subject's runs without haptic guidance, and not parameters chosen to tune the strength of the haptic feedback. The experiment thus had a confounding effect, in that the preference for curve negotiation influenced the strength of the haptic feedback to deviations from the nominal path. This amounts to the third design decision, *how strong should be the feedback to deviations from the nominal path?* A related decision is the weighting of what constitutes a "deviation". In this case, deviation can be expressed in lateral position and in heading deviation. A control law needs to determine how these are weighed; such a control law depends on the dynamics of the controlled system, required performance and the comfort levels that one wants to attain.

The final design decision concerns the *authority of the haptic controller*. In analogy to the Level of Authority in supervisory control (Parasuraman, Sheridan, & Wickens, 2000; Sheridan & Parasuraman, 2005), Abbink et al. (Abbink, Mulder, & Boer, 2012) coined the phrase Level of Haptic Authority (LoHA). By selecting the control device's stiffness settings – fixed or possibly variable – one can influence the weight of the automation in determining the final control input. Given that the human operator has the means to adjust the settings of their neuromuscular system, this will result in a range of division of LoHA between automation and human.

Note that with a high level of haptic authority, in a system designed without haptic support, the haptic interface will still push the system towards the reference, however before that happens a control error needs to build up in the system, and the path that results will no longer match the subject's curve negotiation strategy. This behavior annoyed some of the subjects in our experiments, since when subjects implement the proper control strategy, the automation does not contribute to the control signal, except to correct any deviations.

Conclusion

Haptic shared control is common practice in training settings in aircraft; typically the instructor's and student's controls in a trainer aircraft are mechanically linked, and a good instructor can make a student feel the necessary inputs, reduce their LoHA – both in generating feed-forward and corrective feed-back inputs, and there is a common and compatible (visual frame of) reference. Such a situation can be seen as a reference for haptic shared control. An implementation of haptic shared control with automation requires that such an instructor's behaviour be made explicit with a number of design choices:

Human Compatible Reference (HCR) Generation of a reference for the control, compatible with user strategies and the device and environment constraints.

Level of Haptic Support (LoHS) A choice for the Level of Haptic Support; i.e., by how much will the automated system contribute to implementing a path that follows the reference (feed-forward).

Strength and Strategy of Haptic Feeback (SoHF) A choice for the strength of the haptic feedback and the control law upon which this feedback is based (in this case, weighing lateral and heading error); i.e., by what control law / aggressivenes will the automation provide corrective inputs to reduce the difference between the reference and the vehicle's path.

Level of Haptic Authority (LoHA) A choice for the level of haptic authority; i.e. how is the balance between human input and automation. A high level of authority is implemented by choosing a large base stiffness of the control device. In that case the feedback and autonomy signals (since they are adapted to the joint control device and human operator stiffness) scale too.

The first and fourth issues have been addressed in literature. Independently tuning the level of haptic support and the strength of the haptic feedback is a step that is still lacking in many designs.

References

Abbink, D. A., Cleij, D., Mulder, M., & van Paassen, M. M. (2012, October). The Importance of Including Knowledge of Neuromuscular Behaviour in Haptic Shared Control. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics* (pp. 3350–3355). Seoul, Korea: IEEE. doi: 10.1109/ICSMC.2012.6378309

Abbink, D. A., & Mulder, M. (2010, April). Neuromuscular Analysis as a Guideline in designing Shared Control. In M. Hosseini (Ed.), *Advances in Haptics.* InTech.

Abbink, D. A., Mulder, M., & Boer, E. R. (2012, March). Haptic Shared Control: Smoothly Shifting Control Authority? *Cognition, Technology & Work*, *14*(1), 19–28. doi: 10.1007/s10111-011-0192-5

Boink, R., van Paassen, M. M., Mulder, M., & Abbink, D. A. (2014, October). Understanding and Reducing Conflicts between Driver and Haptic Shared Control. In C. L. P. Chen & W. A. Gruver (Eds.), *IEEE Systems, Man and Cybernetics Conference* (pp. 1529–1534). San Diego (CA): IEEE.

de Stigter, S., Mulder, M., & van Paassen, M. M. (2007). Design and evaluation of a haptic flight director. *AIAA Journal of Guidance, Control and Dynamics*, *30*(1), 35–46.

Goodrich, K., Schutte, P., & Williams, R. (2011, September). Haptic-Multimodal Flight Control System Update. In *11th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference.* American Institute of Aeronautics and Astronautics. Retrieved 2015-03-02, from http://dx.doi.org/10.2514/6.2011-6984

Griffiths, P. G., & Gillespie, R. B. (2004). Shared Control Between Human and Machine: Haptic Display of Automation During Manual Control of Vehicle Heading. In *12th International Symposium on Haptic Interface for Virtual Engironment and Teleoperator Systems (HAPTICS '04)* (p. 9). Piscataway, NJ: IEEE.

Lam, T. M., Mulder, M., van Paassen, M. M., Mulder, J. A., & van der Helm, F. C. T. (2009, May). Force - Stiffness Feedback in Uninhabited Aerial Vehicle Teleoperation with Time Delay. *Journal of Guidance, Control, and Dynamics*, *32*(3), 821–835. doi: 10.2514/1.40191

Mars, F., Deroo, M., & Hoc, J.-M. (2014, July). Analysis of Human-Machine Cooperation When Driving with Different Degrees of Haptic Shared Control. *IEEE Transactions on Haptics*, *7*(3), 324–333. Retrieved 2015-03-09, from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6710125 doi: 10.1109/TOH.2013.2295095

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *30*(3), 286–297.

Sheridan, T. B., & Parasuraman, R. (2005, January). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129. doi: 10.1518/155723405783703082

Sunil, E., Smisek, J., van Paassen, M. M., & Mulder, M. (2014, October). Validation of Tuning Method for Haptic Shared Control using Neuromuscular System Analysis. In W. A. Gruver & C. L. P. Chen (Eds.), *IEEE Systems, Man and Cybernetics Conference* (pp. 1518–1523). San Diego (CA): IEEE.

DESIGN AND EVALUATION OF A HAPTIC DISPLAY FOR FLIGHT ENVELOPE PROTECTION SYSTEMS

J. Ellerbroek, M. J. M. Rodriguez y Martin, M. M. van Paassen, M. Mulder
Faculty of Aerospace Engineering, Delft University of Technology,
Kluyverweg 1, 2629 HS, Delft, The Netherlands

This paper describes the design and initial evaluation of a haptic display that is aimed to complement a `hard' flight envelope protection system. The evaluation mainly focused on usability of the presented haptic cues, and on the handling qualities of the stick with active feedback. Results are presented for two evaluations, concerning stall protection feedback and load factor protection feedback respectively. They show that while subjects are positive about the added information cue, and are able to correctly identify limiting actions, they are not consistently able to identify changes in the aircraft's condition.

Modern fly-by-wire aircraft, such as the Boeing 777 and the Airbus A380 are equipped with Flight-Envelope Protection (FEP) systems, in order to ensure operation within a specific safe operating domain. These systems are designed to avoid commands that would result in unwanted situations such as stall, over-speed, or excessive load factors. A distinction can be made between so called `soft' protection, where the crew can override the protection system by applying excess force on the controls, and `hard' limits that cannot be overridden (Traverse, Lacaze, & Souyris, 2004).

On the one hand, the arguments for 'hard' envelope protection are clear: excursion of the aircraft beyond these limits leads to unsafe situations that potentially result in structural damage of the aircraft, and can ultimately, lead to unrecoverable loss of control. Indeed, with these flight envelope protection systems, the number of handling and control-related accidents has greatly reduced. On the other hand, extreme maneuvers that take the aircraft beyond the envelope limits can sometimes be necessary as a last resort, where the only alternative is the certainty of a crash. In the China Airlines B747 incident in 1985, for instance, pilots were required to overstress the horizontal tail surfaces to recover from a roll and near-vertical dive (NTSB, 1986). This recovery would have been impossible had a hard envelope protection system been in place.

Similarly, also 'soft' envelope limits have their benefits and drawbacks. While an accident such as the one avoided in the China Airlines incident can, in principle, be avoided with a soft envelope protection system in place, an important disadvantage is that in any situation, pilots have the ability to control the aircraft into dangerous situations. This means that pilots have to be fully aware of the limitations of their aircraft, and experience will play an (even more) important role, especially in non-nominal situations.

The discussion about these two approaches to envelope protection therefore remains a valid and important one, where the optimal solution is likely to lie with a combination of both approaches, rather than one of the above extremes. This report describes an addition to the hard envelope protection system that addresses the problem of lack of Situation Awareness (SA) (with respect to flight envelope limits / limiting) that can occur. An advanced haptic feedback system is proposed, which addresses the communication of flight-envelope boundaries to the pilot, and how they relate to control inputs from the pilot. The haptic system uses force and stiffness feedback to communicate how manoeuvrability is affected by flight envelope boundaries.

Haptic communication of FEP boundaries

In line with Billings' concept of human-centered automation (Billings, 1996), haptic feedback is seen as a way to flexibly share information and control between the human operator and the automation on a physical level (Abbink, Mulder, & Boer, 2012). To address the lack of SA with respect to flight envelope limiting, and more generally to the flight control system state, a haptic display is therefore proposed, that complements the existing `hard' FEP system. The current concept considers longitudinal limits; lateral limits will be added in future iterations.

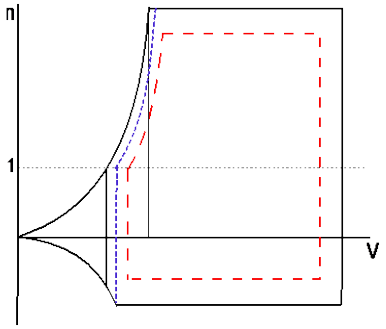


Figure 1. Flight envelope with areas where haptic feedback is active.

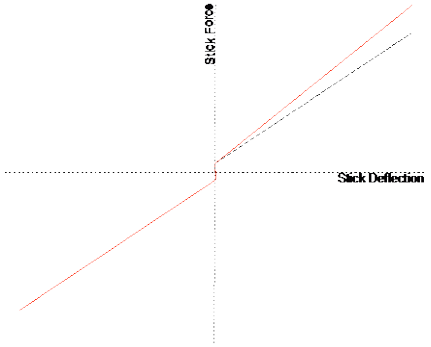


Figure 2. Increased stiffness near stall

Haptic feedback near stall

The haptic feedback provided in near-stall situations is divided into two categories, depending on the severity of the minimum speed incursion. Two areas are defined here, see Figure 1. The inner border (red dashed line) indicates the area beyond which proximity to the stall limit is communicated by increased stiffness on the stick, the effect of which is illustrated in Figure 2. When speed is reduced beyond the second border, (blue dashed line in Figure 1) a vibration (a.k.a. `stick shaker') is felt on the stick.

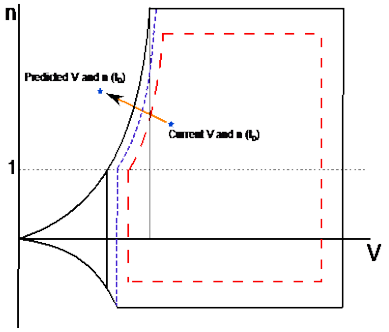


Figure 3. Predicted state outside the envelope.

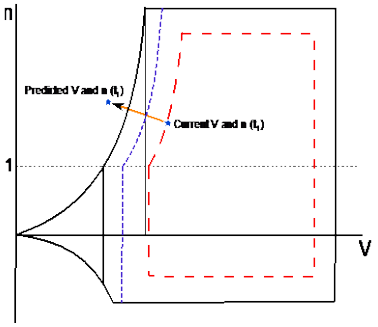


Figure 4. Adapted haptic boundaries.

To be able to translate perceived feedback into a desired action, it is required that the pilot receives the force feedback with sufficient anticipation. If for example the aircraft experiences a sudden increase in load factor while its velocity is rapidly decreasing, the stall speed might be reached very quickly. Examples of such maneuvers are a sustained pull up or a high bank angle coordinated turn. Due to the rapid increase of the stall speed caused by the fast rise in load factor, it is necessary to take into account the time the pilot needs to understand and react to the haptic feedback. To this end, predictions

are made of load factor and speed (see Figure 3) taking into account a certain cognition time. When predictions exceed the envelope limits, the haptic boundaries are shifted to match the aircraft's current velocity, see Figure 4.

In case of a near-stall situation, recovery maneuvers can consist of both reducing load factor and increasing velocity. A change in load factor can be achieved rapidly, by reducing the commanded load factor with the stick. Maintaining a certain load factor and increasing the thrust to gain velocity is a much slower process, and might even be impossible when thrust is at maximum or in case of engine failure. Whether a load factor reduction is sufficient, or whether an increase in speed is required to return to a safe state depends on the location of the unwanted state within the envelope, see Figure 5. Here, case 1 illustrates a situation where a reduction in load factor is sufficient to return to the safe envelope. For case 2 it can be seen that reducing stick output to neutral isn't enough to return to the safe envelope. In addition, a speed increase is required, which can be obtained by a pitch-down command. This is communicated haptically by shifting the neutral point of the stick stiffness curve to the desired deflection, see Figure 6.

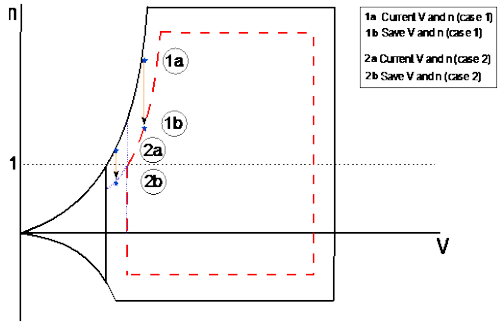


Figure 5. Situations where a load factor reduction is or isn't sufficient.

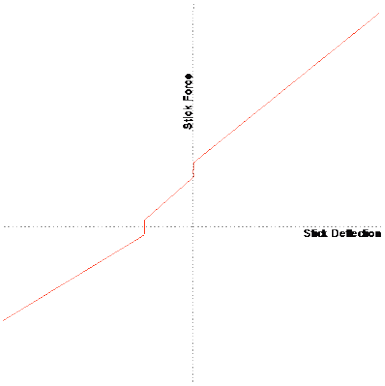


Figure 6. Neutral point shift

High load factor protection

In high load factor situations, an increased stiffness profile is applied which is proportional to the relative proximity of the aircraft state with respect to the applicable load factor limit. The stiffness varies between one times the nominal stiffness at the highest considered commanded load factor where no additional feedback is given, and two times the nominal stiffness when the load factor is equal to either the maximum or minimum allowed load factor. The resulting stiffness profile is similar to the increased stiffness in certain near-stall situations, illustrated in Figure 6.

Other protections

In addition to stall and load factor, Airbus flight-envelope protection systems also implement limitations in near-overspeed situations, and in extreme attitude situations. Both these envelope limitations are implemented in the haptic system using increased stiffness profiles, similar to the near stall stiffness adaptation illustrated in Figure 2. Should the evaluation of the haptic system identify confusion issues due to the similarity of the feedback cues, a future design iteration will investigate possible alternative feedback methods.

Experiment

The experiment has been set-up as a part-task evaluation, a co-pilot was not present during the experiment. To focus analysis on the effects of the haptic feedback system, otherwise present visual and aural warnings regarding the FEP system were absent during the experiment.

In the stall protection evaluation, subjects were requested to perform a wings-level, maximum thrust climb (such as in a go-around). In nominal conditions, the task here is to maintain a constant velocity, controlling the aircraft's pitch angle. Some time after the pilot has reached a stable climb, the stall speed is slowly increased by simulating icing conditions. With (adaptive) envelope protection enabled, the response of the aircraft is to push down its nose.

In the load factor protection evaluation, subjects were requested to find the roll angle that corresponded to the applied load factor limit. Because the current haptic display implementation uses stiffness feedback (i.e., feedback is only felt for non-zero stick deflections), the automatic pitch control during banking was circumvented by applying an additional load factor command.

Two subjects participated in this preliminary experiment, both male, both active A330 captains. Subjects are aged 55 and 49, with an average flight experience of 10,000 hours.

Results

Stall protection evaluation

After establishing a steady, maximum thrust climb, stall speed was increased gradually due to the worsening of aerodynamic properties caused by icing. This increase in stall speed caused the adaptive alpha protection to act, providing a pitch down input (see Figure 7), which both test pilots tried to counter in order to maintain the same pitch attitude, see Figure 8. In other words, neither pilot was able to correctly identify the change in flight envelope. This behavior was consistent between pilots, and between support system conditions. From the questionnaire it became clear that both subjects assumed that the nose down behavior was the result of a reduced elevator authority.

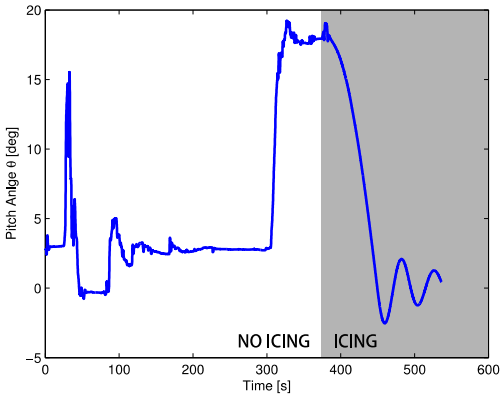


Figure 7. Example of pitch angle before and after icing.

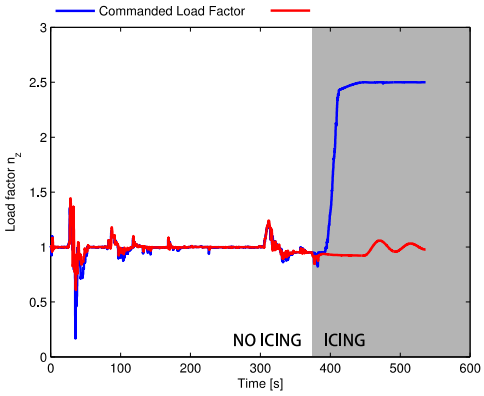


Figure 8. Example of commanded and measured load factor.

The commanded load factor in Figure 8 shows that after the envelope protection system has pushed the nose down, pilots invariably reacted by fully deflecting the side stick contrary to the nose-down command, and maintaining that input. The force-displacement relation for this behavior can be seen in Figure 9, and Figure 10, for the baseline condition and the haptic feedback condition, respectively.

What is clear from these graphs is that the force with which the aft deflection is applied more than saturates the forces applied by both the passive and the active side stick stiffness profiles. This means that any information that is supposed to be communicated through stiffness alterations cannot be detected by the pilot, because he/she is pulling harder than that against the end stop of the side stick.

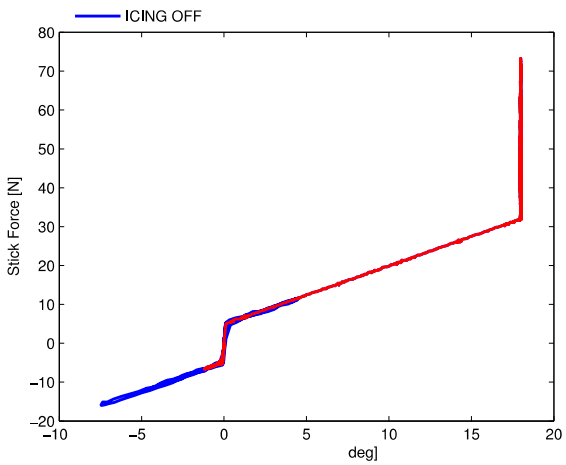


Figure 9. Force-displacement diagram in baseline condition.

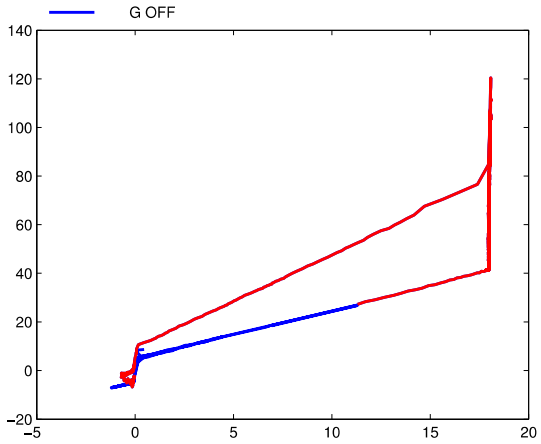


Figure 10. Force-displacement diagram in haptic feedback condition.

Load factor protection evaluation

Load factor limitation evaluation was established by means of a roll angle capture task. The goal in each run was to roll the aircraft, and stabilize at the roll angle where the load factor limitation feedback becomes active. Because this type of evaluation inherently requires some kind of feedback to be present, no baseline condition could be evaluated. The following results therefore only consider the haptic feedback ON condition.

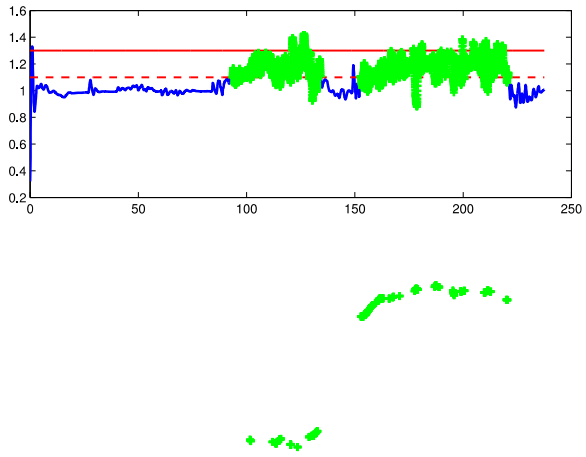


Figure 11. Example of a roll angle capture.

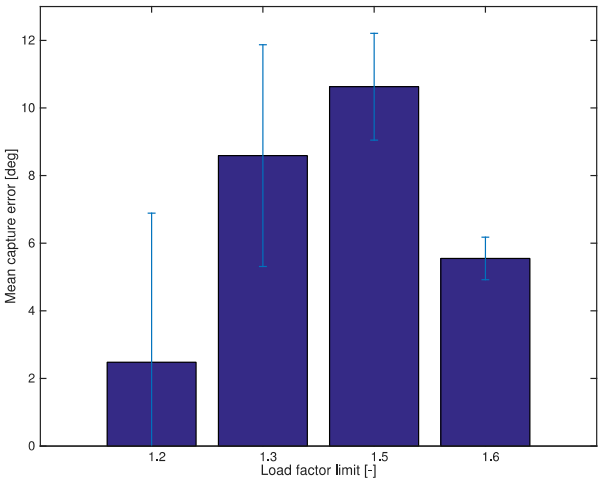


Figure 12. Average capture error.

Figure 11 shows time histories of load factor (top graph) and roll angle (bottom graph), in an example of a roll angle capture task. In this case, a load factor limit of n = 1.3 needed to be identified, corresponding to a maximum roll angle of $\varphi = 40°$, and a protection limit of $\varphi = 25°$. The green sections of the time histories indicate the ranges where the subject has identified the load factor limit feedback. This task has been performed for maximum load factors n = 1.2, n = 1.3, n = 1.5, and n = 1.6 (these correspond approximately with logical roll angles).

Figure 12 shows how well the subjects were able to identify the corresponding roll angles for each of the four load factor limits. From the deviation of the capture error it can be noticed that the subjects managed to perform the task more consistently with higher load factors.

Discussion and conclusions

At first sight, the experiment results seem to suggest that haptic feedback has the potential to improve the pilot's awareness during critical flight conditions. To be able to test this without similar systems confounding the results, this was tested with aural and stall visual warning cues unavailable. The haptic stall protection was tested by means of increasing the stall speed as caused by ice formation. While the haptic cues did work as a good attention catcher to make the pilot aware of the fact that the envelope protection system was acting, subjects could not successfully use this information to identify exactly what was going on. This was amplified by the fact that both subjects saturated the inputs, making it impossible to provide information through stiffness feedback only. A possible addition to combat this would be to look at the combination of stiffness feedback and a discrete force cue, such as a stick shaker. This will be considered in a next iteration of the concept. In the second evaluation experiment, subjects were capable of identifying maximum load factors, by means of levelling at corresponding roll attitudes. In all experiment runs, pilots were able to properly and accurately identify the roll angle corresponding to the maximum load factor.

Finally, both during the experiment, as well as in the post-experiment questionnaire, both subjects indicated that, despite initial scepticism, they were positive about the proposed system, which they considered as having the potential to be a useful addition. A further iteration of the haptic display concept, with an accompanying larger evaluation, are therefore planned to investigate this potential way forward.

Acknowledgements

This research was performed as part of the ACROSS project. This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under FP7-TRANSPORT, grant number 314501.

References

Abbink, D. A., Mulder, M., & Boer, E. R. (2012). Haptic shared control: smoothly shifting control authority? *Cognition, Technology & Work*. doi:10.1007/s10111-011-0192-5

Billings, C. E. (1996). *Aviation Automation: The Search for a Human-Centered Approach* (1st ed.). CRC Press.

NTSB. (1986). *Aircraft Accident Report - China Airlines Boeing 747-SP, N4522V, 300 Nautical Miles Northwest of San Francisco, California, February 19, 1985*. Washington D.C.

Traverse, P., Lacaze, I., & Souyris, J. (2004). Airbus fly-by-wire: A total approach to dependability. In *Proceedings of the 18th IFIP World Computer Congress*. Toulouse, France.

OPEN SOURCE DEVICES FOR HUMAN FACTORS RESEARCH

Kevin M. Gildea & Nelda J. Milburn
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK

The availability of increasingly powerful and versatile open source software and hardware products continues to open new possibilities for the design and development of experimental devices. The declining cost of many proprietary software and hardware solutions has further increased the options available to researchers. These new capabilities have led to an increasing number of people engaging in design and development of devices for research and other purposes. Capabilities that were previously only available to well-funded engineering organizations are now accessible to individuals and small teams with limited resources. The formation and growth of local and online support communities have provided access to existing solutions, guidance, and discussion.

Light emitting diodes (LEDs) are rapidly replacing incandescent sources for aviation signal lighting creating a need to evaluate the implications in terms of human visual processing in color vision normal and deficient individuals. Recent research at the Civil Aerospace Medical Institute (CAMI) evaluated whether individuals with certain color vision deficiencies were able to discriminate between the red and white lights in fielded approach light systems based on luminous intensity, even if unable to detect a difference in color (Milburn et al., 2013; Milburn & Gildea, 2012). Approach lighting systems, including the Vertical Approach Slope Indicator (VASI) and Precision Approach Path Indicator (PAPI) lighting systems present pilots on final approach to runways with a visual indicator of their height—either above or below—an optimal glide slope. Combinations of red and white lights convey visual glideslope information. For this experiment, we designed the device to ascertain whether individuals with certain color vision deficiencies were able to discriminate between the red and white lights in fielded approach lighting systems based on luminous intensity (Gildea & Milburn, 2013; Milburn, Gildea, Perry, Roberts, & Peterson, 2014; Figure 1).

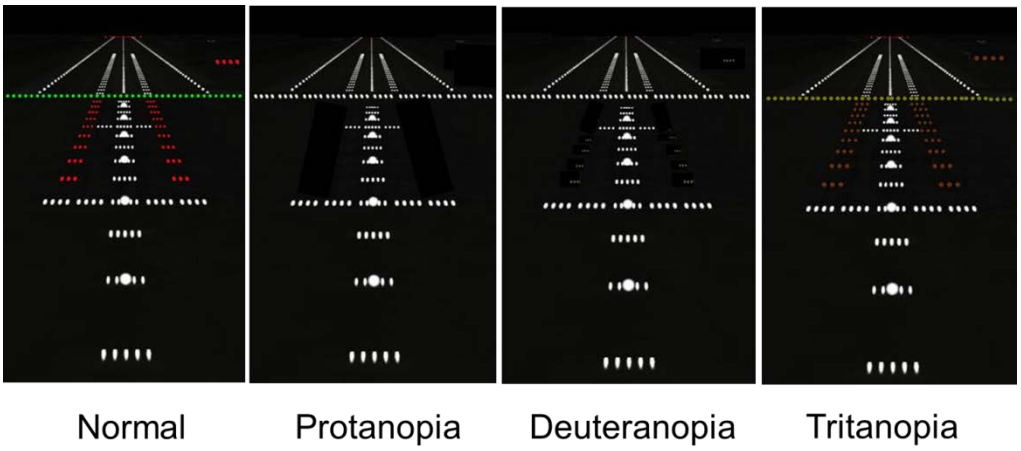


Figure 1. Approach lights as notionally perceived by those with normal and deficient color vision.

Traditionally, integrating controllers and writing software code for use in experiments required a significant amount of training. Now, software solutions, including open source, make programming the device relatively simple. Open source hardware and software are increasing the flexibility and capabilities of devices for controlled experiments. Open source solutions are often more accessible to those with limited research budgets. For instance, processing physiological data, auditory experiments, and vision researchers have utilized open source hardware and software (Christie & Gianaros, 2013; Hillenbrand & Gayvert, 2005; Teikari et al., 2012).

Lessons learned from the design and construction of a manually controlled precision approach path indicator (PAPI) experimental device (Milburn & Gildea, 2012) aided the design process. Early discussions led us to explore microcontrollers that could control LED luminance using pulse width modulation (PWM). This and other studies utilized an open source microcontroller called an Arduino (D'Ausilio, 2012, Teikari et al., 2012). In addition to the Arduino, there are a number of other microcontroller solutions available. Some of these devices are compatible with Arduino hardware and/or software, while others do not assure compatibility with the Arduino, but are similar in concept. There are also single-board computers that operate in conjunction with Arduino or similar devices providing increased capabilities. Some boards use various versions of Linux and provide additional options for programming languages.

The use of rapid development tools, open source options, and freeware solutions extends beyond the capabilities used for construction of the device. Design and prototype development are also areas that are amenable to hardware and software advances. Computer-Aided Design (CAD) programs, in association with Computer-Aided Manufacturing (CAM) tools, can reduce the likelihood of late modifications and reworks. Some CAD programs are capable of generating an associated parts list that can make it easier to estimate costs.

For designing and simulating integrated circuit functionality, there are several versions of SPICE (Simulation Program with Integrated Circuit Emphasis) are available, including the current version available from the University of California-Berkeley. LTspice™, Multisim™,

Ngspice, TINA-TI™, and XSPICE are also available at no cost and are relatively easy to use. These software tools are useful in evaluating circuit designs prior to moving to a physical breadboard to test circuits.

For determining the physical layout on a printed circuit board (PCB), DesignSpark PCB, ExpressPCB™, and similar software packages provide a manufacturing file (e.g., RS-274 Gerber file format, Excellon format) that can be sent to certain custom PCB manufacturers for drilling and etching at a nominal fee (often <$100) with no limit on the number of boards manufactured. These manufacturers can deliver the board to the designer in a bare state with just the circuit conduction paths, or traces, and no components soldered to the PCB. Some software packages, including DesignSpark PCB and ExpressPCB™ prepare a list of components for direct order of the discrete components (e.g., resistors, capacitors, ICs). For an additional fee, some PCB manufacturers will deliver the board fully assembled (e.g., Pad2Pad®, Sunstone Circuits®).

Even with a "clean sheet" design, many components are often available commercially. There are generally a number of viable options for any given component or subsystem, as well as macro level design and implementation. In many instances, it is necessary to modify off-the-shelf components for uses other than for those originally intended. Such off-the-shelf devices are generally much cheaper than custom manufacturing or machining of components.

We needed a device that could present multi-condition stimuli representing: 1) a signal light gun, 2) PAPI lights as their current incandescent sources and as planned LED sources, 3) a tricolor PAPI LED condition, and 4) with the luminances controlled to match in-service incandescent (white 2 times brighter than red) or of equal luminance. Obviously, there was no off-the-shelf apparatus available to present any or all of those stimuli.

PAPI Experimental Device

There were two chromaticity conditions with the LED light source. Half of the LED groupings consisted of unicolor emissions with each of the three LEDs emitting the same wavelength. The other half of the groupings consisted of LEDs, each emitting a separate wavelength for a tricolor condition. The unicolor white LEDs were 5500 Kelvin (5500K), and the tricolor white LEDs were 3000K, 5500K, and 8000K. Unicolor red LEDs were 642nm with the tricolor red LEDs being 628nm, 642nm, and 660nm.

LED selection from commercially available sources is problematic in terms of chromaticity, luminance, and distribution pattern. There is a limited selection of chromaticities available from LEDs. For the purposes of this study, it was necessary to have an LED that had a dominant wavelength of 660nm, and the available LEDs with sufficient luminance for our purposes could provide a peak wavelength of 660nm — but the dominant wavelength from these LEDs was generally 652nm. This was too close to the 642nm emissions of some of the red LEDs for our experimental purposes. To address this issue, a bandpass filter was selected that would only pass 660nm +/- 2nm, and this was placed in front of a 642nm LED to provide the required emissions.

The luminance intensities of the commercially available LEDs also presented challenges. LEDs that presented identical luminous intensities for all of the necessary wavelengths were not

available. The solution was to select LEDs that provided a higher luminous intensity than was necessary and then adjust those luminous intensities with a combination of current-limiting variable resistors/potentiometers and PWM control from the Arduino.

Code for the Arduino microprocessor. As mentioned in the *Hardware* section, The Arduino microcontroller board uses an 8-bit, 16 MHz Atmel AVR microprocessor. The microprocessor is programmed using an open-source, Wiring-based language that is similar to C/C++. The software runs on Windows®, Mac OS X®, and Linux. For a detailed explanation of programing with the Arduino, see http://arduino.cc/en/Tutorial.

Another method of controlling luminous intensity with software is using the ShiftPWM library that creates outputs using shift registers. The code and additional information for the use of ShiftPWM is readily available online. There are additional integrated circuit options for control of luminous intensity. Dedicated LED drivers are available including the LTC3220, TLC5940 16-Channel LED Driver, LM3409HV, and the MY9221 with the addition of external current-limiting resistors. The ARD127D2P Rainbowduino ATmega328 board is an Arduino-compatible board constructed specifically for controlling LED matrices with the use of MY9221 integrated circuits as the modulation controllers.

A SainSmart relay board controlled the incandescent bulbs, although we considered constructing a relay board from discrete components. The cost of a custom relay board, even without considering the labor involved, was significantly more than that of the SainSmart solution. Designers and developers will likely find that to be the case with many devices when weighing options between COTS devices (e.g., microcontroller boards, integrated circuits) and constructing custom devices. However, for those who are unable to find workable commercial solutions, there are often examples of workable circuits online.

Solutions for LED luminous intensity control included constant current devices and the current limiting variable resistors (potentiometers wired using only two of the leads). Several commercial devices are available for constant current control of LEDs; these include the LuxDrive™ 3023 Wired BuckPuck Modules, TLC5940 16-Channel LED Driver, and LM27964 White LED Driver System with I2C Compatible Brightness Control.

The main concern with control of the luminous intensity was stability and repeatability of the stimuli. We were also operating on a very constrained budget, so we sought a solution that was both cost-effective and would provide a stable output. We took repeated measures of the output of the LEDs under the control of several mechanisms with an integrating sphere. The luminous intensity was repeatable across multiple days when using the potentiometers. If we construct another device in the future, we would use constant current devices for LED control.

Ultimately, the authors tested the LED control circuit design using three sample LEDs, each combined with a potentiometer, 200-ohm resistor, and a digital output from the Arduino Mega. The breadboard enabled us to validate the ability to control luminous intensity, timing, and sequence of presentation. Following this test, we constructed the final circuits using perforated circuit boards. Perforated circuit boards used at this stage provided the flexibility to modify the circuit, if necessary, without the challenges of a custom etched circuit board.

Although there are digital devices for controlling of the incandescent bulb luminous intensities, in the final device, bidirectional triode thyrister (TRIAC) switches controlled the incandescent bulbs. The intensities of the bulbs remained constant with repeated measurements using the integrating sphere. The main drawback of using the potentiometers and dimmer switches for luminous intensity control is the labor-intensive task of manually setting each control.

Ultimately, the apparatus we designed and used facilitated the experiment by precisely controlling the stimulus duration, the inter-trial time, luminance, and automatically announcing the trial number. These enhancements were a definite improvement over our previous, manually operated device. Further enhancements could easily incorporate a response device that would automate data collection.

For researchers seeking to design custom devices, there are a number of open source solutions. As with the design of most stimulus-presentation devices for research use, or probably most other unique devices, developers should anticipate modifications of plans, hardware, and software. The challenges are those experienced with any prototype in terms of designing, implementing, testing, and modifying. If a researcher can purchase the necessary device off-the-shelf—buy it; if not, expect to test and modify after the initial design.

References

Christie, I. C. & Gianaros, P. J. (2013). PhysioScripts: An extensible, open source platform for the processing of physiological data. *Behavior Research Methods, 45*, 125-131. doi:10.3758/s13428-012-0233-x

D'Ausilio, A. (2012). Arduino: A low-cost multipurpose lab equipment. Behavior Research Methods, 44, 305-313. doi:10.3758/s13428-011-0163-z

Federal Aviation Administration (2011). Advisory Circular 150/5345-28G, Precision Approach Path Indicator (PAPI) systems. Retrieved from http://www.faa.gov/documentLibrary/media/Advisory_Circular/150_5345_28g.pdf

Gildea, K. M. & Milburn, N. (2013). Open source products for a lighting experiment device. Behavioral Research Methods, doi:10.3758/s13428-013-0423- 1.

Hillenbrand, J. M. & Gayvert, R. T. (2005). Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research, 48,* 45–60. doi:10.1044/1092-4388(2005/005)

Milburn, N., Chidester, T., Peterson, S., Roberts, C., Perry, D., & Gildea, K. M. (2013). Pilot color vision research and recommendations. Presented at the 84th Annual Meeting of the Aerospace Medical Association, Chicago, IL.

Milburn, N., & Gildea, K. M. (2012). Usability of light-emitting diode (LED) Precision Approach Path Indicator (PAPI) simulator by color-deficient and color normal observers.

Presented at the 83rd Annual Meeting of the Aerospace Medical Association, Atlanta, GA.

Milburn, N., Gildea, K., Bullough, J.D., & Yakopcic, C. (2013). Aviation-Related Light-Emitting Diode (LED) Perception Research. *Aviation, Space, and Environmental Medicine, 84,* 876-878.

Milburn, N. J., Gildea, K. M., Perry, D., Roberts, C., & Peterson, L. S. (2014). Usability of Light-Emitting Diodes (LEDs) in Precision Approach Path Indicator (PAPI) systems by individuals with marginal color vision.

Teikari, P., Najjar, R. P., Malkki, H., Knoblauch, K., Dumortier, D., Gronfier, C., & Cooper, H. M. (2012). An inexpensive Arduino-based LED stimulator system for vision research. *Journal of Neuroscience Methods, 211,* 227-236. doi: 10.1016/j.jneumeth.2012.09.012

Acknowledgment

Research reported in this presentation was conducted under the Flight Deck Program Directive / Level of Effort Agreement between the FAA NextGen Human Factors Research and Engineering Division (ANG-C1) and the Aerospace Human Factors Division (AAM-500) of the Civil Aerospace Medical Institute sponsored by the Office of Aerospace Medicine and supported through the FAA NextGen Human Factors Division.

PHYSIOLOGICAL INDICATORS OF WORKLOAD IN A REMOTELY PILOTED AIRCRAFT SIMULATION

Michael Hoepf
Oak Ridge Institute for Science and Education
Matt Middendorf
Middendorf Scientific Services
Samantha Epling
Ball Aerospace & Technologies Corp.
Scott Galster
Air Force Research Laboratory
Dayton, Ohio

Toward preventing performance decrements associated with mental overload in remotely piloted aircraft (RPA) operations, the current research investigated the feasibility of using physiological measures to assess cognitive workload. Two RPA operators were interviewed to identify factors that impact workload in target tracking missions. Performance, subjective workload, cortical, cardiac and eye data were collected. One cardiac and several eye measures were sensitive to changes in workload as evidenced by performance and subjective workload data. Potential future applications of this research include closed loop systems that employ advanced augmentation strategies, such as adaptive automation. Thus, by identifying physiological measures well suited for monitoring workload a realistic simulation, this research advances the literature toward real-time workload mitigation in RPA field operations.

U.S. armed forces are increasingly using remotely piloted aircraft (RPA) to accomplish missions in hostile environments because of their standoff capability in areas that are difficult to access or otherwise considered too hazardous for manned aircraft or personnel on the ground (U.S. Department of Defense, 2011). It has been documented that the military intends to increase the number of RPA in service while simultaneously reducing the number of operators (Dixon, Wickens, & Chang, 2004). One proposal to accomplish this is to allow operators to control multiple aircraft simultaneously (Rose, Arnold, & Howse, 2013). However, piloting one aircraft remotely is a complex task, and operating additional aircraft could increase task demands sharply. This is potentially problematic because cognitive overload can negatively impact performance (Young & Stanton, 2002). One solution to offset this risk is to monitor operator workload in real-time and provide augmentation before performance decrements occur. Physiological measures, which have been shown to reflect changes in cognitive workload in various environments (e.g., Wilson & Russell, 2007), are well suited for this goal.

Before physiological measures can be used to monitor workload in RPA field operations, additional research is needed using realistic task environments. This is because each category of physiological measures (e.g., cortical, cardiac, and eye-based) is sensitive to workload in different situations. Hankins and Wilson (1998), for instance, found that cortical measures were sensitive to workload during mental calculation, cardiac measures were related to workload during flight segments heavily dependent on instrument use, and eye activity was associated with workload during visually demanding flight segments.

To address the research needs outlined above, an experiment was designed using a high-fidelity RPA simulation. In this study workload was experimentally manipulated and several physiological measures were collected while participants performed a target tracking task. Two RPA subject matter experts (SMEs) were interviewed to identify factors that can affect workload in this task. Two factors that were identified were weather (clear vs. hazy) and the route the target would follow (city vs. country). One of the SMEs engaged in a test run of the experiment to test and adjust the implementation of the factors. In addition to these two factors, a third factor was implemented which manipulated the number of targets being tracked (1 vs. 2).

In this study the physiological measures included the electroencephalogram (EEG), the electrocardiogram (ECG) and several eye measures. Researchers have demonstrated that EEG can be used in real-time to assess mental workload in aviation environments (e.g., Wilson & Russell, 2007). ECG was used to obtain heart rate (HR) and heart rate variability (HRV). In both laboratory and field settings, researchers typically observe HR increases and HRV decreases in high workload situations (Wilson, 1992). Eye measures included blink rate, blink duration, and

pupil diameter. Generally, during increased cognitive workload, blink rate and duration decrease (Fogarty & Stern, 1989), and pupil diameter increases (Beatty, 1982).

Methods

Participants

Six people who were either students at a Midwestern university or recent graduates participated in the experiment. They were paid $15 per hour for their participation. Three participants were female and three were male. Age ranged from 19-28 years, with a mean of 22.3. They were screened for motor, perceptual, cognitive, heart, and neurological conditions, as well as hearing impairments. They were fluent in English, right-handed, had normal or corrected-to-normal eyesight with no color blindness, and provided written informed consent in accordance with human research ethics guidelines prior to the start of the experiment. All study procedures were reviewed and approved by the Air Force Research Laboratory Institutional Review Board.

The Task

Primary task. The primary task was developed using the RPA software platform "Vigilant Spirit." This software was produced by the Air Force Research Laboratory System Control Interfaces Branch (RHCI). The goal of the primary task was to track either one or two high value targets (HVTs) depending on the condition. Participants were instructed to keep the RPA sensor positioned over the HVTs, which they accomplished by clicking in the video feed with the mouse, causing the video feed to center on where they had clicked. The sensor slaved tracking feature would then automatically update the aircraft position to fly a loiter circle around this center point, thereby eliminating the need for manual aircraft navigation. Participants actively tracked the HVT(s) for the duration (4.5 minutes) of each trial.

Secondary task. A secondary task was presented concurrently with the primary task. The task consisted of answering cognitively challenging questions. There were three math questions and one mental rotation question per trial. Questions were presented verbally over a headset and transcriptions were displayed. Participants were instructed to press and hold the spacebar while they responded verbally.

Apparatus and Measures

Performance assessment. Performance was assessed using a composite scoring algorithm, which was based on components from both the primary and secondary task. For each trial, the maximum possible score was 1,000 points (800 primary and 200 secondary). To obtain points from the primary task, participants were required to keep the HVT(s) in their video feed(s). Maximum points were accumulated when using the highest two levels of zoom and half as many were accumulated at lower levels of zoom. For the secondary task, there were four questions per trial, each worth a maximum of 50 points. In order to obtain all points, participants had to respond correctly within 10 seconds. After 10 seconds, the participants would lose 1 point per second for the next 10 seconds, and then 2 points per second for the following 10 seconds. After 30 seconds, no points were given. Answering incorrectly resulted in a 5 point penalty.

Subjective workload. Subjective workload was assessed using the National Aeronautics and Space Administration-Task Load Index (NASA-TLX), a multidimensional measure that assesses perceived workload (Hart & Staveland, 1988). Workload was determined by averaging across the six sub-scales (mental demand, physical demand, temporal demand, performance, effort, and frustration). This average has been found to be psychometrically equivalent to the weighted sub-scale averaging suggested by the NASA-TLX authors (Nygren, 1991). Empirically, the weighted averages have not been found to be superior to the simple average of the sub-scales (Christ et al., 1993; Hendy, Hamilton, & Landry, 1993).

Physiological data acquisition and processing. The physiological data collected in this study included the EEG, ECG, vertical EOG (VEOG), and pupil diameter. The EEG, ECG, and VEOG signals were sampled using a Cleveland Medical Devices BioRadio 150. The EEG and VEOG were sampled at 480 Hz, and the ECG signal was sampled at 960 Hz. All signals connected to the BioRadio 150 were subjected to a hardware high pass filter with a break frequency of 0.5 Hz. The sampled data were transmitted wirelessly to a computer for processing and recording.

The EEG data were acquired using electrodes placed directly on the scalp and secured in place with an Electro-Cap manufactured by Electro-Cap International, Inc. EEG was measured at seven sites on the scalp in accordance with the international 10/20 system (Jasper, 1958). The seven sites were the F7, F8, T3, T4, Fz, Pz, and O2. The right and left mastoids were used as the reference and ground for the EEG signals. All initial electrode impedances were measured to be at or below 5 kΩ. The frequency bands (i.e., pass bands) used in the EEG signal processing were delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz), gamma 1 (31-40 Hz), gamma 2 (41-57 Hz) and gamma 3 (63-100 Hz). A two second time domain window was used to process the raw EEG data. The raw data in the two second window was filtered using a 4th order Butterworth band pass filter. A Hanning window was applied to the filtered data and power spectral analysis was performed. The resulting power in the pass band was then averaged. These steps were repeated for each frequency band and electrode site. The two second time domain windows had a 50% overlap, thus yielding one measure of average power every second. This signal processing approach yielded 49 EEG measures per second (7 sites with 7 bands per site).

The ECG data were acquired using two electrodes placed on the sternum and xiphoid process, and the VEOG data were acquired using two electrodes placed above and below the left eye. The initial electrode impedances for the VEOG and ECG were measured to be at or below 20 kΩ. The left mastoid was used as the ground for the ECG and VEOG signal. Interbeat intervals (IBIs) were extracted from the ECG data. The IBIs were used to calculate heart rate and heart rate variability. Blink rate and duration were extracted from the VEOG data using a blink detection algorithm (see Epling, Middendorf, Hoepf, & Galster, this volume). Pupil diameter data were acquired using the Smart Eye Pro 5.9 system with four cameras sampling data at 60Hz.

Competition. A social (no monetary compensation) competition was implemented to maintain motivation and prevent task disengagement (Sherif, Harvey, White, Wood, & Sherif, 1961). Top session average scores were posted on a whiteboard.

Procedure

Participants were brought into the laboratory for one day of training and four days of data collection. For training, participants first viewed a PowerPoint presentation containing a description of the task and measures, and then completed part-task training for the primary and secondary tasks. Training concluded with the completion of eight practice trials. On data collections days, participants were equipped with the physiological measurement devices and then completed eight experimental trials per day, for a total of 32 trials.

Experimental Design

The current investigation utilized a 2 x 2 x 2 full factorial design. There were three manipulations intended to impact workload, each containing two levels (easy and hard). The first manipulation was weather, which included clear (easy) and hazy (hard) conditions. The clear condition was free of clouds and visibility was unobstructed. A layer of fog was present in the hazy condition, which reduced visibility. The second manipulation was route difficulty, which referred to the roads the HVTs traveled. For the country (easy) routes, HVTs simply traveled back and forth along a long straight road, the view of the HVT was generally unobstructed. Conversely, for the city (hard) routes, the HVTs took many turns and sometimes became occluded by buildings. The third manipulation was the number of HVTs, which was either one (easy) or two (hard). In single HVT conditions, participants needed to track only one HVT, requiring only one RPA. Conversely, in the two HVT conditions, participants were required to utilize two RPA to track two HVTs simultaneously. Latin squares were used for counterbalancing such that each condition preceded every other condition an equal number of times, thereby combatting order effects.

Results

Performance. The performance, subjective workload, and physiological data were statistically evaluated using a three-way (weather, HVT, route) repeated-measures ANOVA. Performance in hazy conditions (M = 785.0, SE = 25.6) was not significantly different than the performance in clear conditions (M = 776.2, SE = 23.4). Performance was higher in conditions with country routes (M = 814.5, SE = 19.2) than in conditions with city routes (M = 746.7, SE = 31.6), $F(1, 5) = 10.18$, $p < .05$, and higher in one HVT conditions (M = 873.6, SE = 24.1) than two HVT conditions (M = 687.6, SE = 25.4), $F(1,5) = 220.30$, $p < .001$.

Subjective workload. Subjective workload in hazy conditions (M = 43.5, SE = 4.3) was not significantly different than clear conditions (M = 43.3, SE = 5.0). Subjective workload was higher in city conditions (M = 47.6, SE = 5.3) than country conditions (M = 39.1, SE = 4.1), $F(1, 5) = 18.52$, $p < .01$, and higher in two HVTs conditions (M = 54.6, SE = 6.1) than one HVT conditions (M = 32.1, SE = 4.2), $F(1, 5) = 18.97$, $p < .01$.

Cortical measures. The EEG measures (power at each site and frequency band) were analyzed for each manipulation, but for conciseness only the significant ($p < .05$) results are reported and the means, standard errors, and F values are not included. In regards to the weather manipulation, there was less power in hazy conditions than clear conditions at the O2 site in the alpha band. For the route manipulation, there was less power in city conditions than in country conditions at 7 sites, including F7, Fz, F8, T3, T4, and Pz in the delta band, and F7 in the theta band. For the HVT manipulation, there was more power for two HVT conditions than one HVT conditions at 19 sites, including F7, F8, T3, and O2 in the delta band, F7, F8, T3, T4, Pz, and O2 in the theta band, F7, F8, T3, Pz, and O2 in the alpha band, and T4 in the beta, and all three gamma bands. These effects may not be due to neural activity in the brain, but rather artifacts from eye activity (see discussion section).

Cardiac measures. HR was not significantly impacted by any of the experimental manipulations. HRV in hazy conditions (M = 0.0539, SE = 0.0050) was not significantly different than HRV in clear conditions (M = 0.0533, SE = 0.0046). HRV was significantly lower in city conditions (M = 0.0530, SE = 0.0049) than in country conditions (M = 0.0542, SE = 0.0047), $F(1,5) = 7.44$, $p < .05$, and significantly lower in two HVT conditions (M = 0.0517, SE = 0.0046) than one HVT conditions (M = 0.0555, SE = 0.0050), $F(1,5) = 19.46$, $p < .01$.

Eye measures. The weather manipulation did not significantly impact blink rate or duration. Blink rate was lower in city conditions (M = 18.34 bpm, SE = 4.88) than in country conditions (M = 19.59 bpm, SE = 5.23), $F(1,5) = 8.23$, $p < .05$. Blink rate was also lower in two the HVT conditions (M = 16.28 bpm, SE = 4.50) than in the one HVT conditions (M = 21.65 bpm, SE = 5.87), but this difference was not statistically significant $F(1,5) = 3.98$, $p = .10$. Blink duration was significantly shorter in city conditions (M = 0.1041s, SE = 0.0042) than in country conditions (M = 0.1064s, SE = 0.0043), $F(1,5) = 16.77$, $p < .01$, and shorter in two HVT conditions (M = 0.1005s, SE = 0.0047) than one HVT conditions (M = 0.1099s, SE = 0.0041), $F(1,5) = 13.81$, $p < .05$.

Interestingly, pupil diameter was significantly larger during clear conditions (M = 4.06mm, SE = 0.68) than hazy conditions (M = 3.87mm, SE = 0.59), $F(1,5) = 14.85$, $p < .05$. To investigate this finding, a Minolta Chroma-Meter CS-100 was used to assess the luminance of the two conditions. Results suggested that the pupil light reflex was most likely responsible for this difference, as hazy conditions were brighter than clear conditions. Pupil diameter was larger during city routes (M = 4.01mm, SE = 0.63) than during country routes (M = 3.92mm, SE = 0.66), although this difference was not significant, $F(1,5) = 3.38$, $p = .125$, and larger during two HVT conditions (M = 4.09mm, SE = 0.62) than during one HVT conditions (M = 3.84mm, SE = 0.65), $F(1,5) = 44.33$, $p < .01$.

Discussion

To meet increasing demand for RPA operations, future systems are envisioned in which single operators control multiple aircraft (Dixon et al., 2004). Such systems would allow an efficient use of resources during low workload operations. However, a concern is that workload could become excessive due to increased mental demand from managing multiple aircraft, possibly leading to performance decrements and mission failure. One solution to address excessive workload from controlling multiple vehicles, as well as existing challenges RPA operators experience, is to monitor operator state in real-time so that mental overload can be identified and handled appropriately. That is, accurate workload assessment would allow the implementation of augmentation strategies *before* performance decrements occur. By examining the feasibly of using physiological measures to monitor workload, this project advances the literature toward real-time workload mitigation in field operations.

In regard to the experimental manipulations, the results indicated that the weather manipulation did not impact workload. The HVTs were the only motorcycles in the simulation (other traffic consisted of cars, trucks, vans etc.), so it could be that the haze did not sufficiently obscure the visual cues necessary to track them. The route manipulation did effectively impact workload, so a strong point of this research is that this factor (identified in SME interviews) has physiological correlates that can be used for workload assessment. Similarly, results showed that tracking two targets had a strong impact on performance, workload, and physiological measures. The control of

multiple semi-autonomous air vehicles is not a current capability, and so the present research is valuable in that it provides a preview of what may be expected if such a capability is implemented in future RPA workstations.

Physiological results revealed that one cardiac and several eye measures were sensitive to the same workload manipulations as the performance and subjective workload measures. Specifically, HRV, blink duration, blink rate, and pupil diameter were generally sensitive to the route and HVT manipulations. The haze manipulation did not impact performance or subjective workload, and the physiological measures generally did not reflect a change in workload either. These physiological findings are consistent with previous research (e.g., Beatty, 1982; Fogarty & Stern, 1989; Wilson, 1992), suggesting that the results did not occur by chance, and that these physiological measures may be well suited for real-time workload monitoring in RPA field operations.

Despite these promising results, other physiological measures did not function as expected. HR, for instance, did not demonstrate sensitivity to any workload manipulation. Additionally, the EEG data was difficult to interpret, as results were not always in the expected direction. Alpha power, for instance, increased at several sites in two HVT conditions, which were shown to be the more difficult conditions as evidenced by the performance and subjective workload data. This is in contrast to the classic concept that alpha activity is an idling rhythm of humans at rest, which becomes desynchronized during cognitive processes (Pfurtscheller & Lopes da Silva, 1999). One possible explanation is that EOG artifacts (see Fatourechi, Bashashati, Ward, & Birch, 2007) were present in the EEG data due to the additional saccades associated with two target tracking conditions. Although an effort was made to remove these artifacts, some residual effects remained in the data. Specifically, the band pass filter used to process the EEG data continues to ring for up to one second after the artifact occurs. This indicates a need for an improvement to the EEG signal processing software for future research.

Admittedly, a limitation of the current study is the small sample size, such that these findings could be specific to the sample. In addition, several measures (i.e., blink rate, pupil diameter) sometimes failed to yield significant differences, despite trending in the expected direction. A larger sample size would help to either confirm or deny the utility of these physiological measures.

Conclusion

The current research investigated the feasibility of using physiological measures to monitor workload in a high fidelity RPA tracking task. One cardiac (HRV) and several eye measures (blink rate, blink duration, and pupil diameter) demonstrated sensitivity to changes in workload, and thereby appear well suited to real-time workload monitoring. In the future, these measures could be used in physiologically driven adaptive automation (see Scerbo, 1996) systems to prevent performance decrements. Given the promising nature of these results, future research is encouraged on the topic.

Acknowledgements

The authors would like to thank Chelsey Credlebaugh and Jonathan Mead for their assistance in data reduction and preparation of this manuscript, as well as Chuck Goodyear for his help with statistical analysis. We would also like to thank Kevin Durkee, Noah DePriest, and Mark Squire for their technical support. This research was supported in part by an appointment to the Student Research Participation Program at the U.S. Air Force Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and USAFRL. The views expressed in this report are solely those of the authors and do not necessarily reflect the views of the employers or granting organizations.

References

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276-292.

Christ, R. E., Hill., S. G., Ayers, J. C., Iavecchia, H. M., Zaklad, A. L., & Bittner, A. (1993). *Application and validation of workload assessment techniques* (Technical Report 974). Alexandria, VA: U.S. Army Research Institute for the Behavioral Sciences.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 152-156). New Orleans, LA: Human Factors and Ergonomics Society.

Epling, S., Middendorf, M., Hoepf, M., & Galster, S. (this volume). The electrooculogram and a new blink detection algorithm. In *Proceedings of the 18th International Symposium on Aviation Psychology.* Dayton, OH: Wright State University.

Fatourechi, M., Bashashati, A., Ward, R. K., & Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology, 118*(3), 480-494.

Fogarty, C., & Stern, J. (1989). Eye movements and blinks: Their relationship to higher cognitive processes. *International Journal of Psychophysiology, 8*, 35-42.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69,* 360-367.

Hart, S. G., & Staveland, L. E. (1988). Development of the NASA- TLX (Task Load Index): Results of experimental and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp.138-183). Amsterdam: North-Holland Press.

Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 35*, 579-601.

Jasper, H. (1958). Report of the committee on methods of clinical examination. *Electroencephalography and Clinical Neurophysiology, 10*, 370-375.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33*, 17-31.

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology, 110*, 1842-1857.

Rose, M. R., Arnold, R. D., & Howse, W. R. (2013). Unmanned aircraft systems selection practices: Current research and future directions. *Military Psychology*, *25*(5), 413-427.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*. Mahwah, NJ: Erlbaum.

Sherif, M., Harvey, O. J., White, B. J., Wood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The robber's cave experiment.* Normal, OK: Institute of Group Studies.

U.S. Department of Defense. (2011). *Unmanned systems integrated roadmap: FY2011-2036* (Reference No. 11-S-3613). Washington, DC: Department of Defense.

Wilson, G. F., (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology, 34*(2-3), 163-178.

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors, 49*(6), 1005-1018.

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science, 3*(2), 178-194.

EEG DATA ANALYSIS USING ARTIFACT SEPARATION

Chelsey Credlebaugh
Ball Aerospace and Technologies Corp.
Matthew Middendorf
Middendorf Scientific Services
Michael Hoepf
Oak Ridge Institute for Science and Education
Scott Galster
Air Force Research Laboratory

It has been postulated that physiological measures can be a positive indicator of mental workload. One such measure is the electroencephalogram (EEG). It is well known that the EEG signal is easily affected by artifacts. One prominent source of artifacts is eye activity, including blinks and saccades. These contaminates coincide directly with EEG signals, making it difficult to obtain artifact-free data. This paper discusses a methodology that performs artifact separation at the data analysis stage. This technique was used to analyze data from a recent experiment. Workload was manipulated by varying the difficulty of the primary task while responding to mathematical communications on the secondary task. Our findings demonstrate the importance of distinguishing between statistical significances found in the EEG signal as caused by neuronal activity versus those caused by artifacts. The artifact separation approach facilitates this investigation.

Mental workload has been described as an intervening variable that reflects the extent to which the information processing abilities of a participant are engaged during task performance (Gopher & Donchin, 1986). The ability to reliably assess mental workload is important due to the effect increased workload can have on human operator performance. This is vital due to the ever increasing complexities of technology and systems, and the higher demand they place on the human operator (Hankins & Wilson, 1998). The most basic issue in the study of cognitive workload is the problem of how to effectively measure it (Gevins & Smith, 2003). Tsang & Wilson (1997) classified workload measurements into three general categories, which include: performance, subjective evaluation and physiological measures, including electroencephalography (EEG) and electrooculography (EOG).

The Electroencephalogram

EEG is a noninvasive electrical sensing technique that uses electrodes placed on the scalp to measure brain activity. Dependent upon the research, different sites may be used. The locations of these sites are based on the International 10-20 system (Jasper, 1958). Researchers have reported the sensitivity of EEG to changes in mental workload (Gevins & Smith, 2003). It has been shown that the delta band (1-3 Hz) and theta band (4-7 Hz) spectral peaks increase in power during high workload related tasks (Gevins & Smith, 2003). In contrast, multiple studies have shown that power decreases in the alpha band (8-12 Hz) during high workload (Gevins & Smith, 2003).

Although EEG has often been used as a measure of cognitive workload, it has some functional and practical limitations that must be carefully considered before being applied to operational settings. EEG signals are easily corrupted by a number of artifacts. That is, in addition to the brain's electrical activity recorded at the scalp, the EEG signal can include contaminating potentials from rapid eye movements and blinks. (Gevins & Smith, 2003).

The Electrooculogram

The electrooculogram (EOG) is a measure of electrical signals associated with eye activity, including blinks and rapid eye movement (saccades). The vertical EOG (VEOG) is a sensing technique that uses electrodes placed above and below one eye to measure vertical eye activity. The VEOG signal is processed by algorithms to detect blinks and saccades. It has been reported that these eye-based measures can be used to assess changes in cognitive workload (Fogarty & Stern, 1989).

Typical blink measures include: amplitude, duration and frequency. It has been reported that when faced with increased cognitive workload; participants will blink with reduced duration and frequency (Recarte, Perez, Conchillo & Nunes, 2008). Typical saccade measures include: amplitude, velocity, and length. Many studies have reported that the peak saccade velocity will increase as workload increases (Wang & Zhou, 2013).

Among EOG artifacts, blinks cause the largest distortions, mainly because of the movement of the eyelids

across the surface of the eyes. It is often the case in research that experimental manipulations can result in changes in eye activity. Therefore it is very important that the associated artifacts be dealt with effectively or else the EEG results could be obscured or misleading.

Artifact Mediation Approaches

Considering the effects of artifacts on the EEG signal, a great deal of research has been directed towards artifact mediation (Gevins & Smith, 2003). Common methods of dealing with artifacts in the EEG are artifact avoidance, artifact rejection, and artifact removal. The artifact avoidance method consists of avoiding their occurrence by issuing instructions to the participants to not blink. Designing tasks that do not require gaze changes, thus avoiding saccades, is another way artifact avoidance can be achieved. Artifact avoidance has the advantage of being the least computationally demanding, since it is assumed that no artifact is present in the signal (Fatourechi, Bashashati, Ward & Birch, 2006). It also has several drawbacks including, the inability to control eye and body movements and the understanding that artifacts will always be present in brain signals.

Artifact rejection refers to the process of rejecting the data affected by artifacts (Fatourechi, Bashashati, Ward & Birch, 2006). Artifact rejection can be done manually or automatically. During the manual rejection method, data is visually checked by an expert and the contaminated EEG data are removed from the analysis (Fatourechi, Bashashati, Ward & Birch, 2006). Manually rejecting data is not computationally demanding but faces many disadvantages. These disadvantages include the cost of intense labor while the process of selecting the artifact-free data may become subjective and the rejection of artifact-contaminated data may lead to a loss of data (Fatourechi, Bashashati, Ward & Birch, 2006). While manual rejection focuses on human correction, automatic rejection discards segments that are contaminated automatically using the EOG signals or by using EEG signals contaminated with artifacts (Gratton, 1998). Both approaches are less labor intensive but still suffer from sampling bias and loss of valuable data.

Artifact removal is the process of reducing the impact of the artifact on the EEG signal. This may be thought of as an attempt to 'fix' the signal in the time domain so that it remains continuous. This artifact mediation approach is relatively simple and involves using mathematical solutions to remove the artifacts. Common methods for artifact removal include: linear filtering, linear combination, regression, blind source separation (e.g., independent component analysis) and principle component analysis. These methods, however, fail when the EOG artifacts lie in the frequency bands of interest. Subtracting the EOG signal may remove part of the EEG signal.

Experimental Background

In our work we explore the use of EEG as an indicator of cognitive workload. In this study we found there was a significant effect of workload on frontal delta. However, we were concerned about this finding because the effect was in the wrong direction. Specifically, spectral power in the delta band decreased in the high workload condition. It has been reported that blink rate decreases under high workload conditions, so it was unclear if the significant frontal delta effects were due to brain activity or EOG artifacts (Wang & Zhou, 2013).

To investigate the concern above, a blink detection algorithm (Epling et al., this volume) was written to process the VEOG data. This algorithm was used to support a technique for addressing artifacts that we refer to as artifact separation. Specifically, the EEG spectral measures that are blink-free are separated from the contaminated measures at the data analysis stage. When this technique was applied to EEG measures, many of the significant effects on frontal delta disappeared. A second algorithm was written to detect saccades using EEG data (see discussion section). When the saccades were separated, the remaining significance in frontal delta disappeared. Each EEG spectral measure is accompanied by two flags to indicate the presence of artifacts (blinks and saccades). We are intending that the contribution of this paper will focus on the artifact separation technique. However, the site-specific saccade detection approach, and a robust blink detection algorithm are also noteworthy.

Methods

Participants

There were a total of 6 participants in this study, with 3 males and 3 females. The age of participants ranged from 19-28 (M=22.3). Participants were recruited from a local mid-western university. They read and signed the informed consent document before participating and were compensated for their time. All study procedures were reviewed and approved by the Air Force Research Laboratory Institutional Review Board.

Task

In this experiment, the primary task was to track one or two high value targets (HVTs). The task was implemented using Vigilant Spirit 3.14, which is a remotely piloted aircraft (RPA) simulator. This software was produced by the Air Force Research Laboratory System Control Interfaces Branch (RHCI). Participants were instructed to track the HVT(s) by continuously clicking in each video feed while the HVT traveled by motorcycle. Dependent upon the condition, the HVT on the motorcycle would either take a route through the city or country, during clear or hazy visibility. Half of the trials consisted of tracking one HVT and the other half consisted of tracking two HVTs. The secondary task consisted of answering operationally relevant questions. The composite scoring algorithm was based on components of both the primary and secondary task. For each trial, the maximum possible score was 1,000 points (with 800 primary and 200 secondary). Note: for additional information on the actual task, design and procedure of the experiment, see Hoepf, Middendorf, Epling & Galster, this volume.

Apparatus and Measures

Seven channels of EEG data were recorded during this study which included: F7, Fz, F8, T3, T4, Pz and O2. The frequency ranges of the seven bands of EEG were delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz), gamma 1 (31-40 Hz), gamma 2 (41-57 Hz) and gamma 3 (63-100 Hz). The VEOG data were acquired using two electrodes placed above and below the left eye, Mastoids were used as reference and ground points. Electrode impedances were below 5kΩ for EEG and 20kΩ for VEOG. The EEG data and the VEOG data were sampled at 480 Hz using the Cleveland Medical Devices BioRadio 150. This device has hardware high pass filters with break frequencies of 0.5 Hz.

Analysis Approach

EEG signal processing. The raw EEG data were split into two-second windows and filtered using a 4^{th} order Butterworth band pass filter with pass bands set as described earlier. A Hanning window was applied and a power spectral analysis was performed. The resulting power in each window was then averaged. The two-second time domain windows had a 50% overlap, thus yielding one average power measure every second for each frequency band and site. This produced a total of 49 measures per second (7 frequency bands at 7 sites).

Blink detection algorithm. The blink detection algorithm uses VEOG to identify blinks in real-time. The main features computed for each blink are its amplitude and duration. After two or more blinks are found, blink rate can be computed. The major components of the blink detection algorithm are threshold generation, feature extraction state machine, scoring & classification and blink save/false detection logic. The blink detection algorithm was validated using truth data (Epling et al., this volume).

Saccade detection algorithm. Due to horizontal EOG not being recorded, an EEG-based saccade detection algorithm was developed. The two-second window of data used for EEG signal processing is evaluated to find the largest saccade in the window, if one exists. A sliding linear fit is performed that is 25 milliseconds long and must have an R^2 value greater than 0.9. The linear fit must also have a high slope (greater than 550 microvolts per second). Once an initial fit is found, its length is allowed to grow until the R^2 value fails. The length, amplitude and velocity of the saccade are then computed from the final linear fit.

Procedure

Participants were brought into the laboratory for one training session and four data collection sessions. For training, participants were asked to read through a PowerPoint presentation briefing them on task instructions. The researchers then provided training on each individual task, followed by eight practice trials. Each participant received performance feedback from the composite score after each trial. At the end of each trial, self-reported workload assessments were obtained using the NASA Task Load Index (TLX) (Hart & Staveland, 1988). On data collection days, participants were equipped with physiological sensors which included EEG and VEOG. Participants then completed eight trials per day, for a total of 32 trials.

Design

There were three independent variables in this study, each containing two levels. The three variables were visibility (clear/hazy), number of high value targets (one/two) and route type (city/country). We utilized a 2 x 2 x 2 full factorial repeated measures design. The performance, workload, and physiological data were statistically evaluated using a three-way (weather, HVT, route) repeated-measures ANOVA.

Results

Performance

Performance in hazy conditions (M = 785.0, SE = 25.6) was not significantly different than the performance in clear conditions (M = 776.2, SE = 23.4). Performance score was higher in conditions with country routes (M = 814.5, SE = 19.2) than in conditions with city routes (M = 746.7, SE = 31.6), $F(1, 5) = 10.18$, $p < .05$, and higher in one HVT conditions (M = 873.6, SE = 24.1) than two HVT conditions (M = 687.6, SE = 25.4), $F(1,5) = 220.30$, $p < .001$.

Subjective Workload

Workload in hazy conditions (M = 43.5, SE = 4.3) was not significantly different than clear conditions (M = 43.3, SE = 5.0). Workload was higher in city conditions (M = 47.6, SE = 5.3) than country conditions (M = 39.1, SE = 4.1), $F(1, 5) = 18.52$, $p < .01$, and higher in two HVTs conditions (M = 54.6, SE = 6.1) than one HVT conditions (M = 32.1, SE = 4.2), $F(1, 5) = 18.97$, $p < .01$.

Cortical Measures

The EEG measures (power at each site and frequency band) were analyzed for each manipulation, but for conciseness only the significant ($p < .05$) results are reported and the means, standard errors, and F values are not included. There was less power in hazy conditions than clear conditions at the O2 site in the alpha band. For the route manipulation, there was less power in city conditions than in country conditions at 7 sites, including F7, Fz, F8, T3, T4, and Pz in the delta band, and F7 in the theta band. For the HVT manipulation, there was more power for two HVT conditions than one HVT conditions at 15 sites, see Figure 3 (top row).These effects may not be due to neural activity in the brain, but rather artifacts from eye activity (see discussion section).

Eye-Measures

The weather manipulation did not significantly impact blink rate or duration. However, blink rate was lower in city conditions (M = 18.34 bpm, SE = 4.88) than in country conditions (M = 19.59 bpm, SE = 5.23), $F(1,5) = 8.23$, $p < .05$. Blink rate was also lower in the two HVT conditions (M = 16.28 bpm, SE = 4.50) than in the one HVT conditions (M = 21.65 bpm, SE = 5.87), but this difference was not statistically significant $F(1,5) = 3.98$, $p = .10$. Blink duration was significantly shorter in city conditions (M = 0.1041s, SE = 0.0042) than in country conditions (M = 0.1064s, SE = 0.0043), $F(1,5) = 16.77$, $p < .01$, and shorter in two HVT conditions (M = 0.1005s, SE = 0.0047) than in the one HVT conditions (M = 0.1099s, SE = 0.0041), $F(1,5) = 13.81$, $p < .05$.

Discussion

The focus of this paper is on an analysis methodology based on artifact separation. One could reasonably argue that artifact separation is the same thing as automatic artifact rejection. One big difference is artifact rejection is typically done in the time domain, and our artifact separation approach is done on the spectral results. Another nuance is that it's up to the consumer of the data to decide what to do with the artifact flags. The EEG spectral results are available in real time for such applications as machine learning models. In this case a model could decide if it wants artifact free data using the flags.

The number of HVTs manipulation introduced a task-related effect in the EOG data. When one target was being tracked, the participants would focus on one video feed. When two targets were being tracked, the participants had to regularly shift their gaze between the two video feeds, thus introducing substantially more saccades. The original intent in this study was to use EOG to detect blinks, so only the vertical EOG was collected. Due to the task-related effect, most of the saccades were in the horizontal axis. Therefore the VEOG data was insufficient for saccade detection. A new approach was implemented to detect saccades directly in the EEG data.

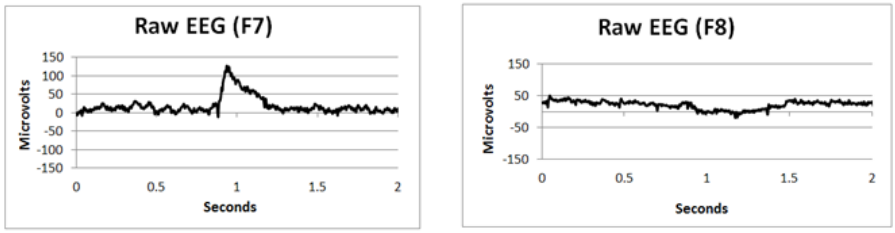


Figure 1. This data illustrates that saccade artifacts can be site-specific

One benefit of EEG-based saccade detection is that it is site-specific. An EOG-based approach would suggest that all sites contain the artifact. Based on examination of raw EEG data it is concluded that a saccade can contaminate one site but not another. Figure 1 shows that F7 is contaminated by a saccade while F8 is not. This is likely due to the angle of the saccade. EEG-based saccade detection results in less data being flagged as contaminated.

The artifact separation technique was applied to see if significant frontal delta effects for the route (country vs. city) manipulation were due to eye activity, or if they were due to an actual neurological phenomenon. When all of the data were used (no artifact separation), there was a significant effect of increased workload at six EEG sites. When blinks were separated, only two sites remained significant. These two sites lost significance with both blinks and saccades were separated (Figure 2).

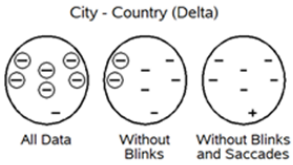


Figure 2. The sign shows the direction of the difference in log power. The size of the sign is the relative absolute value of

The task-related effect led to widespread significant effects in the EEG data due to the number of targets manipulation (Figure 3). Applying the artifact separation technique had little impact on the widespread effects. We believe this is a side effect of the band pass filter that is used in the EEG signal processing. When a saccade passes through the filter it will ring at, or near, the center frequency of the pass band. The power due to the ringing of the filter overwhelms the power found in EEG signal alone (Figure 4). The so called "artifact free" data is not truly artifact free because only the big saccades are detected and flagged. Therefore, many smaller saccades go undetected and the task-related effect is still prominent.

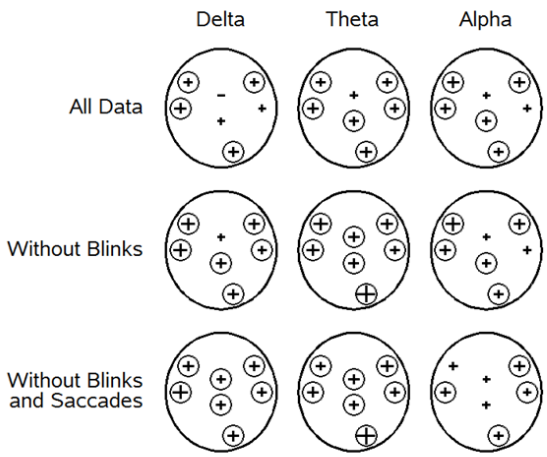


Figure 3: The sign shows the direction of the difference in log power. The size of the sign is the relative absolute value of the t statistic. If the sign is circled then $p \leq 0.05$.

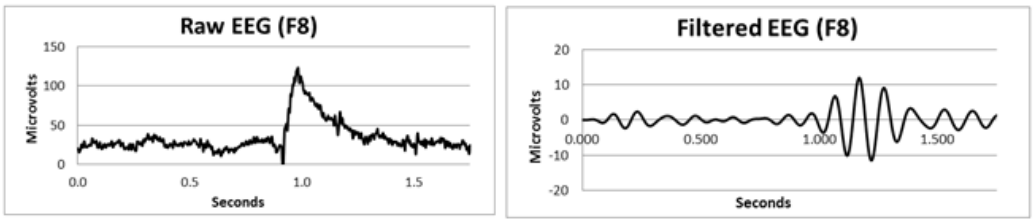


Figure 4. Increase in alpha power due to the ringing of the band pass filter.

Conclusions

The artifact separation technique seems to have real promise. In particular, it does not attempt to 'fix' the signal in the time domain. One drawback is that it can result in less data being used in the analysis stage. Secondly, the EEG signal processing algorithm needs to be enhanced to more effectively account for power from artifacts crossing window boundaries. The choice of filter types and order should be systematically evaluated.

The EEG-based saccade detection algorithm has the benefit of retaining more artifact free data because it is site-specific. One downside to this approach is it is good at finding big saccades and not as reliable for the smaller ones. The performance of this algorithm could be substantially improved if it is coupled with a polar-based saccade detection algorithm using EOG data (Middendorf, Epling, Hoepf & Galster, this volume). This enhancement and others are planned for future research. Lastly, when attempting to use EEG to assess cognitive workload, carefully evaluate the experimental manipulations to see if they introduce a task-related effect that systematically changes eye activity.

Acknowledgements

The authors wish to express gratitude to Kevin Durkee, Samantha Epling, Chuck Goodyear, Christina Gruenwald and Lucas Stork for assistance with data analysis and preparation of this manuscript. The views expressed in this report are solely those of the authors and do not necessarily reflect the views of their organizations.

References

Epling, S., Middendorf, M., Hoepf, M., Galster, S., Gruenwald, C., & Stork, L. (this volume). The Electrooculogram and a new blink detection algorithm. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Fatourechi, M., Bashashati, A., Ward, R.K., & Birch, G.E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology, 118*(3), 480-494.

Fogarty, C., & Stern, J. (1989). Eye movements and blinks: Their relationship to higher cognitive processes. *International Journal of Psychophysiology, 8,* 35-42.

Gevins, A.S., & Smith, M.E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4,* 113-131.

Gratton, G. (1998). Dealing with artifacts: The EOG contamination of event-related brain potential. *Behavior Research Methods, Instruments & Computers, 30,* 44-53.

Gopher, G., & Donchin, E. (1986). Workload – An examination of the concept. In K.R. Boff L. Kaufman & J.P. Thomas (Ed.), *Handbook of Perception and Human Performance, Vol. 2,* New York: John Wiley.

Hankins, T.C., & Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG and subjective measure of pilot mental workload during flight. *Aviation, Space and Environment Medicine, 69,* 360-367.

Hart, S.G. & Staveland, L.E. (1988). Development of the NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 138-183). Amsterdam: North-Holland Press.

Hoepf, M., Middendorf, M., Epling, S. & Galster, S. (this volume). Physiological indicators of workload in a remotely piloted aircraft simulation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State Universtiy.

Jasper, H. (1958). Report of the committee on methods of clinical examination. *Electroencephalography and Clinical Neurophysiology, 10,* 370-375.

Middendorf, M., Epling, S., Hoepf, M., & Galster, S. (this volume). Saccade detection using polar coordinates – A new algorithm. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Recarte, M., Perez, E., Conchillo, A., & Nunes, L. (2008). Mental workload and visual impairment: Differences between pupil, blink and subjective rating. *The Spanish Journal of Psychology, 2,* 374-385.

Tsang, P., & Wilson, G.F. (1997). Mental workload. In: Salvendy G (Ed.), *Handbook of Human Factors and Ergonomics.* Baltimore, MD: John Wiley.

Wang, Y., & Zhou, J. (2013). Literature review on physiological measure of cognitive workload. Machine learning research group – NICTA.

A COALITION STUDY OF WARFIGHTER ACCEPTANCE OF WEARABLE PHYSIOLOGICAL SENSORS

Lauren E. Menke[1], Christopher Best[2], Gregory J. Funke[3], Adam J. Strang[3]

[1]Ball Aerospace & Technologies Corporation, Fairborn, Ohio
[2]Defence Science and Technology Organisation, Melbourne, Australia
[3]Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio

Combat operations are often high tempo, resulting in undesirable levels of operator workload and stress. Adaptive automation has been suggested as a solution to these issues. However, this augmentation approach is predicated on operator consent to monitoring. Acceptance of such systems may be influenced by concerns regarding the use of monitor data and mistrust of automation technology. The purpose of the current investigation was to examine operator acceptance of physiological monitoring and future augmentation strategies after limited exposure to one device. During a simulated exercise, eleven command and control operators were equipped with a physiological monitor prior to each mission. Following the exercise, operators were surveyed regarding their acceptance of monitoring and several potential augmentation strategies. The results of the survey suggested that the operators were generally open to both monitoring and augmentation, but that they may also be insensitive to the limitations of current augmentation technology.

Military teams face increasingly difficult situations, characterized by high tempo operations, distributed team environments, long shift durations, high information throughput, and decision making under uncertainty (Chappelle et al., 2013). Concurrently, technological advances (e.g., for surveillance and monitoring, and cyber defense) are increasingly providing capabilities that will require rapid data processing and decision speeds that exceed human capabilities (e.g., Dahm, 2010). It has been suggested that factors such as these may result in human operators becoming a "bottleneck" in future military operations (e.g., Dorneich, Whitlow, Ververs, & Rogers, 2003).

In response to this challenge, military strategic guidance and planning documents (e.g., Dahm, 2010) suggest that human augmentation solutions need to be developed. A potential solution that has been suggested is adaptive automation (e.g., Dorneich et al., 2003). Adaptive automation is predicated on activation of assistive functions based on cues derived from operator behavior or physiology. Of particular interest is automation that is part of a data-driven feedback loop, wherein monitoring technologies track and assess physio-behavioral changes indicative of negative operator states (e.g., mental workload and fatigue; Galster & Johnson, 2013). This information can then be shared (e.g., with operators, mission commanders, other automated systems, etc.) as part of an augmentation strategy, perhaps resulting in dynamic task reallocation. However, for this approach to be viable, operators will need to be monitored during task performance. If risk factors such as stress and fatigue are to be considered (Caldwell, Caldwell, & Schmidt, 2008), extensive monitoring may be required, possibly including off-duty hours. These monitoring approaches will certainly require operator consent. This need resonates with the concept of *ubiquitous monitoring*, suggested by Moran and colleagues (2013), wherein behavioral, and potentially physiological, data are collected continuously from individuals for the purpose of monitoring and targeted intervention. Therefore, the success of adaptive automation as an augmentation strategy is contingent on operators' acceptance of both monitoring and automation.

With regard to monitoring, research suggests that operator acceptance is likely to fall on a continuum of responses. At the "low" end of acceptance, participants may feel that monitoring is intrusive and reduces privacy (similar concerns have been raised regarding telemedicine, e.g., Beckwith, 2003; Rackett, 1997). For example, operators may fear the unwanted disclosure of health related information as a result of monitoring, particularly amid ongoing concerns regarding data privacy (e.g., Ahamed, Talukder, & Kameas, 2007). In addition, operators (e.g., aircrews) may fear that their duty status could be negatively affected by the discovery of ill-health.

A further concern may be feelings of discomfort or anxiety associated with perceptions of the presence of an evaluative "other," such as a superior, colleague, or the monitoring system itself (e.g., Zeidner & Matthews, 2005). Research on evaluation anxiety (e.g., Zeidner & Matthews, 2005) suggests that under some circumstances, operator worries about evaluation may result in sufficient distraction to negatively impact task performance. The behavioral or physiological symptoms of such worries could result in the activation of the augmentation system, which in turn could reinforce and amplify their initial worries – creating an ongoing cycle of distraction and poor task performance.

At the "high" end of the acceptance continuum, operators may respond positively to continuous monitoring, particularly if they perceive that the benefits of the technology outweigh the risks (Moran et al., 2013). This may well be the case for military operators, considering that they are likely to a) be aware of emerging military doctrine concerning current and future reliance on automated systems, and b) have been affected by the difficult circumstances of current combat operations described previously. Operators may also endorse monitoring technologies if they are offered the opportunity to utilize the recorded physiological information for their own purposes, such as fitness or health management (e.g., Heron & Smyth, 2010).

An additional influencer of operator attitudes may be past experience with automated systems. For some operators, negative experiences with automation reliability (e.g., Parasuraman & Riley, 1997) and automation surprise (e.g., Sarter, Woods, & Billings, 1997) may elicit a general distrust in automation. There is also evidence that, under some circumstances, automation may actually increase operator workload, potentially resulting in operator underuse or disuse of augmentation technologies (Parasuraman & Riley, 1997). Finally, operators may have little understanding of the state of current automation technologies, and therefore have unrealistic expectations concerning system capabilities. Informed by popular media coverage, movies, and television, operators may believe that contemporary monitoring and automated augmentation technologies are more robust and advanced than they actually are. Similar beliefs have been expressed regarding perceptions of the capabilities of modern robots (e.g., Adams & Skubic, 2005).

Given these concerns, the purpose of the current experiment was to gauge operator opinions regarding their acceptance of monitoring and endorsement of several potential augmentation approaches. Participants in this study were a small group of Air Battle Manager (ABM) operators from the Royal Australian Air Force (RAAF) selected to take part in Exercise Black Skies (EBS; Best, Jia, & Simpkin, 2013). As part of the exercise, operators consented to physiological monitoring, providing them (limited) experience with monitoring upon which to base their ratings. We expected that operators would express general agreement to monitoring while performing their duties, and more limited approval of several augmentation strategies. Furthermore, we expected that support for specific strategies would be moderated by operational environment, with the highest endorsement during training, and reduced acceptance in more "real world" settings, such as combat missions. This would indicate a general openness of operators to emerging technologies, tempered by veridical assessment of current monitoring and augmentation capabilities and limitations.

Methods

Overview of Exercise Black Skies 2014 (EBS14)

Exercise Black Skies is a 5-day simulation training research exercise hosted by the Defence Science and Technology Organisation (DSTO) at their Air Operations Simulation Centre in Melbourne, Australia. While the specific training audience and scenarios are unique for each biannual instantiation of EBS, the broader objectives remain the same, which are to: 1) provide high-fidelity training to prepare ABM operators for a subsequent multinational, live training exercise (Exercise Pitch Black), and 2) serve as a test-bed for the development and evaluation of emerging technologies that might benefit current and future ABM operations and training.

The training audience for EBS14 included two sub-teams of ABM operators: a ground-based ABM unit (specifically, an Air Defence Ground Environment, or ADGE, unit) and an airborne unit (a mission crew from the E-7A "Wedgetail" airborne early warning and control aircraft). Participants in the exercise were 10 men and 1 woman. Their average age was 29.64 years (SD = 6.40; M_{ADGE} = 27.50, SD_{ADGE} = 6.47; $M_{Wedgetail}$ = 32.20, $SD_{Wedgetail}$ = 5.89). The ADGE team was composed of an Air Battle Director (ABD), a Tactical Director (TD), two Fighter Controllers (FCs), and two Picture Managers (PICMAN). The Wedgetail team was composed of a Mission Commander (MC), a Senior Surveillance and Control Officer (SSCO), and three Surveillance and Control Officers (SCOs).

Within these teams, the ABD, TD, MC and SSCO roles were leadership/supervisory roles, with the ABD and MC roles filled by the most experienced members (with 4,500 and 2,000 hours of controlling experience, respectively). The TD and SSCO roles were filled by the next most experienced operators (with 837 and 700 hours of controlling experience, respectively). The FC and SCO roles were tasked with tactical control of the aircraft within the team's assigned airspace. Operators in these roles had less experience (averaging approximately 250 hours of controlling experience). The operators filling the PICMAN roles reported an average of approximately 4,000 hours experience.

While the functions and mission objectives of the two teams were mostly similar, there were several notable differences in their working environments. First, because different Command and Control interface systems are used by the RAAF in real ground-based and airborne environments, these systems were also different for the two EBS14 teams. Second, the physical configuration of the simulation facilities reflected those of each team's typical work environment; the ADGE team sat in a semi-circular "weapons pit" arrangement of two rows (with team leaders

seated behind members responsible for tactical control) while the Wedgetail team sat side-by-side (as is typical of the seating arrangement on the aircraft).

During EBS14, other command and control elements, as well as friendly and adversary airborne assets (e.g., fighter aircraft, air-lift aircraft, tankers), were simulated by an exercise "White Force" consisting of RAAF personnel and ex-military contractors. An important characteristic of EBS14 was that the mission scenarios used during the exercise were designed to simulate, in terms of airspace structure, airfield, target and sensor locations, friendly and adversary order of battle, mission types and unit roles, those that the operators would encounter several weeks later during the live exercise Pitch Black. This is noteworthy since Pitch Black is the RAAF's largest and most complex air-combat exercise, making EBS14 a large, complex, and realistic simulation training event.

Physiological Monitoring System

During EBS14, operators consented to physiological monitoring of their responses to events in the simulation. They were told the information would be used to shape future simulation exercises and to develop augmentation technologies. It should be noted that although operators were provided an explanation for the physiological monitoring they experienced during EBS14, they were not provided information or feedback about their or their teammates' particular physiological responses during the exercise, nor were they provided information about specific future augmentation technologies that might rely on such data.

Each operator wore a Zephyr BioHarness 3 (model BH3) during the exercise. The BioHarness is a lightweight physiological sensor designed to be worn against the wearer's chest by means of a flexible synthetic strap (see Figure 1 for an illustration). The device was applied in accord with Zephyr's instructions, i.e., the chest strap was aligned with the bottom of the operator's sternum, and the recording module was located on the left side of the body in line with the operators' armpit or slightly rotated to the back for comfort. The BioHarness records electrocardiographic (ECG), respiration, and accelerometry data (at 250, 100, and 25 Hz, respectively) and provides summary statistics once per second. Raw and summary data were recorded throughout each session to the onboard memory of the recording module. At the end of each session, data were downloaded from each operator's module to a central database.

Figure 1. Zephyr BioHarness. The left image portrays the recording module (circular disk) and harness. The right image depicts proper placement of the harness.

Device Comfort Questionnaire (DCQ)

Following the final trial of EBS14, participants completed a novel measure, the *Device Comfort Questionnaire* (DCQ; see Appendix A). The DCQ is comprised of 19 items, representing 5 related subscales. Items of the first subscale, *device ergonomics*, relate to fit factors, such as simplicity of application and interference with task performance. The second subscale, *acceptance of physiological monitoring*, includes items related to operators' perceptions of discomfort and intrusiveness associated with being monitored. Items of the final 3 subscales, *endorsement of use during simulation training exercises, live training exercises,* and *real operations*, ask operators to rate their degree of predicted acceptance of a future augmentation technology designed to utilize physiological monitoring data for a variety of purposes, including automatic adjustment of task difficulty, performance assessment, and workload monitoring. Items on the DCQ are rated on a scale of 1 ("Completely Disagree") to 10 ("Completely Agree"). After reverse scoring relevant items (see Appendix A), subscale scores on the DCQ are computed by averaging across the pertinent item ratings.

Results

Mean operator ratings on the DCQ are presented below in Table 1, which depicts ratings aggregated (based on team role) into the categories of *lead* (ABD and TD, MC and SSCO) and *tactical* (FC and PICMAN, SCO) for the ADGE and Wedgetail teams, respectively.

Table 1.
Mean DCQ ratings by sub-team and team role.

	Team						
	ADGE			Wedgetail			
Mean Subscale and Item Responses	Lead	Tactical	**Mean$_{ADGE}$**	Lead	Tactical	**Mean$_{Wedge}$**	**Grand Mean**
Device ergonomics (Mean)	**8.67**	**7.92**	**8.29**	**7.17**	**7.44**	**7.31**	**7.80**
1. Not hindered performing duties	9.00	7.75	**8.35**	9.50	8.67	**9.08**	**8.73**
2. Device did not cause discomfort	9.00	8.75	**8.88**	5.00	8.00	**6.50**	**7.69**
5. Easy to put on and take off	8.00	7.25	**7.63**	7.00	5.67	**6.33**	**6.98**
Acceptance of physiological monitoring (Mean)	**10.00**	**9.63**	**9.81**	**9.00**	**9.00**	**9.00**	**9.41**
3. Device was not intrusive	10.00	9.75	**9.88**	8.50	8.67	8.58	**9.23**
4. Comfortable being monitored	10.00	9.50	**9.75**	9.50	9.33	9.42	**9.58**
Simulation training (Mean)	**8.17**	**9.50**	**8.83**	**6.17**	**7.44**	**6.81**	**7.82**
6. White force sets training difficulty	9.00	9.50	**9.25**	8.50	7.67	**8.08**	**8.67**
7. Identify debrief points	10.00	9.50	**9.75**	5.50	7.67	**6.58**	**8.17**
8. Automatically set training difficulty	6.00	9.50	**7.75**	5.50	7.67	**6.58**	**7.17**
9. Assessors make judgments	9.50	9.50	**9.50**	6.50	7.00	**6.75**	**8.13**
10. Inform lead about workload	9.00	9.50	**9.25**	6.50	7.67	**7.08**	**8.17**
Field exercise (Mean)	**8.00**	**9.50**	**8.75**	**6.00**	**7.44**	**6.72**	**7.74**
11. White force sets training difficulty	6.00	9.50	**7.75**	7.50	7.67	**7.58**	**7.67**
12. Identify debrief points	10.00	9.50	**9.75**	4.50	7.67	**6.08**	**7.92**
13. Automatically set training difficulty	6.00	9.50	**7.75**	5.00	7.67	**6.33**	**7.04**
14. Assessors make judgments	9.00	9.50	**9.25**	6.50	7.00	**6.75**	**8.00**
15. Inform lead about workload	9.00	9.50	**9.25**	6.50	7.67	**7.08**	**8.17**
Live operation (Mean)	**8.67**	**9.50**	**9.08**	**6.50**	**7.44**	**6.97**	**8.03**
16. Identify debrief points	9.50	9.50	**9.50**	8.00	7.67	**7.83**	**8.67**
17. Assessors make judgments	9.00	9.50	**9.25**	5.00	7.00	**6.00**	**7.63**
18. Inform lead about workload	8.50	9.50	**9.00**	6.50	7.67	**7.08**	**8.04**
19. Inform lead about fatigue	8.50	9.50	**9.00**	8.00	7.67	**7.83**	**8.42**
Grand Mean	**8.68**	**9.26**	**8.97**	**6.82**	**7.67**	**7.24**	

Perusal of Table 1 reveals several interesting effects. First, operators' ratings of the Zephyr BioHarness's *device ergonomics* were relatively high. Second, the ABMs indicated they were overwhelmingly accepting of the physiological monitoring they experienced during EBS. Third, when operators were asked to speculate about the future uses of physiological monitoring for adaptive aiding, they expressed high positive endorsement for the monitoring irrespective of the purpose or operational setting that the monitoring would be employed.

To further examine the data in Table 1 for differences in ABM operator ratings based on team and role across subscales, a 2 (team) × 2 (team role) × 5 (subscale) mixed analysis of variance (ANOVA) was computed. Results indicated a statistically significant main effect of team, F (1, 7) = 31.06, $p < .05$, $\eta_p^2 = .816$. No other effects in the analysis were statistically significant (all $p > .05$). Members of the ADGE team consistently provided higher agreement ratings on DCQ items than members of the Wedgetail team.

Discussion

The purpose of the current study was to provide an initial examination of operator response to physiological monitoring and potential future performance augmentation strategies. We expected that operators would express general agreement to monitoring while performing their duties, and more limited approval of the augmentation strategies. Further, we expected support for specific strategies would be moderated by operational environment. Our results suggest that operators were generally accepting of monitoring and endorsed the prospective augmentation strategies uniformly across operational environments. We also found that operator acceptance and endorsement was moderated by team; ADGE operators indicated higher agreement across items than did Wedgetail operators.

Contrary to initial predictions, the ABM operators were relatively accepting of physiological monitoring and agreed to usage of that data for all of the purposes and environments proposed. This may indicate that the

perceived benefits of the proposed technology outweighed the perceived risks. Alternatively, it could suggest that operators may be unfamiliar with the capabilities and limitations of current (and near future) automated augmentation technologies. Whatever the underlying drivers may be, one consequence is relatively clear: operators are positive about future developments in monitoring and augmentation. It therefore behooves those of us working in the area to ensure that their expectations are appropriately calibrated against the actual capabilities of the systems we develop. Failure to do so is likely to result in violated expectations, mistrust, and disuse of future augmentation solutions.

Surprisingly, we found that Wedgetail operators expressed less acceptance of monitoring and augmentation than ADGE team operators. Though explanation of this effect is speculative, it could be due to disparities in operator experience across teams (ADGE team operators were generally more experienced than their Wedgetail counterparts), or other structural differences between the two groups. For example, the Wedgetail is a relatively new platform in the RAAF, and consequently those operators' attitudes may have been influenced by other factors, such as evolving organizational structure, mission requirements, and the relatively negative history associated with development of the aircraft (see e.g., Bergmann, 2013, for a brief history).

Alternatively, the observed differences in ratings may reflect differences in attitudes regarding deployment of electronic equipment. Wedgetail operators' ratings may be due to worries concerning electronic interference or safety considerations around wearing equipment in flight (e.g., it could hinder movement in the event of an emergency) – these are concerns that ADGE team operators would not necessarily share because of the ground-based nature of the unit. Yet, this explanation does not fully explain Wedgetail operator attitudes, as their ratings on the simulation training subscale of the DCQ were also lower than ADGE operators, even though simulation training exercises are not conducted on an aircraft. This may indicate that Wedgetail aircrew training has broadly sensitized them to issues of electronic interference and/or safety regardless of operational setting. This explanation has implications for other work environments. For example, if the Wedgetail operators' less-positive responses were driven by concerns about restricted movement during a crash or emergency egress, other aircrew may have similar concerns (e.g., fighter aircraft with ejection seat).

Overall, the operators surveyed in this experiment expressed high positive regard for future monitoring and augmentation approaches. Though substantial work is required to mature those technologies, it appears that operators are generally ready to accept them. In developing these devices, care must be taken to ensure that the capabilities and limitations of any such systems are communicated to operators, thereby appropriately calibrating their trust in and expectations of those devices.

References

Adams, J. A., & Skubic, M. (2005). Introduction to the special issue on human-robot interaction. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 35*, 433-437.

Ahamed, S.I., Talukder, N., & Kameas, A.D. (2007, September). *Towards privacy protection in pervasive healthcare*. Paper presented at the 3rd IET International Conference on Intelligent Environments (IE 07), Ulm, Germany. Paper retrieved from https://www.cs.purdue.edu/homes/ntalukde/papers/IE07.pdf

Beckwith, R. (2003). Designing for ubiquity: The perception of privacy. *Pervasive Computing, 3*, 40-46.

Bergmann, K. (2013, April 8). Wedgetail. Asia-Pacific Defense Reporter. Retrieved from http://www.asiapacificdefencereporter.com/articles/298/Wedgetail

Best, C., Jia, D., & Simpkin, G. (2013, October). *Air force synthetic training effectiveness research in the Australian context*. Proceedings of the NATO STO MSG 111 Multi-Workshop, Sydney, Australia.

Caldwell, J. A., Caldwell, J. L., & Schmidt, R. M. (2008). Alertness management strategies for operational contexts. *Sleep Medicine Reviews, 12*, 257-273.

Chappelle, W., McDonald, K., Christensen, J., Prince, L., Goodman, T., Thompson, W., & Hayes, W. (2013). Sources of occupational stress and prevalence of burnout and clinical distress among U.S. Air Force Cyber Warfare Operators (Report No. AFRL-SA-WP-TR-2013-0006). Wright-Patterson Air Force Base: Air Force Research Laboratory, Human Effectiveness Directorate.

Dahm, W.J.A. (2010). *Technology horizons: A vision for air force science & technology during 2010-2030* (Report No. AF/ST-TR-10-01). USAF HQ, Arlington, VA.

Dorneich, M.C., Whitlow, S.D., Ververs, P.M., & Rogers, W.H. (2003, October). *Mitigating cognitive bottlenecks via an augmented cognition adaptive system*. Proceedings of the 2003 IEEE International Conference on Systems, Man, and Cybernetics, Washington, DC.

Galster, S.M., & Johnson, E.M. (2013). *Sense-assess-augment: A taxonomy for human effectiveness* (Report No. AFRL-RH-WP-TM-2013-0002). Wright-Patterson Air Force Base: Air Force Research Laboratory, Human Effectiveness Directorate.

Heron, K.E., & Smyth, J.M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behavior treatments. *British Journal of Health Psychology, 15*, 1-39.
Moran, S., Jaeger, N., Schnadelbach, H., & Glover, K. (2013, June). *Using adaptive architecture to probe attitudes towards ubiquitous monitoring*. Proceedings of the 2013 IEEE International Symposium on Technology and Society (ISTAS), Toronto, Canada.
Parasuraman, R., & Riley, R. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*, 230-253.
Rackett, C.M. (1997). Telemedicine today and tomorrow: Why "virtual" privacy is not enough. *Fordham Urban Law Journal, 25*, 167-191.
Sarter NB, Woods DD, Billings CE. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 1926-1943). New York, NY: Wiley.
Zeidner, M., & Matthews, M. (2005). Evaluation anxiety: Current theory and research. In A.J. Elliot & C.S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 141-163). New York, NY: The Guilford Press.

Appendix A: Device Comfort Questionnaire (DCQ)

Instructions to participants:

You wore a physiological monitoring device (i.e., a Zephyr BioHarness 3) during each VUL [trial] in EBS14. The purpose of these devices was to help us monitor how hard you were working during each VUL, with the idea that we could use that information to help shape future Black Skies exercises. Given that, in the following questions we are interested in your level of comfort wearing the system.

Please rate the following statements about the device on a scale from:
1 = "Completely Disagree" to **10 = "Completely Agree"**

1. I was not hindered by the device while performing my duties.
2. I felt that wearing the device caused discomfort.*
3. I felt that wearing the device was intrusive.*
4. I felt uncomfortable being monitored.*
5. I felt that the device was easy to put on and take off.

I would feel comfortable having physiological data, such as that collected during EBS14, used in a future **simulation training exercise** (e.g. Black Skies) to:

6. Help white force set or change training difficulty.
7. Help identify debrief points for after action review.
8. Automatically set or change training difficulty.
9. Help expert assessors make judgments about my performance.
10. Help inform my team leader about my workload during a mission.

I would feel comfortable having the physiological data used in a **live training exercise** (e.g. Pitch Black) to:

11. Help white force set or change training difficulty.
12. Help identify debrief points for after action review.
13. Automatically set or change training difficulty.
14. Help expert assessors make judgments about my performance.
15. Help inform my team leader about my workload during a mission.

I would feel comfortable having the physiological data used during **real operations** to:

16. Help identify debrief points for after action review.
17. Help expert assessors make judgments about my performance.
18. Help inform my team leader about my workload during a mission.
19. Help inform my team leader about my level of fatigue during a mission.

Note. Items marked with an asterisk (*) are reverse scored.

Scoring the DCQ

The DCQ includes five dimensions: *device ergonomics* (mean rating of items 1, 2, & 5); *acceptance of physiological monitoring* (mean rating of items 3 & 4); *endorsement of use during simulation training exercises* (mean rating of items 6-10); *endorsement of use during live training exercises* (mean rating of items 11-15); *endorsement of use during real operations* (mean rating of items 12-19).

MODELING TASK PRIORITIZATION BEHAIVORS IN A TIME-PRESSURED MULTITASKING ENVIRONMENT

Takeaki Toma
School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University
Corvallis, Oregon
Kenneth H. Funk II
School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University
Corvallis, Oregon

Cockpit task management (CTM) theory is structurally consistent with cognitive multitasking models. Based on the CTM framework, it is hypothesized that aviation task prioritization behavior in human multitasking may be influenced by importance, urgency, performance status, salience, and workload of tasks in a cockpit. A middle fidelity flight simulation study was conducted to test the above hypotheses. Questionnaire data indicated that the perceived task importance, the perceived task urgency and the perceived task salience had significant relationships with the perceived task priority after taking the individual difference and flight situational difference into account. The perceived task priority was related to the task execution time and task performance, but not correlated with task awareness level in the flight simulation.

Introduction

Human multitasking in transportation can be dangerous. A recent National Safety Council (NSC) white paper (2010) reported that 25% of all car crashes were caused by the use of a mobile device while driving, and an estimated 1.4 million crashes and 645,000 injuries were related to multitasking. Relatedly, many aviation incidents and accidents are reported to occur during multitasking. Chou, Madhavan, and Funk (1996) reviewed National Transportation Safety Board (NTSB) aircraft accident reports and NASA Aviation Safety Reporting System incident reports using cockpit task management (CTM) theory (Funk, 1991); they reported that 23% of aviation accidents and 49% of aviation incidents were rooted in CTM errors.

One reason for dangerous multitasking in transportation operations might be the difficulty of task prioritization. A well-known example is the crash of Eastern Air Lines Flight 401 on December 29, 1972 in the Florida Everglades. When the airplane was approaching the Miami airport, the pilots noticed that a landing gear indicator light did not turn on. The pilots communicated with the approach controller who gave a clearance to maintain an altitude of 2000 feet and hold west over the Everglades. The cockpit crew proceeded to put the plane in autopilot control and believed that the airplane was holding at an altitude of 2000 feet. However, the pilots did not notice that the autopilot setting had changed (probably due to a pilot mistakenly moving the yoke) and the airplane gradually descended towards the ground (NTSB, 1973). The root cause of the accident can be interpreted as a task prioritization error because the pilots wrongfully prioritized their attention to the landing gear problem instead of controlling the airplane; the pilots did not pay attention to the altitude because they focused too much on the landing gear problem (NTSB, 1973). This incident raises several questions: Why did the pilots prioritize diagnosing the landing gear problem higher than controlling the airplane? What is the mechanism behind task prioritization in aviation multitasking environments? We hypothesized that the aviation task prioritization process is influenced by five factors: task importance, task urgency, task status, task salience, and multitasking workload based on the antecedent study conducted by Colvin, Funk and Braune (2005). The objective of this research was to test the above hypotheses.

Potential Task Prioritization Factors

How should pilots manage multiple tasks in time-pressured and dynamic situations? How should pilots prioritize tasks in cockpits? Using a systems engineering approach, Funk (1991) developed Cockpit Task

Management (CTM) theory as "the process by which the flight crew manages an agenda of cockpit tasks" (p. 277). Funk's (1991) CTM theory is structurally consistent with cognitive models, such as Wicken et al.'s (2013) human information processing stage model and Endsley's (1995) situation awareness (SA) model. For example, CTM theory has situation awareness stages (CTM 2.a, 2.c, and 2.e) that correspond to Level-2 and Level-3 in the SA model. On the other hand, while the resource scope in CTM theory includes human resources (pilots) and equipment resources (autopilots, radios, displays and controls), the scope of SA theory and other human cognitive theories focus only on human sensory cognitive and motor resource. Chou et al. reported that at least two types of CTM errors (task initiation errors and task prioritization errors) occurred when the required cognitive resource was large and the number of concurrent tasks and task difficulty (flight path complexity) were high. Chou et al. (1996) concluded that high workload generated CTM errors and proposed the necessity for pilots to develop strategies to predict and handle high workload situations. Wilson (1998) and Funk, et al. (1999) reported that the automation of pilots' tasks might increase task prioritization errors. For example, inappropriately designed automation may make it difficult for pilots to detect, diagnose and evaluate the consequence of automation failures.

Then what is the limit of human ability to compute task priority? Shakeri and Funk (2007) tested how people can calculate the tradeoff among their CTM task prioritization criteria (importance, urgency and status of tasks) in multitasking with a juggler's paradigm. A participant monitored six hypothetical tasks on a computer screen. The status and urgency of each task was displayed with a bar similar to a battery icon charging level bar. Each task had a different level of importance, urgency, and status property to be taken into consideration for task prioritization, and reported these four main findings. First, the participants were not able to achieve perfect task-prioritization (they scored 71% to 87% in task prioritization performance to perfect task-prioritization of a near-optimal algorithm). Second, the participants were more aware of the importance of tasks that were static and displayed on the screen and failed to recognize the dynamically changing urgency or status of tasks that required mental computation and prediction. Third, the participants overemphasized the penalty score of the task-prioritization decision, which indicated the difficulty participants had ignoring the salient task; and fourth, participants "learned" which tasks should be prioritized more highly than others (strategic task management).

What are the most relevant and useful factors for the rational design of the aircraft cockpit? Colvin, Funk and Braune (2005) conducted a flight simulator study in which they asked pilot participants to report the six primal candidate factors they used for task prioritization. The first factor reported was the perceived salience of stimuli that relates to a task. Colvin et al., hypothesized that "the priority of a task is directly proportional to its salience" . Shakeri and Funk (2007) reported that people could not ignore salient stimuli in task prioritization. Furthermore, if task-related stimuli are not salient, inattentional/change blindness phenomena may occur (Simons & Chabris, 1999; Simons and Levin 1998; Strayer, Drews & Johnston, 2003). The salience of task-related stimuli are also a factor in visual attention prioritization in the SEEV model (Wickens, Helleberg, Horry & Talleur, 2003). The second factor participants reported in Colvin et al.'s study was the perceived importance (or value) of a task. They hypothesized that "the priority of a task is directly proportional to its importance" (Colvin et al., p334). The importance (or value) criterion is often used for tradeoffs in multi-attribute utility decisions between pros (importance or value) against costs. For example, Kushleyeva, Dario, Salvucci, and Frank (2005) constructed a task-prioritization model that trades off a cost factor and the value factor multiplied by their probabilities, and Wickens et al. (2003) and Funk (1991) used importance / value factor for a visual attention prioritization model and the CTM theory, respectively. The third factor reported by Colvin et al.'s participants was the perceived performance status of a task. Colvin et al. hypothesized that "the priority of a task is directly proportional to its importance" (p335). The perceived status of a task may influence the task prioritization decision, because understanding of the current status is situation awareness Level-2, which is a foundation of sound decision-making (Endsley, 1995). Wickens (2003) noted that pilots need to have situation awareness of task status for sound decision-making along with spatial awareness and system awareness. Furthermore, Altman and Trafton (2002) reported that people tend to forget to recall, resume, and execute tasks in a suspended status. For example, the pilots of Spanair Flight 5022 forgot to complete a flaps checklist item while taxiing, resulting in its crash. Thus even high priority tasks may be forgotten and observers

(e.g., accident investigators) may regard it as an "inappropriate task prioritization decision". Furthermore, the status of tasks (CTM 2.c) is used as a task prioritization factor in CTM theory (Funk, 1991). The fourth potential task prioritization factor is the perceived urgency of a task. Colvin et al. hypothesized that "the priority of a task is directly proportional to its urgency" (p335). The perceived urgency of task may be defined as the buffer time, or time remaining until the deadline of the task (Wickens, et al., 2013), which reflects the projected situation awareness of Level-3 SA (Endsley, 1995). The lack or inappropriate perception of task urgency may lead to fatal aviation accidents (e.g., In the Flight 401 accident case, the pilots did not notice the urgency of the task). The Threaded Cognition Multitasking model uses the task urgency factor for task prioritization (Salvucci & Taatgen, 2011; Salvucci, Taatgen and Borst, 2009), and Funk (1991) also use it as the task prioritization factor in CTM. The fifth potential factor is the expectation of a task. Colvin et al. (2005) hypothesized that "the priority of a task is directly proportional to its consistency with procedure or with other pilot expectations" (Colvin et al., p335). Simons and Chabris (1999), and Simons and Levin (1997) showed that lacking the expectation of stimuli may generate inattention/change blindness cognitive phenomena. Endsley (1995) argued that an expected mental model (i.e., expectation) affects situation awareness that will influence where attention is directed and how perceived information is interpreted in a top-down cognitive process. Furthermore, the projected future situation of the environment (SA level-3) will become an expectation that affects the above points (Endsley, 1995). Thus, the expectation factor is used in prioritization models (e.g., Wickens et al., 2003). The sixth potential factor is the perceived cost or effort of a task or its workload. Colvin et al. (2005) stated that "the priority of a task is proportional to the time/effort required to perform it" (Colvin et al., p336). Chou et al. (1996) reported that high workload adversely affects multitasking performance. Furthermore, high workload or high switching costs of tasks (e.g., Allport 1994) will delay the execution of prioritized tasks in time (e.g., Lee, McGehee, Brown, & Reyes, 2002). When people cannot make adequate progress in concurrent multitasking they may suspend one or more tasks for later resumption (e.g., Altman & Trafton, 2002; Salvucci, Taatgen & Borst, 2009). As mentioned before, outside observers (e.g., accident investigators) may potentially regard delayed execution as inappropriate task prioritization decisions in incidents or accidents (e.g., Flight 401; NTSB, 1973). Generally, cost (effort or workload) is considered as a tradeoff against value (i.e., importance) in multi-attribute utility decisions, and several human multitasking models utilize "cost" as a task prioritization factor (e.g., Wickens et al., 2013, Kushleyeva et al. 2005).

Research Questions and Methodology

Based on important insight from the above literature, the following research questions were raised about task prioritization behavior in aviation human multitasking. **Research question-1**: Can perceived task priority be explained by the following five factors? 1. perceived importance of tasks, 2. perceived urgency of tasks, 3. perceived status of tasks, 4. perceived salience of tasks, 5. perceived workload. **Research question-2**: Is the perceived priority of a task consistent in the level of task awareness and task execution performance?

Sixteen pilots (15 males and 1 female) were recruited for a flight simulation experiment. All the pilots possessed a private pilot license with an average of 4,508 hours of total flying experience (minimum 65 hours, maximum 31,000 hours) and an average of 2,274 hours of single pilot hours. Their ages were between 24 and 82 and average age was 49.3. They were compensated for 2 hours of data collection (duration of the experiment) with a $25 gift card. The experiment was conducted at the Human Factors Engineering Laboratory at Oregon State University. Subjects flew a Cessna 172 RG airplane in a X-plane®-based general aviation flight simulator. In order to run the experiment and collect multitasking behavioral data, the following instruments were prepared: a computer-synthesized voice for the ATC communication (operated by the experimenter), a flight checklist, and flight charts. Each participant practiced the X-plane flight simulator, Air Traffic Control (ATC) communication system, the flight plan, and aircraft checklists to become familiar with the system before the start of the simulation. In this practice session, the participant was allowed to fly in the simulator in a similar condition to the experiment. After the practice was completed, the participant could ask for clarification regarding simulator operation. The second session was data collection using the flight simulator. Flight data (e.g., headings, altitudes, airspeeds, engine

parameters, radio frequencies, and flight control movements) were automatically recorded by the simulator. Every few minutes the flight simulation was frozen and the participant was asked to rate his or her perception of the importance of tasks, urgency of tasks, status of tasks, salience of tasks, workload of tasks, and perceived priority of tasks at that moment. Here, we followed Endsley (1995b)'s query guideline such that the timing of each freeze for query was randomly determined at each experimental block. Because of its unpredictable interruption, it was assumed that the participants could not anticipate or prepare for queries beforehand, which could provide unbiased estimates of the participant's task-prioritization decisions. Furthermore, behavioral audio and video data were recorded throughout the flight. A simple but challenging flight scenario was prepared. After becoming familiar with the flight simulation in the practice session, pilots conducted a simulated flight scenario using two VORs (VHF Omni Directional Radio Range). Pilot participants communicated with the experimenter who played the role of the ATC controller with the computer-synthesized voice based on a predetermined communication script. The pilot was reminded that in the simulation, flight safety was the ultimate goal, and it was assumed that the participating pilots prioritized among four tasks: Aviate (vertical control), Navigate (lateral control), Communicate, and Manage Systems. In each of eight situations, expected and unexpected flight instrument problems challenged the participants. Repeating problems included those problems that pilots could expect to happen because they reoccurred multiple times in the scenario. Those problems were Pitot tube clog that caused a airspeed indicator malfunction in situations 2, 4, and 7; a low fuel problem occurring in situations 6, 7, and 8; and an altimeter malfunction in situations 4 and 8. Non-repeating problems included those problems that pilots could not easily expect to happen because each problem occurred only once. Those problems were artificial horizontal indicator malfunction, vacuum pump indicator, vertical speed indicator, and navigation instrument malfunctions. No information was given to pilots about which problem(s) might occur or repeat in the scenario.

To answer the first research question, each pilot's perceptions of task priority and five hypothesized prioritization criteria were obtained by asking the probe questions on Table 1. The obtained perceived task priority, all response variables and explanatory variables were numerically coded for statistical hypotheses tests. The perceived task priority score was modeled using five explanatory variables with regression models, and each factor's effect was estimated with the corresponding coefficient in mixed models. Here, each of the five task-prioritization decision criteria was used as the fixed effect. Intercepts for subjects and by-subjects slopes were used as random effects in the model. A likelihood ratio test was used for testing a linear relationship between the perceived priority task priority and each of the five task prioritization decision criteria.

Table 1.
Variables and Probe Questions in the questionnaires.

Variables	Probe Questions
Perceived Task Priority Score (Y)	Which task did you prioritize at this moment? (pair-wise comparison)
Perceived Task Importance Score ($X1$)	Based on your comprehension of the current situation, which task was more important? (pair-wise comparison)
Perceived Task Urgency Score ($X2$)	Based on your projection of the future status, rate the urgency of each task by its "buffer time"; the amount of time you could delay the task before it requires your attention to maintain safe flight.
Perceived Task Performance Score ($X3$)	Based on your comprehension of the current situation, rate the performance of task, how successful you believe in accomplishing the goal of the task set by yourself?
Perceived Task Salience Score ($X4$)	Based on your current perception, which task was more salient and drew your attention at the moment? (pairwise comparison)
Perceived Task Workload Score ($X5$)	NASA's TLX workload questionnaire applied to each task singly

To answer the second research question regarding the relationships between perceived task priority and task awareness and task execution, subjects were instructed to verbally report as soon as they noticed any problems or

abnormal situations during the flight simulation. The unnoticed time period was measured from the recorded video for estimating the level of task related problem awareness. The directional deviation and altitude deviation from the target path were obtained from the flight recorder.

Results

There were clear linear relationships between the perceived task priority and the perceived task importance score, the perceived buffer time (i.e., task urgency), and the perceived salience score. There was not enough evidence to conclude the existence of linear relationships on the perceived performance status score and the perceived workload score Table 2 summarizes the linear relationships between the four perceived task priority scores and the five potential prioritization criteria scores. Each P-value shows the linearity test result between the perceived task priority and the perceived task prioritization criterion. The relative weights methodology revealed that the perceived importance and the perceived salience scores explained most of the variance of the perceived task priority scores.

Table 2.
P-Values of five potential task prioritization decision criteria for four tasks.

Y: Priority of each task	**X1: Perceived Importance**	**X2: Perceived Urgency**	**X3: Perceived Performance Status**	**X4: Perceived Salience of tasks**	**X5: Perceived Workload**
Aviate task	P-Value<0.001 $R^2 = 0.70$	P-value<0.03 $R^2 = 0.55$	P-Value=0.88 $R^2 = 0.47$	P-Value<0.001 $R^2 = 0.60$	P-Value=0.03 $R^2 = 0.48$
Navigate task	P-Value<0.001 $R^2 = 0.58$	P-Value=0.07 $R^2 = 0.29$	P-Value<0.06 $R^2 = 0.34$	P-Value<0.001 $R^2 = 0.40$	P-Value=0.29 $R^2 = 0.26$
Communicate	P-Value<0.005 $R^2 = 0.66$	P-Value=0.02 $R^2 = 0.41$	P-Value=0.66 $R^2 = 0.34$	P-Value<0.001 $R^2 = 52$	P-Value=0.23 $R^2 = 0.17$
Manage Systems	P-Value=0.59 $R^2 = 0.58$	P-Value=0.002 $R^2 = 0.43$	P-Value=0.11 $R^2 = 0.40$	P-Value<0.001 $R^2 = 0.50$	P-Value=0.22 $R^2 = 0.33$

The second research question result was such that the perceived task priority score had a linear relationship with the actual task execution time and the task execution performance while the task awareness level was not improved by the higher perceived task priority score. For example, the higher perceived Navigate task priority (P=0.06) improved the directional deviation, and it was worsened as the number of concurrent tasks increased (P=0.13). On the other hand, the time period of inattention/change blindness was not improved by the perceived task priority score (P-Value=0.68), and it was worsened as the number of concurrent tasks increased (P-Value=0.02).

General Discussion

Following is our preliminary interpretation of those results. Scrutinizing the perceived urgency criteria data revealed two possible mechanisms of task awareness (i.e., inattention/change blindness). The first possible mechanism was a task-prioritization decision problem in sequential multitasking. Some pilots rated larger task buffer times (i.e., not urgent), suggesting that the high priority task was postponed; deterministically it took a longer time to notice the task related problem. Thus, task importance, task salience, and task urgency criteria should be addressed in cockpit design and pilot training for better task prioritization decisions. The second possible mechanism was the cognitive resource interference problem during concurrent multitasking. Subjects worked hard to conduct concurrent multitasking because they reported multiple tasks were simultaneously urgent. When pilots perceived multiple tasks to have high urgency and high priority, they would start concurrent multitasking. During concurrent multitasking, the task awareness level dropped in a stochastic way. When the instrument panel malfunction signals were neither expected nor salient, the mean time to notice the task signal was 137.8 seconds and no significant effect of number of concurrent tasks was observed (P-value=0.9). When the instrument panel

malfunction signals were expected but without salient stimuli, the mean time to notice the task signal was 122.5 seconds but it increased as the number of concurrent multitasks increased (P-value=0.01). When the instrument panel malfunction produced a salient signal but it was not expected, the mean time to notice the task signal was 24.8 seconds, but it increased as the number of concurrent multitask increased (slope P-value=0.03). When a problem was expected and had a salient signal, the mean time to notice the task signal was 23.9 seconds and it did not increase with more concurrent multitasks (slope P-value=0.7). This indicate that the bottom-up factors (signal salience, signal expectation), and the number of concurrent multitask should be taken into consideration in cockpit design and pilot training to increase awareness of tasks related information.

References

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. Cognitive science, 26(1), 39-83.

Chou, C. C., Madhavan, D., & Funk, K. (1996). Studies of cockpit task management errors. The International Journal of Aviation Psychology, 6(4), 307-320.

Colvin, K., Funk, K., & Braune, R. (2005). Task Prioritization Factors: Two Part-Task Simulator Studies. The International Journal of Aviation Psychology, 15(4), 321-338.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 32-64.

Funk, K. (1991). Cockpit task management: Preliminary definitions, normative theory, error taxonomy, and design recommendations. The International Journal of Aviation Psychology, 1(4), 271-285.

Funk, K., Suroteguh, C., Wilson, J., & Lyall, B. (1998, October). Flight deck automation and task management. In IEEE INTERNATIONAL CONFERENCE ON SYSTEMS MAN AND CYBERNETICS (Vol. 1, pp. 863-868). INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE).

Kushleyeva, Yelena, Dario D. Salvucci, and Frank J. Lee. "Deciding when to switch tasks in time-critical multitasking." Cognitive Systems Research 6.1 (2005): 41-49.

Lee, J. D., McGehee, D. V., Brown, T. L., & Reyes, M. L. (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. Human Factors: The Journal of the Human Factors and Ergonomics Society, 44(2), 314-334.

National Safety Council White Paper. (2010). Understanding the distracted brain: Why driving while using hands-free cell phones is risky behavior. (on-line publication)

NTSB. (1973). Aircraft accident report. Eastern Air Lines, Incorporated, L-1011, N310EA, Miami, Florida, December 29, 1972. Report No. NTSB-AAR-73-14. Washington, DC: National Transportation Safety Board

NTSB (1987). Aircraft accident report. Northwest Airlines, Incorporated, MD-82, Detroit, Michigan, August 16, 1987. Report No. Washington, DC: National Transportation Safety Board

Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption.

Salvucci, D. D., & Taatgen, N. A. (2010). The multitasking mind. Oxford University Press.

Shakeri, S., & Funk, K. (2007). A comparison of human and near-optimal task management behavior. Human Factors: The Journal of the Human Factors and Ergonomics Society, 49(3), 400-416.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception-London, 28(9), 1059-1074.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. Psychonomic Bulletin & Review, 5(4), 644-649.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional models of multitask pilot performance using advanced display technology. Human Factors: The Journal of the Human Factors and Ergonomics Society, 45(3), 360-380.

Wickens, C. D. & Hollands, J. G. (2013). Engineering psychology and human performance: Prentice Hall New Jersey.

Wilson, J. R. (1998). The effect of automation on the frequency of task prioritization errors on commercial aircraft decks: An ASRS incident report study.

USING AUGMENTED REALITY AND COMPUTER-GENERATED THREE-DIMENSIONAL MODELS TO IMPROVE TRAINING AND TECHNICAL TASKS IN AVIATION

Amadou Anne
Purdue University - Aviation Technology
Yu Wang
Purdue University - Aviation Technology
Timothy D. Ropp
Purdue University – Aviation Technology

Augmented Reality (AR) technology has considerably improved since its inception, and especially over the last few years, to the point of becoming a relatively reliable and potentially cost-effective tool in many fields. Significant advances have been made by software developers to improve the quality of AR tools and the hardware necessary to access these tools has become common. However, AR is mostly unknown or underused in the aviation industry, either in education and training or in professional environments. This paper aims to demonstrate the potential of AR for training and professional technical applications in aviation, especially when combined with 3D model visualization tools. The current state of AR is discussed, and a technical project showcasing AR is used as a demonstration.

Augmented Reality (AR) is defined as the technology that overlays computer-generated data on top of a real image or view through a digital piece of hardware (MacMillan Dictionary). The type of data displayed can vary from simple text lines to videos and even interactive 3D models. In order to access and visualize the AR content, the hardware required consists of a camera, a display and a data processing unit equipped with the proper software to achieve the task [5]. This equipment is currently integrated – albeit at different quality and performance levels – in personal computers, cell phones, tablets and certain head-mounted devices (HMD) among others. Augmented reality has been used – or at least experimented with – in many domains, such as marketing, entertainment, as well as in medical and technical fields (Hincapie, 2011).

Currently in the aviation industry, data related to performing training or field technical operations is delivered primarily in one of two ways: through paper-based instructions and manuals or in digital format. The digital data is either similar in content and format to the paper-based instructions or may contain enhanced visuals and some level of interactivity. Each of these methods of data delivery has advantages, but also severe drawbacks, which lead to technical documentation being considered the primary human factors challenge in aviation maintenance (FAA, 2012). In fact, 45 to 60% of safety incidents were procedure related or involved technical documentation. Thus, there is a clear need for the implementation of new ways to deliver information to technicians for training and in the field that will facilitate retention of knowledge and execution of the tasks while decreasing errors and thereby diminishing safety hazards for the worker and ultimately the users of the aircrafts.

This paper will demonstrate that applying augmented reality to the delivery of technical information can procure benefits unattainable with classic methods, while solving many of the issues associated with the latter.

The paper is organized as follows. An overview of the current data delivery methods is given, as well as a presentation of the current state of augmented reality technology. Section 4 details the developmental steps for AR-enhanced data delivery and demonstrates the potential use of augmented reality in aviation. The subsequent sections of the paper analyze the benefits and drawbacks of augmented reality as applied to technical tasks in aviation.

Background

Current Education, Manufacturing and Maintenance in Aviation

Over the past few decades, the volume of air travel has considerably increased and aircrafts have become increasingly modern, complex and inclusive of numerous and diverse auxiliary systems. Thus, the concepts and processes associated with manufacturing, maintenance and training – of aviation professionals – have also significantly increased in breadth and complexity. In turn, this has led to a large volume of reference material being necessary to perform tasks in the aforementioned fields. As highlighted in previous publications (Nee, 2012), manufacturing for instance has become much more complex and demanding and in virtually all cases requires exchange of information in real time between different units of production.

Traditional methods of information delivery and exchange – specifically standard printed or digital texts and manuals – are still very widely used in industry, and although they are cost-effective and well implanted throughout the industry, they have several disadvantages when considered in the modern aviation world. Indeed, if one considers the recurring need to update information (through Advisory Circulars, Airworthiness Directives or manufacturer publications), it is clear that traditional methods -consisting of end-user additions to publications for instance – incur undeniable inefficiencies and thus potential safety risks.

Also, workers in the field that use traditional methods of information delivery typically experience many issues when performing given tasks. Indeed, instructions are usually detached from the equipment that the technicians are performing work upon, which leads to the need to constantly switch focus between their instructions and work platform (Ong, 2008). This causes a – sometimes high – loss of time and productivity, as well as a higher potential for errors and injuries or damage.

In addition, as the authors describe in *An Introduction to Augmented Reality with Applications in Aeronautical Maintenance* (Hincapie, 2011), the information in traditional methods can be challenging to locate and extract. Indeed, workers and students – especially if inexperienced - can be led to frustration, poor performance and potentially costly mistakes when trying to find information in traditional texts and manuals.

The issues outlined here are even more critical when paired with the high volume of work and time pressure that the aviation industry imposes on its workers. Thus, with the advent of modern technologies, it is imperative to search for new methods that would solve some or all of these concerns and concurrently have relatively low costs of implementation (economic, human and technological). In this paper, augmented reality is evaluated as a potential new method of information delivery that could supplement the current infrastructure while solving the problems discussed above.

Augmented Reality

Definition. Augmented Reality (AR) consists of the display of information (text, images, videos, interactive content) that augments a scene that is actively captured by a camera (De Crescenzio, 2011). Thus, the three basic components needed to display AR content are a camera, a memory/processing unit and a display surface. Nowadays, these are found in a plethora of portable devices (smartphones, tablets, etc.) and even on wearable technology (glasses, head mounted displays -HMDs). However, it is important to note that the quality and quantity of the data overlaid is very dependent on that of the hardware and software contained in the unit used to display.

Technological Advance. Augmented Reality has been the subject of much research and development over the past decade. This has led to the technology being tested and used in many sectors. As Ong and Nee illustrate (Nee, 2012 & Ong, 2008), AR is being widely used in marketing and advertisement, and has been successfully demonstrated and used in medical, military, entertainment, maintenance and manufacturing fields.

In addition, hardware and software tools that can display AR content continuously gain in computing power and camera and display quality while maintaining or even reducing their size. Indeed, Hincapie and his co-authors observe that modern smartphones for instance boast state of the art sensors (compass, gyroscopes, GPS sensors) which could easily be used to provide higher quality AR content (Hincapie, 2011).

Issues with Augmented Reality. Traditionally, one of the biggest issues with Augmented Reality has been the size and weight of the hardware needed (Hincapie, 2011). Indeed, head-mounted displays for instance can be relatively uncomfortable to wear, especially for extended time periods. They typically lead to fatigue and limited range of movement, which in turn can cause errors and safety issues. However, there are multiple options to compute and display AR content, and considering technological advances, it is possible to find or design ideal platforms for the aviation industry.

Another concern related to AR is that computing power is still limited (especially for the display of complex 3D models) (Hincapie, 2011). Along the same line, considering that the AR data is usually stored on servers and accessed (usually wirelessly) through networks by the end-users, there is the issue that real-time data access, tracking and computation for correct display can be hampered by slow or faulty connections. These are valid concerns, which need to be considered and addressed by implementing the proper network infrastructure and choosing the adequate hardware for any given application.

Potential of Augmented Reality in Aviation Manufacturing, Maintenance and Education

Research, Experiments and Trials. There have been many research projects involving AR in manufacturing, maintenance and education. Many of these have focused on comparing augmented-reality methods of information delivery to currently common ones such as text, images and video.

Regarding educational applications, one such project conducted by Ong, Yuan and Nee, has demonstrated that AR is more effective than other forms of instructions, as it reduces errors and makes tasks easier (Ong, 2008). Another study by Macchiarella and Vicenzi (Macchiarella, 2004), which was designed to compare AR to video and text-based learning methods, compared short-term and long-term recollection of a topic in an aviation setting. The results obtained showed that AR produced significantly better long-term retention of information and thus was a better learning platform.

Industry applications for AR have also been tested by researchers. In *Augmented Reality for Aircraft Maintenance Training and Operations Support* for instance, the development process for an AR project is highlighted as well as the need to analyze the risks associated with each technical step in order to mitigate them using augmented reality. A case study by the authors validated this with subjects that properly followed the given procedures (in this case for an oil check) and did not commit errors or perform unneeded operations. Another experiment also tested the application of AR to industry practices, but focused on inspection procedures (Chung, 1999). Groups of participants measured the thickness of a part using either manual, computer or AR-aided methods. These two types of tasks (procedural and inspections) are the most common in the aviation industry, and therefore AR would be advantageous if it were integrated into professional task and information delivery.

Advantages of Augmented Reality. Beyond some of the efficiency and safety improvements discussed above, AR can provide multiple new ways to enhance information delivery in aviation. In fact, as Kesim and Ozarslan note, it allows for much better visualization and manipulation of objects and figures displayed on-screen (Kesim, 2012). In addition, information is displayed in the user's field of view, which gives them the ability to assimilate it better and concentrate more on the tasks to perform (Ong, 2008). The flexibility of AR also makes it applicable to several different types of processes (Hincapie, 2011), and AR-enhanced information is virtually always physically smaller (in weight and volume) than comparable information in print or other computer-based formats (Ong, 2008). This translates into more mobility, but also less time wasted accessing and retrieving information since the right information can be shown when and where it is needed (Ong, 2008). In practical tasks then, AR provides the benefits of added efficiency, safety and reduced waste of resources.

In training and education, augmented reality has been proven as a more effective learning tool than text or video-based methods (Macchiarella, 2004), and could for instance help reduce training time and costs in maintenance, which typically amount to about 2000 hours (Hincapie, 2011).

In design and manufacturing, AR could be used to simulate and improve products and processes before or during their implementation, and thus ensure their proper execution with minimal to no repetition or rework (Ong, 2008). In all of the previously mentioned domains, another proven advantage to AR is its collaborative potential: with the modern network technologies, design, approval, manufacturing and maintenance information and procedures can be shared and visualized by multiple entities in real-time in order to enhance information transfer (Kesim, 2004). One example would be the remote diagnosis of an aircraft system by experts who later guide a less-experienced maintenance worker through a complex repair that the latter would otherwise not be able to complete (Gautier, 2007). In this case, the aircraft could be dispatched again much faster than would have otherwise been possible through phone and text communication. In addition, there would be cost-saving implications since the experts would not have to travel to the aircraft's location to perform the required maintenance.

How Augmented Reality and 3D can be used in training and tech task delivery in aviation

Methodology for Development of AR Tasks

Creating Augmented Reality scenes for use in educational or professional environments is typically a four-step process:

- Planning of the AR scene
- Preparation of object or environment to be augmented
- Addition of content to be overlaid
- Save or upload of the created AR scene.

Planning of an Augmented Reality scene. As mentioned previously, there are many software, hardware and content options available to produce and access augmented reality content. Choosing the right combination of these elements is essential for the successful deployment of the scene. Many variables dictate this choice, among which:

- The profile of the user: when developing an AR scene, it is important to keep in mind the intended user's level of knowledge on the topic, familiarity with the technology, and even physical limitations among other attributes. For instance, an AR scene that is intended to present an overview of a turbine engine's main sections may be produced in different ways depending on its audience. If it were to be used by students who are familiar with AR technology and the basics of powerplant theory, the interface would be more detailed and content-rich than if it were destined to the general public, in which case there would be less technical content and more on-screen guidance on the use of the technology.
- The user's environment: this is a critical factor for the successful deployment of an AR scene because the user must be able to access and use the AR content with maximal ease and comfort, and minimal potential damage to the equipment or their environment. Some factors to consider are lighting conditions, distance from a network access point if applicable, amount of physical space around the user and noise concerns.
- The user's task or objective: it is important to visualize the user and their intended use for the technology. For instance, an AR project intended to provide instructions for a complex part removal may require the use of AR glasses and on-demand instructions in order to allow for full mobility of the user.
- There are many other factors to consider related to the development of Augmented Reality content, such as available hardware, network access or lack thereof and software limitations.

Preparation of Object or Environment. After planning the AR scene, the next step in the development of an AR project is to prepare the object or environment that will be augmented. The basis to achieve this is to recreate a model of the object or environment that the computing platform can recognize and visually augment. This can be done in different ways depending on the software/hardware platforms being used, but as Nee explains, it is more common to use software-based scanning and tracking methods (Nee, 2012). Below are the main methods used to save objects and environments for augmentation, as outlined by Nee (Nee, 2012):

- Marker-based technologies: using this technique, the software platform transforms certain features of the objects or environments into fixed reference points for augmentation. These features can be two-dimensional (QR codes for instance) or three-dimensional points (Metaio, n.d.).
- CAD model: it is also possible to use computer-generated three-dimensional models to activate augmented reality content. Indeed, this method is similar to using markers, except in the sense that the reference points are generated by the digital objects.
- Location-based technologies: This method involves using the location of the user to trigger augmented reality content.

Addition of content. After the object or environment has been prepared, the AR scene developer can add content that will be overlaid on the user's interface. This is typically done using a dedicated software platform which is compatible with that of the user.

There is a multitude of content types that can be overlaid using augmented reality. Some of the most useful in an aviation context are:

- Text
- Images
- Videos
- 3D models
- Links

In many of the software platforms currently in use, it is also possible to animate the overlaid content or allow the user to interact with it in order to enhance the usability and efficiency of the scene.

Save or upload of the scene. Once the AR scene is ready for deployment to the user, the last step is to save or upload it, depending on the retrieval method of the user. In fact, AR projects can be transferred to the user's platform either directly (with physical device connections or through a network), or through a third-party service (augmented reality application for instance).

Demonstration of an AR application in aviation training and task instruction delivery

The Hangar of the Future Research Laboratory in the Aviation Technology department at Purdue University has been conducting research on using Augmented Reality applications to enhance training and work instructions for a few years (Hangar of the future). Researchers in this laboratory have developed dozens of AR-enhanced projects and demonstrations that are applicable to training and industry tasks. One of these, which uses a Pratt & Whitney 4000-series turbine engine as a platform will be detailed below.

The purpose of this AR project was to demonstrate the capabilities and versatility of the technology in an education environment, but also in a professional manufacturing or maintenance setting. Thus, the user interface was designed to include informational content about the systems, as well as step-by-step instructions to perform certain tasks. This project was created using the Metaio suite of AR software. This includes the scanning application *Toolbox*, the AR content creation platform *Metaio Creator* and *Junaio*, an application which allows the users to access the created content (Metaio). This Hangar of the Future project was primarily intended to be accessed on tablets and smartphones. Those devices were chosen because of their relatively low cost, high computing power, portability and popularity.

The project essentially consists of two parts: one that can be used for familiarization with turbine engine components and functioning, and a second that provides step-by-step instructions to perform certain hands-on laboratory projects. In both of these sections, several different data formats were used to convey information clearly and efficiently. Text boxes and images were displayed for descriptions and illustrations, as well as for users to select in order to navigate to displays that contain additional information. For instance, text and images were used to describe the different sections of the turbine engine and illustrate the air flow through them. In addition, videos were used in multiple cases to provide supplemental audio-visual information. This was the case to demonstrate proper cable routing and attachment for example. Finally, computer-generated three-dimensional models (3D) were used to enhance the visualization of certain parts, as they provided the possibility for the users to manipulate the object on-screen. This capability was used to provide visual details of the full authority digital engine control (FADEC) unit of the engine.

This augmented reality project constitutes phase I of this research, and has demonstrated the feasibility and applicability of augmented reality as a tool for education and delivery of task instructions in an aviation setting. The next stage will be to design and execute an experiment that will test and measure students' perception of this technology.

Conclusion

This research paper has shown that augmented reality has the potential to positively impact information delivery in aviation in many ways, both for training and professional purposes. However, developing successful augmented reality project requires a careful and methodical approach, which was followed by this research team and the Hangar of the Future laboratory at Purdue University to create an application for students in the Aviation Technology department. This project demonstrates that incorporating augmented reality in educational and professional fields is a realistic and feasible possibility, and its impact – as well as the students' perception – will be evaluated in the second phase of this research.

References

Augmented Reality (n.d.). In MacMillan Dictionary Online. Retrieved from http://www.macmillandictionary.com/dictionary/british/augmented-reality

Chung, K. H., Shewchuk, J. P., & Williges, R. C., (1999). An Application of Augmented Reality to Thickness Inspection. *Human Factors and Ergonomics in Manufacturing, 9* (4), 331-342.

De Crescenzio, F., Fantini, M., Persiani, F., Di Stefano, L., Azzari, P., & Salti, S. (2011). Augmented Reality for Aircraft Maintenance Training and Operations Support. *Computer Graphics and Applications, 31* (1), 96-101.

Federal Aviation Administration. (2012). *Technical Documentation Challenges in Aviation Maintenance – A Proceedings Report* (FAA Publication No. AM-12/16). Washington, DC: Office of Aerospace Medicine.

Furmanski, C., Azuma, R., & Daily, M. (2002). Augmented-reality visualizations guided by cognition: Perceptual heuristics for combining visible and obscured information. Proceedings of the *International Symposium on Mixed and Augmented Reality (ISMAR'02)*. Retrieved from http://monet.cs.columbia.edu/courses/mobwear/resources/furmanski-ismar02.pdf

Gautier, G., Fernando, L., Piddington, C., Hinrichs, E., Buchholz, H., Cros, P.-H., … Vincent, D. (2008). Collaborative Workspace for Aircraft Maintenance. Proceedings from *3rd International Conference on Virtual and Rapid Manufacturing: Advanced Research in Virtual and Rapid Prototyping 2007*. Leiria, Portugal.

Hangar of the Future (n.d.). *Hangar of the Future Research Laboratory.* Retrieved from https://tech.purdue.edu/facilities/hangar-of-future

Hincapie, M., Caponio, A., Rios, H., Mendivil, E.G. (2011). An Introduction to Augmented Reality with Applications in Aeronautical Maintenance. Proceedings from *13th International Conference on Transparent Optical Networks (ICTON)*. IEEE: Stockholm.

Kesim, M., Ozarsalan, Y. (2012). Augmented Reality in Education: Current Technologies and the Potential for Education. *Procedia – Social and Behavioral Sciences, 47*, 297-302. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877042812023907

Macchiarella, N. D., Vincenzi, D. A. (2004). Augmented Reality in a Learning Paradigm for Flight and Aerospace Maintenance Training. Proceedings from *The 23rd Digital Avionics Systems Conference (DASC 04)*. IEEE.

Metaio (n.d.). Augmented Reality for Service and Maintenance. Retieved from http://www.metaio.com/fileadmin/upload/documents/pdf/case-study/A4-service_maintenance-2013.pdf

Nee, A. Y. C., Ong, S. K., Chryssolouris, G., & Mourtzis, D. (2012). Augmented Reality Applications in Design and Manufacturing. *CIRP Annals – Manufacturing Technology, 61*.

Ong, S.K., Yuan, M.L., & Nee, A.Y.C. (2008). Augmented Reality Applications in Manufacturing: a Survey. *International Journal of Production Research, 46:10*, 2707-2742, DOI: 10.1080/00207540601064773

Ong, S.K., Zhang, J., Shen, Y., & Nee, A.Y.C. (2011). Augmented Reality in Product Development and Manufacturing. *Handbook of Augmented Reality*, 651-669, DOI: 10.1007/978-1-4614-0064-6_30

Tang, A., Owen, C., Biocca, F., & Mou, W. (2002). Experimental Evaluation of Augmented Reality in Object Task Assembly. Proceedings of the *International Symposium on Mixed and Augmented Reality (ISMAR'02)*. Retrieved from http://www.computer.org/csdl/proceedings/ismar/2002/1781/00/17810265.pdf

Tschirner, P., Hillers, B., & Graser, A. (2002). A Concept for the Application of Augmented Reality in Manual Gas Metal Arc Welding. Proceedings of the *International Symposium on Mixed and Augmented Reality (ISMAR'02)*. Retrieved from http://www.computer.org/csdl/proceedings/ismar/2002/1781/00/17810257.pdf

IMPACT OF TASK LOAD AND GAZE ON SITUATION AWARENESS IN UNMANNED AERIAL VEHICLE CONTROL

Joseph T. Coyne
Ciara M. Sibley
Naval Research Laboratory
Washington, DC

Increasing levels of automation and rising costs of manpower are pushing the DoD towards a supervisory control paradigm for future unmanned aerial vehicle (UAV) missions. Using the Supervisory Control Operations User Testbed, a group of 20 participants completed two twenty minute supervisory control missions where eye tracking and performance data were collected. Each mission had 3 levels of task load; which were manipulated by varying the frequency of events to which the user responded. During each level, the simulation paused and a situation awareness (SA) probe appeared with all UAVs and targets randomly placed on the map. Participants were tasked to reconstruct the map. Results showed higher load was associated with a significant decrease in SA. Additionally, participants spent significantly less time looking at the map when the task load was high. These results suggest that eye gaze may be a useful predictor of SA within a supervisory control task.

Unmanned Aerial Vehicles (UAVs) accounted for only 5% of the Department of Defense's (DoD) aircraft inventory in 2005, however by 2012 that number has surged to 41% (Gertler, 2012). The rapid rise in vehicles and the simultaneous reduction in force within the DoD has increased the interest in shifting the UAV operations paradigm away from direct control of specific vehicle functions (e.g., piloting and payload) and towards a supervisory control model where a single operator monitors multiple vehicles (DoD, 2013). Increases in UAV automation, in addition to open system architectures will help drive this change. UAV control has already begun to shift from control via "stick and rudder" to waypoint navigation. The majority of an operator's tasking in this future environment will likely be to assign vehicles to different targets and objectives, monitor the progress of those vehicles, and update plans given new opportunities and changing information.

Problems associated with increased automation and decreased situation awareness (SA) have been discussed in a number of studies (e.g., Calhoun et al., Endsley & Kaber, 1999). Reduced SA primarily results from a combination of complacency and lack of interaction with the system, and these SA lapses can lead to an inability to detect problems when automation fails, as well as increased time to recover after an error. For example, Calhoun et al. (2011) investigated varying levels of automation and reliability within a UAV supervisory control environment and found that after several sessions in which automation performed perfectly, every participant missed a route error introduced by the automation, despite being told that the automation could make mistakes. While the importance of SA within UAV supervisory control has often been suggested to be a potential issue, the assessment of SA has been limited. Calhoun et al (2001) inferred poor SA based upon user's lack of a correct response, this represents an implicit measure of SA and is only one of several types of SA measures (Sarter & Woods, 1995). The focus of this paper is to investigate two other methods of assessing SA within a UAV supervisory control environment, specifically through SA probes and physiological measurement.

SA probes such as SAGAT (Endsley, 1998) are one of the most common methods of SA assessment. In these probes the environment is paused and hidden and the individual is asked a specific series of questions regarding its current state. While probe measures provide a direct means of measuring SA they can be disruptive, only provide data at discrete points in time, and can only be used within controlled studies. The ability to have continuous measures of SA, which also provide insight into the process behind acquiring and maintain SA are highly desirable. The use of physiological data,

specifically eye tracking based metrics, are beginning to gain traction as a continuous noninvasive SA measurement technique (e.g., Moore & Gugerty, 2010; Van de Merwe et al., 2012). While Van de Merwe couples gaze information with an implicit measures of SA (i.e., detection of an aircraft failure), Moore and Gugerty's research was one of the only studies to include both direct measures of SA (using SAGAT) and eye tracking. Mooore and Gugerty found that the percentage of time looking within areas of interest (aircraft within an ATC display) accounted for a significant portion of the variance within the SA probe, suggesting that eye tracking data is an effective means of assessing SA implicitly.

The purpose of the present study is to further investigate the utility of eye tracking metrics within a supervisory control environment. Specifically the authors hope to replicate the findings of previous researchers and demonstrate a relationship between eye tracking data and performance within a SA probe.

Method

Supervisory Control Operations User Testbed

The Supervisory Control Operations User Testbed (SCOUT) was developed by the Naval Research Laboratory to replicate the tasks that a UAV mission commander and air vehicle pilot will perform in future operations with increased automation. SCOUT was designed as a two screen game environment that enables a single operator to control 3 heterogeneous UAVs. The vehicles differ in their speed and sensor range (which influences the time to complete a target search). The tasks within SCOUT include navigation and route planning; airspace management (requesting access to controlled airspace); communication (responding to requests for information); and adjusting air vehicle parameters (i.e., altitude and speed) and target parameters (i.e., location and search radius) as new information and commands are issued. The route planning task was the participant's main priority and main source of points within the game. The task required the participant to assign multiple targets of varying values, deadlines, and uncertainty (potential time required to locate the target) to the three UAVs they were supervising. Each vehicle would automatically search for their assigned target once they arrived within the target area. The payload task was entirely automated where the target would be found and the corresponding points awarded once the vehicle was within sensor range of the target. In addition, the participant received points for responding to requests for information and vehicle commands. Points were reduced when the participant's vehicles entered into restricted airspace without requesting access. Figure 1 depicts SCOUT's main mission management screen. During the experiment, the map was locked in position and represented an area of approximately 65 x 45 kilometers.

Equipment

A SmartEye Pro 6.1 five camera system was used to capture eye tracking data from participants at a frequency of 60Hz. The data from the SmartEye system was sent via network packets to the computer running SCOUT and was integrated with the participants' behavioral and simulation data.

Experimental Design

Twenty civilian employees and summer interns at the Naval Research Laboratory volunteered for participation in the experiment. Participants completed a 30 minute SCOUT training session consisting of a series of videos and sample tasks. Upon completing the training, there were two experimental sessions. Each experimental session began with a planning phase in which participants had up to ten minutes to select their initial vehicle and target assignments before the vehicles began moving. After the planning phase, the participant completed an 18 minute experimental block which consisted of a six minute easy, medium, and hard segment, in which events were presented approximately every 75,45, and 15 seconds respectively. A situation awareness probe (depicted in Figure 2) was presented within each 6 minute block. During the SA probe the SCOUT control displays would disappear and the probe would

appear with the vehicles and targets randomly placed on the map. The participant had to two minutes to move the vehicles and targets to their estimated position on the map and then indicate which target each vehicle was currently pursuing and its value . Once participants completed the probe, the simulation would return and vehicles would only begin moving again once a "resume mission" button was pressed.

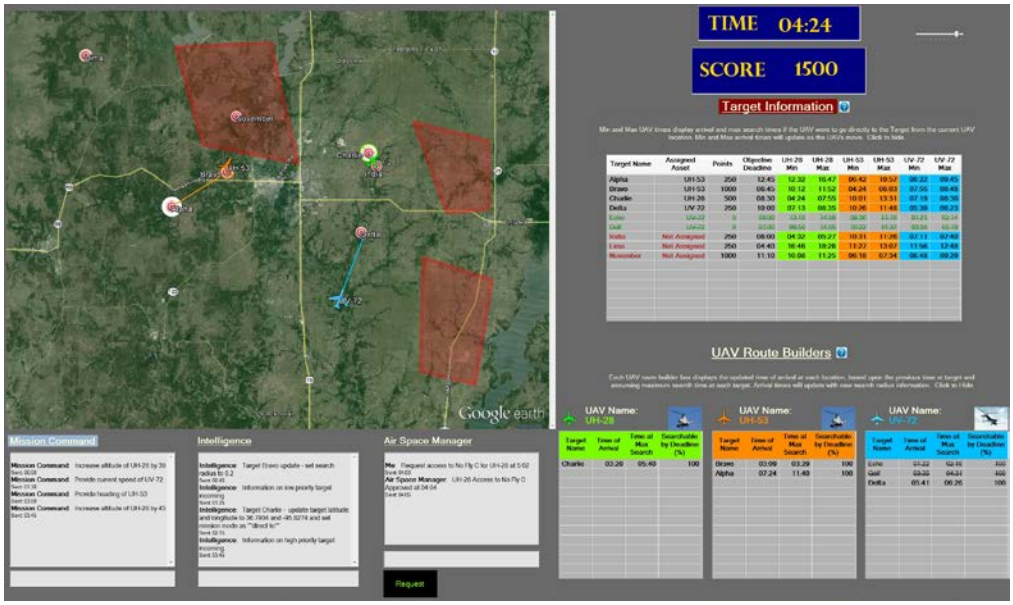


Figure 1. SCOUT's mission management screen where operators assigned targets to their different vehicles.

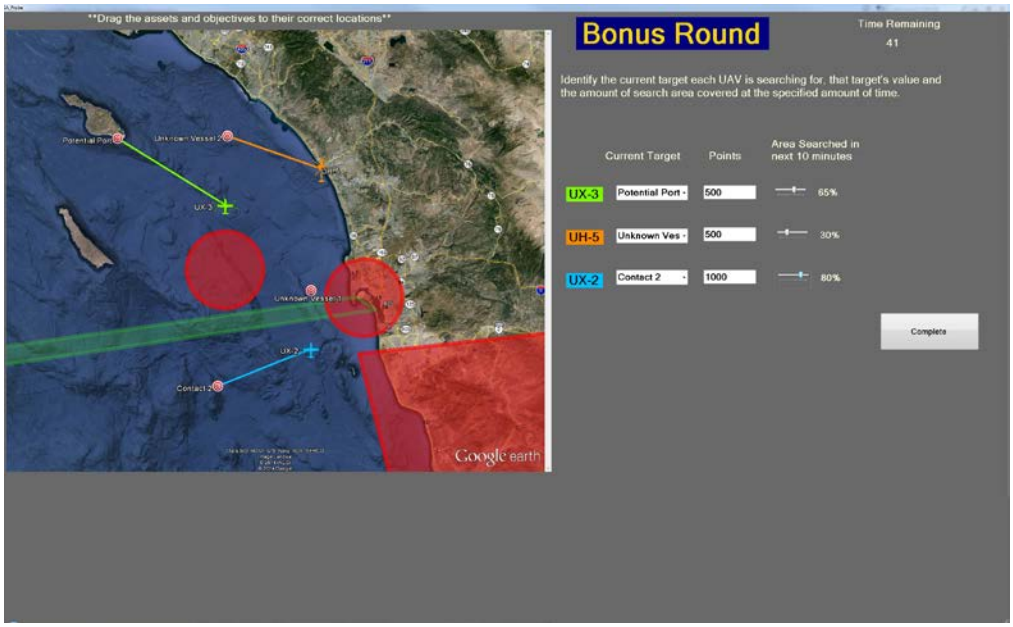


Figure 2. SCOUT's situation awareness probe.

Results

Situation Awareness Probe Results

The primary metric from the SA probe was the distance between the participant's placement of the vehicles and targets on the map and the actual position of those objects in the simulation, immediately preceding the probe. The maximum possible error for each object was limited to 35 km. A two-way (session x difficulty) repeated measures ANOVA was run on the SA probe error data. There was a significant main effect of difficulty $F(2,38) = 3.382$, $p = .044$ (see Figure 3). A Tukey HSD post hoc analysis revealed that the SA probe error was significantly smaller in the easy task load compared to the difficult task load. There was no effect of session ($F(1,19) = 2.109$, $p = .163$) or interaction of session and difficulty on SA probe error ($F(2,38) = .834$, $p = .442$). An additional metric from the SA probe was the ability to correctly identify which target each vehicle was pursuing. The results of this analysis mirrored those of the error distance with only a main effect of difficulty being present $F(2,38) = 5.729$, $p = .007$. Performance within the east condition (61%) was significantly better than the hard condition (36%).

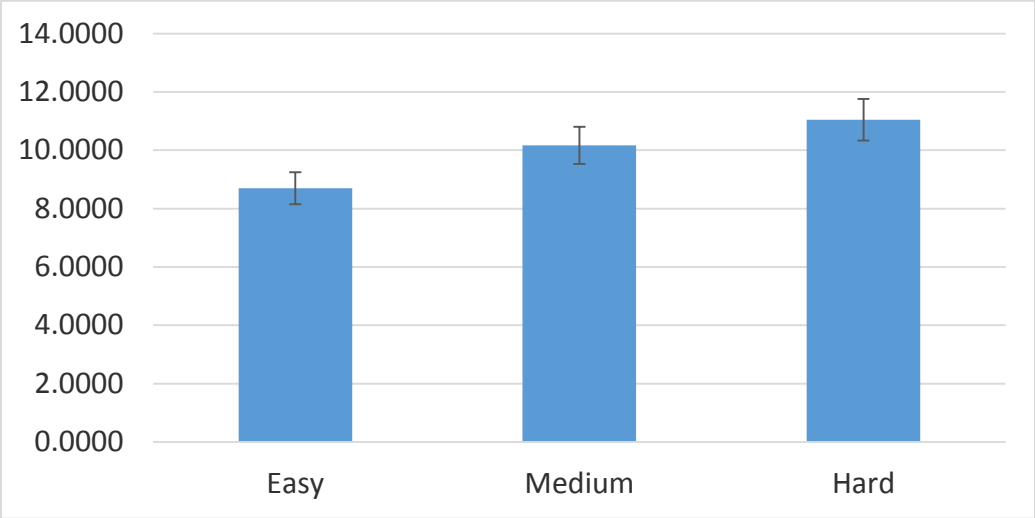


Figure 3. SA Probe Error distance across the three levels of task difficulty. Error bars represent the standard error of the mean.

Gaze Results

The eye tracking analysis focused on the data collected one minute prior to each SA probe. Specifically, the principal eye tracking metric was percentage of time spent looking at the map prior to the probe. A two-way (session x difficulty) repeated measures ANOVA was run on the percentage of map dwell time data. There was a significant main effect of difficulty $F(2,38) = 5.174$, $p = .010$ (see Figure 4). A Tukey HSD post hoc analysis revealed that participants spent significantly less time looking at the map in the hard condition compared to both the easy and medium conditions. There was no effect of session ($F(1,19) = 1.693$, $p = .209$) or interaction of session and difficulty ($F(2,38) = .755$, $p = .478$) on percentage of map dwell time.

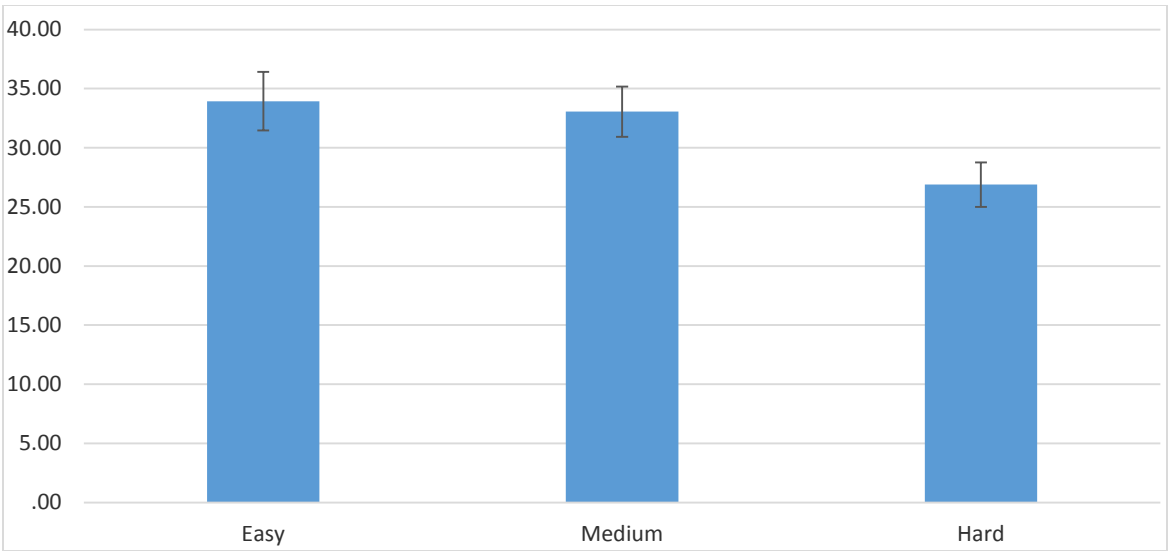


Figure 4. Percent of time spent looking at the map 1 minute prior to probe. Error bars represent the standard error of the mean.

Variable Correlations

Correlations were run to understand the relationships between the different variables of interest. In addition to Difficulty level, SA Error and Time in Map, the average amount of eye movement over the 1 minute time period in pixels was investigated as an additional measure (see Table 1).

Table 1.
Correlation matrix for SA Error, eye metrics and difficulty

	SA Error	Time in Map	Eye Movement	Difficulty
SA Error	1			
Time in Map	-.229*	1		
Eye Movement	.112	-.212*	1	
Difficulty	.215*	-.209*	.110	1

Note. * Indicates significance at .05 level

Discussion

The results of this study indicate that both performance on an SA probe and gaze were significantly impacted by task demands. As task demands increased, operators spent significantly less time looking at the map display (which aided in their ability to avoid restricted airspace and identify new targets of opportunity) and showed significant reductions in their SA. Further, these results support those of previous researchers (Moore & Gugerty, 2010) and provide evidence that eye tracking measures can be used as a potential supplement to direct measures of SA. While direct measures of SA offer high face validity, they are disruptive and difficult to implement in many environments. As such, identification and use of a non-obtrusive continuous measure of SA is of great value. Such a measure could help drive adaptive automation, such as tailored alerting (Ratwani, Mccurry & Trafton, 2010) or be useful in assessing operator performance with new displays or automation. Although the correlations between SA probe error and Time in Map were significant, they were small, suggesting that eye tracking alone cannot account for an individual's SA probe values and that such measures are still valuable in understanding SA.

It is important to note that within the present study, task demands were driven by chat communication as well as the inclusion of new targets of opportunities. These tasks often took the operator's attention away from the map display. The eye movement correlation data shows that as time spent looking at the map increased the amount of overall movement tended to decrease. It is unclear how queries or tasking that pushed attention to the map display would have impacted either the eye tracking metrics or the performance on the SA probe. Additionally the eye tracking analysis described here treated the entire map as a single area of interest. Future research may look at time on each object on the map and SA probe performance for that specific object. The version of SCOUT used in this experiment made such analysis challenging, however improvements to the map functionality now allow for easy conversion of pixel location to both map objects and latitude and longitude. This new functionality will aid in a more precise measure of eye tracking and SA.

Requesting that participants reposition the icons on the map of the same dimension and size represents the most basic SA, i.e. level 1 perception of elements (Endsley, 1995). Future experiments within SCOUT will look at higher levels of SA. Potential changes such as repositioning the area within the map so that user can no longer rely on screen position but must interpret location may can help investigate level 2 SA (Comprehension), additionally requesting individuals draw the position of their aircraft at a future time can help assess level 3 SA (Projection). The use of position based SA probes provides an advantage over traditional SA queries within SAGAT (e.g., where was the aircraft landing) in that the measure is continuous.

Overall the present study helped confirm the utility of using eye tracking as a supplemental measure of operator state and awareness.

References

Calhoun, G.L., Draper, M.H., and Ruff, H.A. (2009). "Effect of level of automation on unmanned aerial vehicle routing task", in: *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (pp. 197-201) San Antonio, Texas.

Department of Defense (2013). "Unmanned Systems Integrated Roadmap FY2013-2038". (Washington, DC: Department of Defense).

Endsley, M.R. (1988). "Situation awareness global assessment technique (SAGAT)", in: *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference* (Dayton, OH: *IEEE*), 789-795.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors,* 37(1)**,** 32-64.

Endsley, M.R., and Kaber, D.B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42**,** 462-492.

Gertler, J. (2012). "U.S. Unmanned Aerial Systems". (Washington, DC: Congressional Research Service).

Moore, K. and Gugerty, L. (2010) Development of a novel measure of situation awareness: The case for eye movement analysis, in *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting* (pp. 1650-1654) San Francisco, CA.

Ratwani, R.M., Mccurry, J.M., and Trafton, J.G. (2010). "Single Operator, Multiple Robots: An Eye Movement Based Theoretic Model of Operator Situation Awareness", in: *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction* (pp. 235-242) Osaka, Japan.

Sarter, N. B. and Woods, D. D. (1995) How in the world did we get into that mode? Mode error and awareness in supervisory control. *Human Factors,* 37(1), 5-19.

Van De Merwe, K., Van Dijk, H., and Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology* 22**,** 78-95.

FUSION: A FRAMEWORK FOR HUMAN INTERACTION WITH FLEXIBLE-ADAPTIVE AUTOMATION ACROSS MULTIPLE UNMANNED SYSTEMS

Allen J. Rowe, Sarah E. Spriggs
Air Force Research Laboratory
Dayton, Ohio

Daylond J. Hooper
Infoscitex
Dayton, Ohio

Future unmanned systems operators and heterogeneous unmanned systems must be able to work as agile synchronous teams to complete tactical reconnaissance, surveillance, and target acquisition related missions requiring the use of automation to assist the human operator. Interface research in this area is critical to the success of human-automation teaming, thus requiring a research test bed that brings together humans, autonomy, and systems. Fusion is a framework that enables natural human interaction with flexible and adaptive automation via the use of intelligent agents reasoning among disparate domain knowledge sources, machine learning providing monitoring services and intelligent aids to the operator, cooperative planners and advanced simulation through an instrumented, goal oriented operator interface that empowers scientific experimentation and technology advancement across multiple systems. There are four primary research threads that the Fusion framework is addressing to accomplish these goals: cloud-based simulation architecture; software extensibility; interface instrumentation; and, human-autonomy dialog through retrospection.

Autonomous systems are becoming an increasingly critical aspect of military operations. To enable these technologies, ever more capable autonomy frameworks are required, providing task-based management of a multitude of heterogeneous unmanned systems (UxSs) (USAF Chief Scientist, 2013). These technologies are still in their infancy, requiring extensive research to target increased understanding of how human operators can effectively coordinate and communicate with autonomous systems, among other issues in this domain. Robust autonomy-based frameworks enable evaluation of cooperation and coordination among widely disparate platforms such as remotely piloted unmanned systems (RPAs), autonomous unmanned systems, and adversarial components for ground, air, cyber, space, maritime, and submarine entities. Tying these interactions into an immersive user interface will improve evaluation of user behaviors and confidence in a low-risk environment. However, a unique challenge exists in unification of user interactions, autonomous platforms, and intelligent aids. A common drive is to push towards more autonomy, diminishing the user's involvement. Users can provide useful information to autonomous systems, and autonomy can be used to augment user capabilities, so an alternative is to develop and support symbiosis between users and systems. This symbiosis can be realized via a robust framework that provides user-tunable accessibility into this autonomy. This enables evaluation of user comfort, trust, and confidence with autonomous components. The associated ability to tune autonomy also drives future requirements for user interface design and accessibility.

Fusion Overview

Fusion Framework

Fusion is a framework that enables natural human interaction with flexible and adaptive automation. It employs multiple components: intelligent agents, reasoning among disparate domain knowledge sources (Douglass, 2013); machine learning, providing monitoring services and intelligent aids to the operator (Vernacsics, 2013); cooperative planners (Kingston, 2009); and advanced simulation via an instrumented, goal-oriented operator interface (Miller, 2012). These empower scientific experimentation and technology advancement across multiple systems (see Figure 1). The Fusion Framework consists of a layered architecture supporting disparate research projects with a development kit to explore a variety

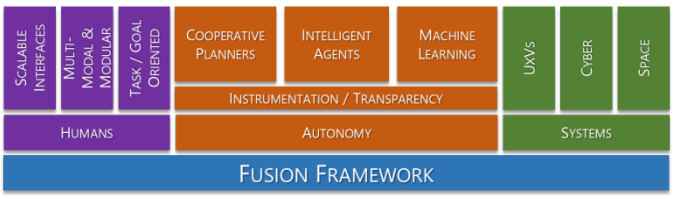


Figure 1. Fusion High Level Framework.

of research goals. The framework consists of four fundamental layers: (a) the core framework layer, (b) the extensibility and API layer, (c) the module / messaging layer, and (d) the application layer (see Figure 2). The core framework layer provides foundational software classes and an application programming interface (API). This layer enables functionality for module lifecycle, user profile, and display layout management. Additional features of this layer include system level notifications, multi-modal interactions and feedback, workspace management, asset management (vehicles, tracks, sensors, named areas of interest, etc.), global information services (GIS) data and earth mapping capability, and user interface elements. All software modules have a public framework API to support interface extensibility. This is accomplished in the extensibility and API framework layer. The module and messaging layer contains code written for single and specific purposes. This is the layer that contains user interfaces, utility classes, and messaging protocol support for communication to external software components. Finally, the application layer contains code related to executable applications such as a test-bed, utility application, or test operator console. All code is written utilizing agile software development principles, namely the SOLID principles (Single Responsibility, Open/Closed, Liskov Substitution, Interface Segregation, and Dependency Inversion) (Martin, 2012).

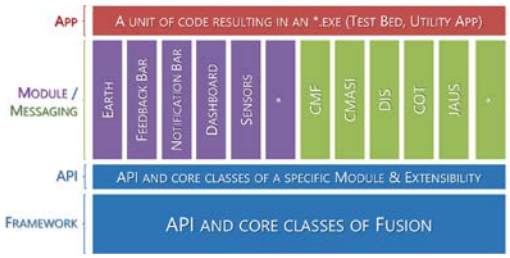


Figure 2. Fusion Layered Architecture.

The Fusion visual framework is broken into six key concepts: (a) Login, (b) Layout, (c) Notification, (d) Feedback, (e) Canvas, and (f) Tiles (see Figure 3).

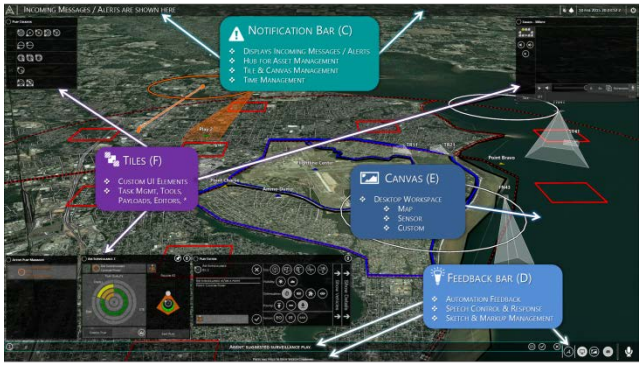


Figure 3. Fusion Visual Framework Components.

Virtual Distributed Lab

The notion of a virtual distributed laboratory (VDL) connecting various DoD and contractor sites throughout the CONUS was paramount to foster a more cohesive and distributed development and research environment. Fusion has adopted a DoD open source model, enabling joint development across a variety of projects and collaborators, all contributing to a single source repository. The core development team is located at Wright-Patterson AFB, and there are currently several offsite laboratory development teams (see Figure 4). Fusion is hosted on a secure web server and program access can be requested at https://www.vdl.afrl.af.mil/ (please contact the authors for further instructions). The software is developed on a standard Windows 8.1 PC platform in Microsoft Visual Studio and several third party developer libraries (See Figure 4). All distributed laboratory sites have similar hardware configurations and the software developers use a common set of software development and configuration management tools.

Figure 4. Fusion Virtual Laboratory Concept.

The Fusion software development team leverages SCRUM, an agile software development process. The Fusion source code repository is hosted on VDL and a strict configuration management process is followed. Once a week, offsite developers submit their changes, and the core Fusion team integrates those changes and posts a new version of Fusion on VDL for the offsite developers and researchers. In the coming year, the team will transition to a fully on-line software development cycle utilizing Git (a software configuration repository structure) and the secure Defense Research & Engineering Network (DREN). This process allows all offsite labs to keep up to date with the core Fusion team as well as keep their software well maintained. The concept of a virtual distributed laboratory has been successful due to Fusion providing a robust and flexible software architecture.

Fusion Keystone Program Application

The Fusion program formed to support two ASD R&E Autonomy Pilot Research Initiative (ARPI) projects. The purpose of this initiative was to foster research and push the envelope for autonomy-based researcg. One such project, Realizing Autonomy via Intelligent Adaptive Hybrid Control, is developing an "Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies" (IMPACT). This is a three year effort (FY14-FY16) with a focus on maximizing human-autonomy team agility. The first year of this effort focused on designing and implementing the user interface for higher level, goal-oriented plays (analogous to the sports ontology), which included asset management and integrating the various autonomous components. This "play" centric concept allows operators to focus on higher level goals for the vehicles, leveraging autonomous aids in accomplishing those goals, thus reducing the need for the user to direct and control vehicles manually.

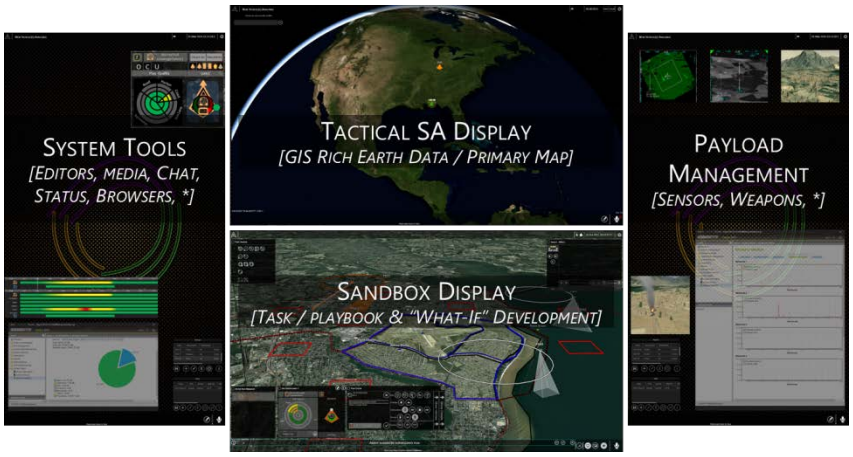


Figure 5. IMPACT instantiation of Fusion.

The IMPACT instantiation of Fusion, as shown in Figure 5, employs a four screen layout: system tools, real time tactical situation awareness (SA) display, sandbox tactical SA display, and sensor management. The operator uses the sandbox display to perform all of their play calling and chat monitoring tasks. The other screens display tools to enhance the operator's SA.

All of the goals of the Fusion framework were critical to the success of the first year of IMPACT. Fusion, autonomous agents, external simulations, a cooperative control planner (CCA), and machine learning algorithms were combined to form a comprehensive, richly interactive environment. This was enabled by Fusion's flexible software architecture through a robust simulation environment, software extensibility, and interface instrumentation.

Flexible Software Architecture

There are four primary research threads that Fusion is addressing to accomplish the goals of developing a framework for human interaction with flexible-adaptive automation across multiple UxSs: (1) developing a software system that can generalize disparate and similar messaging protocols to be protocol-agnostic while allowing a many-to-many relationship between networked systems for the generation, distribution and consumption of network messages; (2) developing a software framework where every public element, regardless of its role as a model or user-interface element, is customizable, extendable and override-able by any other software developer in the system; (3) developing a software system that is fully instrumented to gather real-time user/machine interactions and system details for use in experimentation, software agents, and machine learning; and finally, (4) developing a software system that records the state of each of its components at a rate near 30hz and makes it user-accessible to enable discrete and continuous retrospection of the system in real-time.

Cloud-based Simulation Architecture

The development team has established an API for external software components to communicate and interact with Fusion. To date, vehicle simulations, intelligent task allocation agents, vehicle planners, speech interpreters, chat systems, sensor visualization, operator assistance components, map layer data, and monitoring components have been incorporated into the Fusion network API. These networked components employ various connection modalities (e.g., UDP, TCP/IP, ZeroMQ) and communicate using various messaging protocols (LMCP, JSON, DIS, and custom protocols). In some form, all the components are linked together in their communications modalities by use of a hub. Where appropriate, the connections and protocols are also realized into appropriate interface components in Fusion, and are intended to aid in creating a more immersive and interactive system for human-autonomy teaming.

The goal of this network API is to make the incorporation of external software as transparent and natural as possible while leveraging data efficiently. All of the instrumentation data (discussed in detail later in this section) is distributed to the communications hub, and any component that wishes to consume the data can do so with a simple subscription. Likewise, communication messages from the other components are delivered to the same hub, and Fusion (or any other component) can subscribe and receive those messages. Each of the networked components may also communicate with another networked component using this same network structure. The Fusion team has worked closely with developers of other software components in order to ensure a seamless integration. The publish/subscribe architecture present on the communications hub makes for a natural assembly: all the associated data published by any software entity is available to any other entity that needs to leverage it, thus enabling great flexibility in the potential interactions between the entities, including Fusion and its operator(s). It also establishes the framework that will be needed for our near-future extension of Fusion to a MOMU (multiple operator/multiple unmanned system) interface.

With Fusion as an operator's interface, the communication with the components are more transparent and natural. For example, in the IMPACT work that Fusion has supported, a user can define high-level goals that Fusion dispatches to an intelligent agent. The agent allocates the vehicle or vehicles to be used for the achievement of the goals and submits the realized requirement to the vehicle planner. The planner reports the plan back to Fusion, which is then shown as a candidate solution. When accepted, the plan is delivered to the vehicle simulation for execution, while another component monitors the plan execution to alert the operator in case of issues. Each of these pieces communicate using a different set of message protocols. To the operator, it appears as if all these components are part of Fusion: the operator defines a goal and approves a plan, and it appears to the operator that the deliberative activity occurs exclusively within Fusion. In this way, Fusion enhanced the operator's teaming with the autonomous system components while ensuring that the communication and feedback are transparent to the operator. Since the operator is able to define high-level goals for the IMPACT project, the system enables a realization of an enhancement in the dialog between the operator and the various software entities that provide access and control of autonomous vehicles. This enhancement was exercised in the first IMPACT evaluation, which had positive feedback from all the subjects. The flexibility and extensibility of the communications framework has provided a baseline from which the development team will further extend and enhance the human-machine teaming, creating a more immersive and flexible interactive environment.

Software Extensibility

Fusion is being used in several different projects, all of which share the goal of improving operator interactions with highly autonomous systems (with future extensions to MOMU), but have vastly different human machine interface (HMI) designs and algorithms. Due to this, Fusion was built with the goal of extensibility.

Fusion's infrastructure allows developers to override aspects of the HMI easily. Fusion adopted a layered architecture in which the framework contains the building blocks for HMI tools and services. Developers add new HMI tools and services by overriding those building blocks and developing new modules. Thus, developers can override or extend aspects of Fusion without altering the original or previous extensions. Modules can be either universal or project-specific. Through this, the user can choose which modules are loaded, and therefore affect how the Fusion user interface looks and reacts.

One example of the extensibility currently realized in Fusion is the vehicle symbol. In test beds that allow operators to control or supervise unmanned systems, vehicle symbols are important and appear in many different places in the user interface. In Fusion, vehicle symbols appear on the map, in various notifications, on the vehicle status tool, in tasking tools, in many project specific tools, and other places. Most projects have their own vehicle symbol design and it require a lot of work to replace that symbol everywhere for every project. Fusion was built with this in mind, therefore, there is a default simplistic vehicle symbol included in the framework. Every project has the ability to design and implement their own vehicle symbol. With a single line of code in the project-specific vehicle symbol code, all vehicle symbols in the entire Fusion test bed can be replaced. An additional feature is that developers do not even need to consider this during implementation: the framework handles it at run time.

Extensibility saves a great amount of development time and allows designers to try out multiple solutions. A user interface can be designed and implemented in multiple ways and, depending on which modules the user loads, a specific design is realized. This facilitates experimentation to determine the best design.

Interface Instrumentation

Data collection, agents, and machine learning all require capturing of data, which must be stored or packaged up and sent across the network. User interface interactions is one of these critical data sources. This capability was built into the Fusion framework. It is fairly non-invasive to the developers and provides a host of information, both after the fact and real time. Every user interaction, such as button clicks, typing, and mouse clicks are recorded and saved to a file.

All instrumented data is also packaged up and sent through the network to any agents, machine learning algorithms, cognitive modeling services, or other automated services that are subscribing to the data source. Instrumentation of all operator interactions is critical for human-autonomy teaming. There are many uses for this feature of Fusion. Agents use it to better understand user behavior and take or recommend actions. Machine learning employs instrumentation data to learn how individual users perform, and potentially recommend an interface change either after the fact or in real time (e.g., if a user uses certain buttons more often, the buttons can be reorganized to better suit their use). The data can inform cognitive modeling services, improving researchers' understanding of how the operators are performing.

During evaluations, all instrumentation data is recorded into a comma delimited log file. The experimenter can go back after the fact to analyze performance data. They can determine reaction times and accuracies based on the times of various user actions saved in the file. This was utilized in the IMPACT year one evaluation. For example, a chat module was employed to request tasks from the user. The chat requests and responses were instrumented, as were the user reactions. The experimenter leveraged this data to analyze how quickly and effectively the task was performed. The experimenter also noted any extra steps the operator performed, the sequence of steps taken, and the modality of their actions. All of this data is being used to analyze and improve the user interface as well as any automated services.

Human-Autonomy Dialog through Retrospection

All of the instrumentation data can be used for retrospection. Since all the data is stored, it is natural to allow it to be re-played post process or played back during runtime. Retrospection has two main applications (and potentially more): experimenters can observe what was occurring to analyze why an operator performed an action or series of actions, and operators can "pause" and "rewind" the scenario to get another look at something that occurred in the past, further enhancing the human-autonomy dialog.

The concept of an operator being able to rewind the scenario introduced the concept of a sandbox. The sandbox is an area of the user interface where the operator can invoke actions that aren't instantly carried out by the UxSs. The sandbox allows the user to evaluate autonomy-proposed actions and tweak various parameters before committing to them. Other displays within Fusion still depict current vehicle activities in real time, so the operator maintains effective SA. This can give the operator more insight into the autonomous component actions and reasoning. Another use of the sandbox is to play back the scenario using the instrumented data to see what occurred at some point in the past. This could possibly help operators make more informed, quicker decisions in the future.

Conclusion

Fusion fills a needed role in human-autonomy teaming. Since humans will remain a critical part of autonomous systems development and deployment for the foreseeable future, a clear representation and extension of an operator's intent is required. Fusion was developed to support this, and it continues to expand within the human interface role as it pertains to the Air Force Research Laboratory's goals. It is well positioned to aid in achieving many goals related to autonomy defined by AFRL, the Air Force, and the Department of Defense. Fusion directly addresses two of the four goals established by the AFRL Autonomy Science and Technology Strategy (AFRL, 2013): delivering flexible autonomy systems with highly effective human-machine teaming and creating actively coordinated teams of multiple machines to achieve mission goals. Fusion also participates in addressing two challenges from the 2010 Technology Horizons (USAF Chief Scientist, 2010) for the Air Force: highly autonomous decision-making systems and fractionated, composable, survivable, autonomous systems. Finally, the defense science board (DoD Defense Science Board, 2012) identified perception, planning, learning, human-robot interaction, natural language understanding and multi-agent coordination as key areas that would benefit from improved autonomy. Fusion has encompassed and is continuing work in five of these domains (except perception). Fusion is thus supporting and driving the emerging technologies with which the goals and challenges across all levels of the AFRL hierarchy can be satisfied.

Acknowledgements

This research also supports the "Autonomy for Air Combat Missions: Mixed human/UAV Teams (ATACM)" ARPI that is developing pilot vehicle interfaces for manned/unmanned teaming. The authors would like to acknowledge the software developed at AFRL for the Vigilant Spirit Program (Rowe, 2009) as software engineering principles and concepts have transitioned into the Fusion Framework.

References

Air Force Research Laboratory. (2013, March 5). Autonomy Science and Technology Strategy.

DoD Defense Science Board. (2012, July). Task Force Report: The Role of Autonomy in DoD Systems.

Douglass, S. A., & Mittal, S. (2013). A Framework for Modeling and Simulation of the Artificial. Andreas Tolk (Ed.), Ontology, Epistemology, and Teleology of Modeling and Simulation, Intelligent Systems Series, Springer-Verlag

Kingston, D. B., Rasmussen, S.J., & Mears, M. J. (2009). Base defense using a task assignment framework. *AIAA Guidance, Navigation, and Control Conference*, AIAA-2009-6209

Martin, Robert C., Martin, Micah (2012). Agile Principles, Patterns, and Practices in C#.

Miller, C.A., Hamell, J., Barry, T., Ruff, H., Draper, M.H., & Calhoun, G.L. (2012). Adaptable operator-automation interface for future unmanned aerial systems control: Development of a highly flexible delegation concept demonstration. AIAA Infotech @ Aerospace Conference, AIAA-2012-2529, 1-21.

Rowe, A. J., Liggett, K.K., & Davis, J.E. (2009). Vigilant Spirit Control Station: a research testbed for multi-UAS supervisory control interfaces. Proceedings of the 15th International Symposium on Aviation Psychology.

USAF Chief Scientist (AF/ST). (2010, May 15). Technology Horizons: A Vision for Air Force Science & Technology During 2010-2030, Volume 1.

Vernacsics, P. & Lange, D. (2013). Using autonomics to exercise command and control networks in degraded environments. *18th International Command and Control Research and Technology Symposium*, Alexandria, VA. DTIC ADA 587015.

VISUALIZATION METHODS FOR COMMUNICATING UNMANNED VEHICLE PLAN STATUS

Kyle J. Behymer, Heath A. Ruff
Infoscitex
Dayton, Ohio
Elizabeth M. Mersch
Southwestern Ohio Conference for Higher Education (SOCHE)
Dayton, Ohio
Gloria L. Calhoun, Sarah E. Spriggs
Air Force Research Laboratory
Dayton, Ohio

In order to facilitate a single operator controlling multiple unmanned vehicles, numerous autonomous support tools are being considered. One such candidate tool monitors the situation and alerts the operator when a deviation from the vehicle's plan has occurred. The goal of this research was to develop an effective visualization method for conveying plan deviations. Two interface formats were developed based on a review of the literature: a pie chart and a bar chart. Each format allows an operator to compare values for parameters that have different units, value ranges, and relative priority. Twelve participants were tested using a 2 (chart format) X 3 (number of parameters) X 4 (question type) repeated measures within-participants design. Both objective and subjective data were collected. Participants both preferred and were faster at retrieving parameter state and priority information from the bar chart versus the pie chart.

Intelligent autonomy capabilities are being developed to enable a single operator to control multiple heterogeneous unmanned vehicles (UVs). For example, cooperative control algorithms are being designed that rapidly calculate the most efficient route for a vehicle to take to a specific point while taking into account no fly zones, unpassable terrain, and environmental conditions (Kingston, Rasmussen, & Mears, 2009). Additionally, an intelligent agent is under development (Douglass, 2013) that recommends which vehicle(s) to assign to a specific task based on the UV's probability of success (e.g., detecting the target that the UV is searching for), the UV's estimated time enroute (ETE) to the task location, the time the UV can dwell at the task location once it has arrived, the amount of fuel needed to get to the task location, and the impact assigning the UV to the task will have on other existing tasks. Finally, capabilities are being added to the Rainbow autonomics framework (Verbancsics & Lange, 2013) that monitor the ongoing situation and alert the operator when a deviation from the plan occurs (e.g., a wind shift delays the UV's ETE). Such technologies will allow the human operator to offload manual control of the UVs to the automated system, allowing more time to focus on assigning high-level tasks, monitoring the situation, and adjusting to unexpected events (Draper, 2007).

One of the key challenges facing system developers is designing a human-autonomy interface that allows an operator to monitor the status of assigned high-level tasks and alerts the operator when a deviation from the plan has occurred. A review of the literature provided several potential solutions. One of the most promising was Findler's (2011) Visual Thinking Sprocket. This design was adapted to communicate a given plan's status using the pie chart shown in Figure 1a. This format conveys several types of information to the operator. The size of each pie slice represents the priority weightings that each parameter was given (e.g., ETE was the highest priority, followed by probability of detection [PoD]). The parameters were ordered based on their priority in a clockwise fashion, with the highest priority parameter located at the twelve o'clock position. Color is used to represent a plan's status. If everything is on track for a specific parameter (e.g., the UV is still expected to arrive at task location on time) the middle circular segment of that parameter's pie slice is green and indicates a "normal/ideal" state (see Impact parameter in Figure 1a.). A warning state is represented with three yellow segments and indicates a "slight" deviation from the ideal state. An error state is represented with five red segments and indicates a "severe" deviation from the ideal state. The specific location of the segment with the brighter, more saturated color indicates whether the value exceeds or is less than the desired operating range for that parameter. For example, in Figure 1a the UV is now expected to arrive slightly ahead of the scheduled ETE time (bright yellow segment near pie's center) and with a greatly reduced probability of detection (PoD) than expected (bright red segment near pie's center). This could notify the operator that the UV could possibly get closer to the target for increased sensor quality (since PoD is critically lower than nominal) or could possibly be re-planned to image another target while enroute.

The present study evaluated performance on retrieving UV plan status with this prototype pie chart visualization. The study also evaluated an alternative visualization that encoded the relative priority of each parameter into the width of their respective bars (in contrast to the angle of the wedge used in the pic chart; see Figure 1b). This "bar chart" approach is based on the findings of Cleveland & McGill's (1985) graphical perception task study in which participants were much better at detecting differences in length (e.g., the width of a bar) than differences in angle or area (e.g., the angle/size of the pie wedge).

In the bar chart, the parameters were ordered from left to right based on their priority, with the highest priority parameter at the far left. Otherwise, the coding of the color of each segment of the bars was similar to that employed in each segment of the slices in the pie chart. To summarize, this study compared the effectiveness of the pie chart versus the bar chart at conveying UV plan status information. Additionally, since display clutter can impact information retrieval, the number of parameters presented in the pie and bar charts was varied as well.

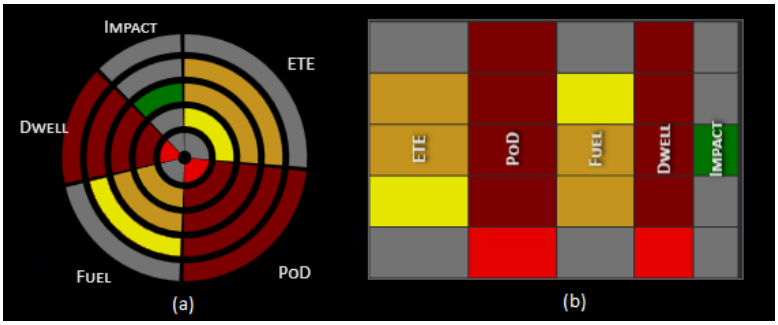


Figure 1. Sample pie chart (a) and bar chart (b) visualizations evaluated to determine effectiveness in conveying the status of parameters related to unmanned vehicle plans.

Method

Participants

A total of twelve volunteer Wright-Patterson Air Force Base employees (8 males, 4 females) between the ages of 22 - 46 (M = 29, SD = 9) participated in this study. All participants reported normal/normal corrected vision and normal color vision.

Experimental Design

Trials were blocked by Chart Format, such that participants completed trials with one Chart Format (pie or bar) before trials with the alternate format. The order of the two Chart Format trial blocks was counterbalanced across participants. Within each of these two blocks, participants completed 96 trials consisting of three 32-trial sets, one with each of three different Number of Parameters conditions tested (3, 5, and 7 parameters), with the order of the sets counterbalanced across participants. Each trial required participants to answer one of four types of questions by retrieving information from a static chart format. Question types included: (1) "What is the state of *parameter X*?" (2) "Which parameter has state of *X*?" (3) "How many parameter(s) have an error or a warning?" and (4) "In comparison to *parameter X*, is *parameter Y* less important, more important, or equally important?" The order in which each question type was posed was randomized with the constraint that each type occurred eight times in each trial set. This resulted in a 2 (Chart Format) X 3 (Number of Parameters) X 4 (Question Type) X 8 (Replication) within-participant factorial design, with each participant completing 192 trials.

Apparatus

Test Stimuli. To generate the Chart Formats (samples shown in Figure 2), 24 data sets were generated, eight for each of the three Number of Parameters (3, 5, and 7) conditions. Each data set specified the priority and operating state (e.g., normal, warning, or error) for each parameter. This step involved defining a unique combination of variables with the goal of representing a range of priorities and operating states, both within each chart and across charts within the trial sets and blocks. Parameter error states had a 50% chance to be nominal, a 15% chance to have a lower warning, a 15% chance to have an upper warning, a 10% chance to have a lower error,

and a 10% chance to have an upper error. For the last step, parameter names were randomly assigned (e.g., for dataset 1, parameter 1 was fuel, for dataset 2, parameter 1 was dwell, etc.).

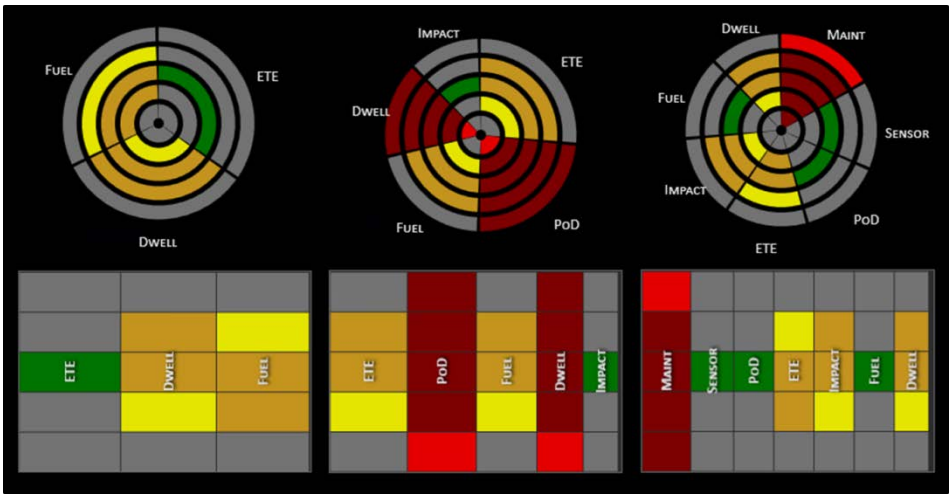


Figure 2. Sample Pie and Bar Chart Formats depicting three, five, and seven parameters of unmanned vehicle plan status. Each chart was approximately 2 X 3 in.

Trial Procedure. The methodology was similar to that employed by Spriggs, Warfield, Calhoun, & Ruff (2010). For each trial, a question was presented along with one chart on a 1920 X 1200 resolution 24 in. widescreen monitor. Participants were trained to click a button labeled "SHOW ANSWERS" when he or she was ready to respond with an answer. Upon button selection, the chart disappeared and candidate responses to the question were presented, as well as the question itself. The Chart Format was removed during the response selection step to prevent the participant from using a process of elimination to answer the question. The participant's task was to select the correct answer as quickly and accurately as possible. Selection of a response blanked the display, except for the presentation of a "NEXT" button; selection of this button initiated the next trial. Thus, progression through the blocks of trials was self-paced with participants selecting buttons via a mouse. Participants could only control progress forward within and across trials; for instance, participants could not return to a previous screen to view the chart again before answering the question. Participants did not receive feedback on their performance during the experimental trials.

Test Sessions

Upon arrival, participants read and signed the informed consent document, filled out a short demographics questionnaire, and were given an overview of the study. Participants were next trained on the specific trial block (Chart Format) they were assigned to complete first. Training consisted of twelve questions (three examples of the four types of questions) using a non-UV scenario in which the state of parameters influencing the success of a hypothetical party were depicted in the charts. Participants were allowed to repeat the training questions until they felt confident in their ability to retrieve information from the chart. Also, participants had to be accurate on the training questions before beginning the experimental trials. Participants were briefed to answer questions as quickly and accurately as possible, and that both speed and accuracy would be recorded.

At the completion of the first trial block, participants were given a Post-Block Questionnaire asking their opinion on the specific Chart Format they just saw. These questions included items asking about their perceived speed, accuracy and general ability to retrieve information, as well as whether the number of parameters made a difference in their ability to answer the questions. Next, the procedures were repeated for the alternate Chart Format. After completion of the training and trial block with the alternate Chart Format, participants were administered another Post-Block Questionnaire. This was followed by a Final Debriefing Questionnaire that included items for participants to compare the two chart formats, indicate their preference, and provide additional feedback. Total session time, per participant, was approximately one hour, with each trial lasting about 5 s.

Results

Data were collapsed across replications. Performance data (response accuracy and time) were analyzed with a repeated measures Analysis of Variance (ANOVA) model. Questionnaire responses were analyzed using paired t-tests and the Kolomogorov-Smirnov nonparametric test of significance.

Response Accuracy. The percent of questions answered correctly was 97% overall. Therefore, data analysis concentrated on mean response time for questions answered correctly.

Response Time. To better reflect the time required to retrieve information to answer the question correctly, response time was calculated from the time that the chart/question was presented until the participant selected the "SHOW ANSWERS" button. The results showed that mean response time to correctly retrieve information was significantly faster with the Bar Chart compared to the Pie Chart Format ($F(1,11) = 5.14$, $p = 0.04$, Figure 3).

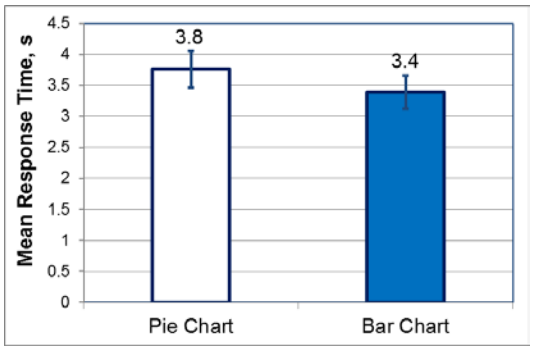


Figure 3. Mean time to retrieve information to answer questions correctly for both the Pie and Bar Chart visualizations. Error bars are the standard errors of the means.

The results also showed significant main effects for Question Type ($F(3,33) = 57.10$, $p < 0.001$) and Number of Parameters ($F(2,22) = 26.20$, $p < 0.001$). These results should be interpreted in light of a significant interaction. Participants' mean response time significantly differed between Question Type as a function of the Number of Parameters ($F(6,66) = 3.18$, $p = 0.045$; Figure 4). Post-hoc Bonferroni t-test results indicated that response time was significantly longer for Question Type #4 (comparison of parameter priorities) than that for the other three question types (all $p < 0.01$). Also, for all question types, mean response time was faster when there were only three parameters depicted in the chart, compared to the seven parameter condition. This result was significant ($p < 0.020$) for three of the four question types (for Question Type #3, count of parameters with errors/warning, $p = 0.080$). There were no significant mean response time differences for tests comparing trials with 3 parameters and 5 parameters, as well as 5 parameters and 7 parameters (all $p > .10$).

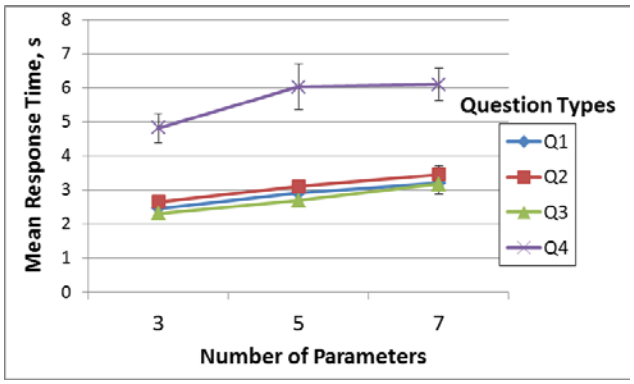


Figure 4. Mean response time for each question type as a function of the number of parameters depicted in the charts. Error bars are the standard errors of the means.

Subjective Data. The results for two items in the Post-Block Questionnaire were aligned with the performance data. Participants' responses on these 5-point rating scales indicated that it was easier to determine the state of a particular parameter as well as the type of warning/error state with the Bar Chart compared to the Pie Chart (respectively, $t(11) = 3.02$, $p = 0.01$ and $t(11) = 2.69$, $p = 0.02$). In contrast, responses to most questions in the Final Debriefing Questionnaire did not significantly differ between the two chart formats: neither format was rated significantly better than the other in terms of speed, accuracy, and general ability in retrieving information to answer the test questions. The only statistically significant finding was in an item addressing information retrieval as a function of the number of parameters. Participants rated the two formats as equal when there were only 3 parameters ($D(12)=.4$, $p<.05$). In contrast, their ratings indicated a slight preference for the Bar Chart format over the Pie Chart when there were 5 or 7 parameters.

Discussion

The results provide empirical support that participants performed better when using the Bar Chart as compared to the Pie Chart. These results are consistent with the results of Cleveland and McGill's (1985) study on basic graphical perception tasks. Though accuracy was extremely high for both chart formats, participants were able to respond more quickly with the Bar Chart. The high level of accuracy might suggest that the questions participants were tasked with answering were relatively easy. Despite this ceiling effect, the fact that there was still a statistically significant difference in participants' mean response time between the Pie and Bar Charts suggests that this difference would likely increase with more difficult questions.

In addition to performing better with the Bar Chart, participants also preferred this visualization method over the Pie Chart for retrieving data when there were five or seven parameters. (Questions were easier to answer in the three parameter condition, explaining why there was no significant difference in response preference). A review of the participants' comments suggests, though, that each chart format had specific advantages. Some participants liked that the Pie Chart was condensed and found the format more intuitive and aesthetically pleasing. Other comments indicated that the Bar Chart format provided a consistent orientation and position of each parameter as well as more separation between parameters. This may explain why some participants said the Bar Chart made it easier to locate parameters and gather a mental picture of parameter state. Another advantage was that comparing parameters was easier since "low" was always on the bottom. Also, the Bar Chart was better to view parameters that were small in value, compared to small Pie Chart slices. One participant summarized the advantages of both chart formats, stating "Pie seemed more aesthetically pleasing than the bar, but the bar seemed ultimately more effective."

The visualization methods examined in this study were designed to enable future UV operations to benefit from automatic monitoring technologies under development, providing the operator near real-time status of a UV plan in progress. It has also been proposed that a similar format be used for autonomy systems to convey one or more *proposed* UV mission plans that the operator should consider. In this manner, the autonomy can illustrate several plans, showing their tradeoffs with respect to different plan parameters. However, employing a Bar Chart similar to that used in the present experiment for each proposed plan would require considerable display space and complicate information retrieval.

For comparison of multiple autonomy-generated plans, the Air Force Research Laboratory has designed a candidate visualization that provides a more concise summary of multiple parameters for multiple plans. With this new Plan Comparison Chart (illustrated in Figure 5; similar to a parallel coordinates plot), each parameter is assigned a column. Parameters are ordered by the priority set for the high-level task (the most important is the leftmost; parameter priorities are represented by column widths). Each of the three plans (A, B, and C) is assigned a unique color and letter. The quality of each plan for specific parameters is mapped onto parameter columns. For example in Figure 5, Plan B is the best plan to use to maximize ETE and Plan C is the best plan to maximize dwell time. Parameter columns are normalized and the yellow and red dashed lines represent threshold levels. This approach depicting multiple plans is under evaluation, as well as other novel displays for transparency into autonomous systems and intuitive interaction methods to support bi-directional human-autonomy dialog.

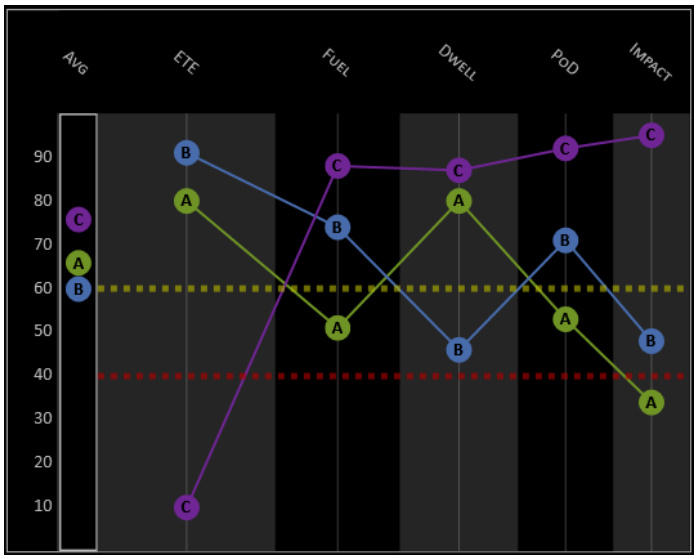


Figure 5. Illustration of Plan Comparison Chart prototype designed and under evaluation by the Air Force Research Laboratory to display the tradeoffs of multiple mission related parameters for multiple autonomy-generated vehicle plans.

Acknowledgements

This research supports the ASD/R&E Autonomy Research Pilot Initiative "Realizing Autonomy via Intelligent Adaptive Hybrid Control" that is developing an Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT).

References

Cleveland, W.S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229,* No. 4716, 828-833.

Douglass, S. (2013). Learner models in the large-scale cognitive modeling initiative. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for adaptive intelligent tutoring systems learner modeling*. Volume1. Orlando, FL: U.S. Army Research Laboratory.

Draper, M.H. (2007). Advanced UMV operator interface. In *Uninhabited Military Vehicles (UMVs): Human Factors Issues in Augmenting the Force*, NATO Report RTO-TR-HFM-078, Chapter 6.

Findler, M. J. (2011). *Cognitively Sensitive User Interface for Command and Control Applications* (Unpublished doctoral dissertation). Wright State University, Dayton, OH.

Kingston, D. B., Rasmussen, S.J., & Mears, M. J. (2009). Base defense using a task assignment framework. *AIAA Guidance, Navigation, and Control Conference*, AIAA-2009-6209.

Spriggs, S., Warfield, L., Calhoun, G., & Ruff, H. (2010). Orientation of a temporal display for multi-unmanned aerial supervisory control. *Proceedings of the HCI (Human Computer Interaction) in Aerospace, Crew Integration Symposium.* Florida: Cape Canaveral.

Verbancsics, P., & Lange, D. (2013). Using autonomics to exercise command and control networks in degraded environments. *18th International Command and Control Research and Technology Symposium*, Alexandria, VA. DTIC ADA 587015.

PANEL ON CROSS-CULTURAL PILOT SELECTION

Diane L. Damos
Damos Aviation Services, Inc.
Gurnee, IL
Mark R. Rose
Air Force Personnel Center
Randolph Air Force Base, TX
Monica Martinussen
The Norwegian Defense University College and
RKBU- Nord, Faculty of Health Sciences, UiT The Arctic University
Tromsø, Norway
Jan L.Lorenz
German Aerospace Center (DLR), Institute of Aerospace Medicine
Hamburg, Germany

Airlines in developing nations often face a pilot shortage. Airlines in such countries either must recruit experienced pilots worldwide or aid in the establishment of national flying schools. The first presentation explains some of the issues the airlines encounter in performing worldwide background screenings on experienced pilots. Selecting either experienced pilots or cadets from a multi-cultural, multi-lingual applicant pool is challenging. The second presentation discusses data comparing native versus immigrant cadet applicants on cognitive tests. The third presentation describes problems associated with adapting a personality assessment to a new culture and language and the resultant predictive validities. Selecting cadets for a national flying school in a multi-cultural developing nation presents additional challenges. The fourth presentation discusses some difficulties in assessing social skills in cadets in a multi-cultural developing country.

Civilian aviation is facing a shortage of both experienced pilots and young people who are interested in flying as a career. The shortage of experienced pilots is particularly acute in parts of Asia and the Middle East. Pilot shortages force airlines in the affected areas to recruit experienced pilots worldwide. Some airlines in the Middle East and Asia now have pilots from more than 50 countries who speak over 30 different native languages. Constructing a valid pilot selection system in a developed country is difficult when the applicants are ethnically mixed with a significant proportion speaking the official or dominant language non-natively. Constructing a selection system for an airline in a developing country is significantly more difficult when the applicant pool is multi-ethnic and multi-lingual.

Airlines in developing nations face several hurdles in developing both an appropriate screening process and a selection process with significant predictive validity. Criminal background checks and examinations of an applicant's driving record often are time-consuming and expensive and sometimes cannot be completed because of the privacy laws of the applicant's home country. The construction of a valid selection system may be hampered by the lack of appropriate selection instruments. Many, if not most, validated selection instruments are administered in a Western European language. Candidates who do not speak a Western European language natively may be at a disadvantage compared to native speakers. Personality tests are even more problematic because they are developed for a specific culture. Using these tests to assess candidates from other cultures may result in misleading results.

A long-term solution for many developing nations is to establish a national flight training school that can meet the country's need for trained pilots. Identifying young men and woman who are likely to be successful in a flying school is difficult for some of the reasons described above. An additional issue, however, is that some developing countries are themselves multi-cultural and multi-lingual.

The four sections comprising this paper describe some of the challenges of background screening and cognitive and personality assessments in multi-cultural environments. The first section deals with international background screening issues for experienced pilots. The next two sections discuss the challenges of using cognitive and personality assessment in multi-lingual, multi-cultural environments. The last section describes the development of a selection system for a cadet program in a multi-cultural nation with no locally validated selection instruments.

Multi-Cultural Pilot Screening Issues

Airlines typically begin the hiring process with background screening. The general purpose of screening is to ensure that an applicant has an acceptable employment history, the required experience minimums, no disqualifying legal events in his/her background, and the educational requirements. To assess the applicant's employment history, airlines typically ask for a list of prior employers and employment dates. The airline also may ask for letters of reference from the pilot's immediate supervisor or from other pilots who are familiar with the applicant. To verify the applicant's flight experience, airlines ask for copies of the pilot's licenses, certificates, and logbooks. Searches for legal issues may take a variety of forms. Airlines will search the pilot's initial civil license issuing authority's (CAA) accident and incident database. Similarly, if the pilot has been employed by different national airlines, the databases of each CAA will be searched. National databases of driving records also will be searched to determine if the applicant was ever convicted of Driving Under the Influence (DUI). If the carrier demands a specific level of education, the applicant may be asked to produce a copy of a diploma.

Screening of domestic applicants is usually straight forward for an airline in a developed country. The applicant's employment history can be relatively easily verified. Current employees who write letters of reference can be contacted. The CAA can be contacted to verify licenses and certificates. A national driving database often is readily available, as is the CAA's accident and incident database. Diplomas often can be verified with a telephone call. If the airline lacks sufficient in-house resources to perform the screening in a timely manner, it may hire specialized companies to perform background searches.

Nevertheless, airlines that hire domestically do encounter some problems. Of these, time may be the biggest issue. For example, the US is currently experiencing a pilot shortage at the regional airline level. Consequently, applicants may be hired and begin initial training before the results of the criminal background and DUI checks are available. Additionally, as the pilot shortage worsens, airlines increasingly must recruit internationally. Smaller airlines may lack the resources to vet international candidates. In the US, this has led the regional airlines to require a prior work history in the US and establish proof of residence. In some cases, the airline may only vet to the extent its resources allow.

Screening for airlines in developing nations is far more problematic. The government may or may not perform an extensive criminal background check before issuing a work permit. The airline may be unable to obtain all of the information from the applicant's CAA; some CAAs do not release pilot license and certification information and accident/incident data to foreign airlines. References from immediate supervisors may be particularly difficult to verify because of high pilot turnover rates at some new airlines in developing nations. An applicant's driving record is often ignored because of the expense of searching national databases, especially if the applicant has worked in several countries.

Multi-Cultural Cognitive Ability Testing

Cognitive ability testing has been a worldwide mainstay of pilot selection systems for more than half a century. Job analyses and validation studies generally support use of cognitive ability testing for evaluating pilot aptitude. However, few cognitive ability tests would meet international testing guidelines for interchangeable use across countries or cultures (International Test Commission, 2001). As noted by Ryan and Tippins (2009), validating a selection tool in one country is a difficult task; the task increases substantially when the process involves multiple countries and cultures. A partial list of factors that may confound the meaning and validity of cognitive test scores across countries includes cultural influences, constructs assessed, specific measures (e.g., item content, item difficulty), translation quality, scoring, study design, method of analysis, abilities of the sample, and the criteria.

This paper presents data from two groups of pilot applicants to the USAF who completed the Air Force Officer Qualifying Test (AFOQT) and the Test of Basic Aviation Skills (TBAS). The first group of consisted of applicants born in the US. The second group consisted of US citizens born outside the country or foreign nationals who later became naturalized US citizens. Because the US is a multi-ethnic, multi-cultural society, membership in these two groups is fuzzy, i.e., applicants have a degree of membership in each of the two groups.

Only the results from AFOQT and TBAS tests that assess constructs commonly included in pilot selection batteries—quantitative ability (Math Knowledge and Arithmetic Reasoning), verbal ability (Verbal Analogies and Word Knowledge), spatial ability (Rotated Blocks and Directional Orientation), and motivation (Aviation Information and Instrument Comprehension)—will be reported here. Table 1 presents descriptive data and correlations between scores on the test and pass/fail with Initial Flight Screening (IFS) and a comparison between the two groups. The results indicate that the cognitive measures (first six rows of Table 1) generally are valid predictors of IFS completion for applicants born inside and outside the US. Cohen's *d* values indicate that average between-group score differences are generally small (.12 to .31), and correlations corrected for dichotomization remain relatively larger for non-US born groups even when taking into account this group's higher IFS attrition rate (24.4% versus 9.3%).

Table 1.
Cognitive Ability Correlations with IFS Graduation/Elimination Across Cultural Groups.

	US (*n* = 4,288)				**Outside US (*n* = 164)**				
Measure	**Mean**	***SD***	**Obs *r***	**Dichot corr *r***	**Mean**	***SD***	**Obs *r***	**Dichot corr *r***	**Cohen's *d***
Math Knowledge	18.53	4.50	.06***	.10	17.80	4.93	.16*	.22	0.16
Arith Reasoning	18.68	4.41	.07***	.12	17.73	4.75	.14	.19	0.22
Verbal Analogies	18.07	3.41	.04*	.07	17.61	3.58	.15	.21	0.13
Word Knowledge	17.65	4.61	.03	.05	16.91	4.88	.16*	.22	0.16
Rotated Blocks	11.56	2.65	.10***	.17	10.74	3.15	.13	.18	0.31
Directional Orientation	0.31	0.88	.15***	.26	0.19	0.81	.25**	.34	0.14
Aviation Information	14.64	3.80	.25***	.44	14.18	4.09	.33***	.45	0.12
Instrument Comprehension	17.04	3.08	.21***	.37	16.45	3.82	.36***	.49	0.19

Note. *$p<.05$ **$p<.01$ ***$p<.001$; Obs *r* = observed point-biserial correlation coefficient; Dichot. corr *r* = observed correlation coefficient corrected for dichotomization of the criterion

Thus, these findings suggest that the same cognitive ability tests may be used to evaluate candidates with different country and cultural backgrounds. Measures of motivation, such as the Aviation Information Test, also appear to be useful predictors for some multi-cultural applicant populations.

Cross-Cultural Use of Trait-Based Personality Measures for Pilot Selection

Trait-based personality tests are frequently used as part of the selection system for both civil and military pilots. Some of these tests are developed for pilot selection, whereas others are measures developed for assessing personality traits in the normal population and are typically based on the Big-five model of personality. How the test results are used may vary between organizations and countries. Sometimes, the test results are used in addition to cognitive ability tests, whereas in other contexts the test results are used in combination with other types of information collected during an interview to assist in hiring decisions. International guidelines for tests and test use for personnel selection (e.g., European Federation of Psychologist's Associations (EFPA), http://www.efpa.eu/professional-development/assessment) state that selection tests should demonstrate reliability and predictive validity in addition to have appropriate norms and consideration for fairness.

Using a personality measure developed in one cultural context for assessing candidates from different language and cultural backgrounds raises a host of issues. Chief among these are problems with translation, appropriate norms, and the effect of biases, such as social desirability, on scores. Additionally, the predictive validity of such a test may decrease when it is used cross-culturally. Several studies have examined the equivalence of Big-five measures in different cultures and there is evidence supporting the five-factor structure across different language and cultural contexts (McCrae, 2002). However, there are also findings indicating that the reliability (in terms of internal consistency) and mean scores may differ between countries (McCrae, 2002). This raises cause for concern when testing applicants with different language and cultural backgrounds. Even for some Scandinavian

countries like Norway, Sweden, and Denmark---which are similar in many ways with a common history, culture and language—applicants may interpret sentences and adjectives in trait-based measures differently, sometimes resulting in invalid personality profiles. Another issue when using personality measures as part of the interview is that there may be cultural differences in self-presentation tactics e.g., in terms of asserting individual excellence and pointing out obstacles where some applicants may be viewed as underselling or overselling themselves compared to the perspective of the interviewer (Sandal et al., 2014). This highlights the importance of cultural competence among those conducting the assessments.

Personality tests in general have demonstrated a relatively modest predictive validity when used for pilot selection , whereas a more recent meta-analysis (K = 8) examining trait-based measures for military pilot selection found mean uncorrected correlations of -.15 and .13 for Neuroticism, and Extroversion, respectively (Campbell, Castaneda, & Pulos, 2009). These findings were supported in a study of US Air Force pilot trainees where some of the Big-five traits predicted training outcomes. However, the uncorrected correlations were generally small (r < .11) (Carretta, Teachout, Ree, Barto, King, & Michaels, 2014).

To conclude, there is some evidence supporting the predictive validity of trait-based measures, but these are mostly based on pilot samples from Europe or the US. In general the correlations are small, but the inclusion of an easy-to-administer, trait-based measure may result in incremental validity in the selection process. In addition, there are studies examining differences between pilot samples and the general population (Meško et al., 2013), and studies linking personality traits to team performance in aviation as well as to accident involvement (for an overview see Ganesh & Joseph, 2005). Taken together these findings indicate that pilot applicants differ from normative samples and that personality traits may indeed be important for pilot performance, even though documenting the predictive validity and generalizability of trait-based measures is still much needed.

Cross-Cultural Differences on Cognitive, Knowledge, and Assessment Center Measures Between Western European and Mauritius Cadet Applicants

A different approach that an airline in a developing country may consider when facing a pilot shortage is sponsoring its own cadet or ab initio program. This approach has a number of challenges. Some of these, however, are not that different from those encountered when hiring experienced pilots from varying nations: In developing states, for instance, even the first step of screening *within* its *own* population proves difficult at times. Frequently, educational and grading systems are not standardized and many young citizens seek educational opportunities abroad – either sponsored by their parents or governmental scholarships. The result is a confusing plethora of degrees, diplomas and grades, etc.

When the German Aerospace Center (DLR) was tasked with selecting viable candidates for the Mauritian Cadet Pilot Programme, we were faced with much the same problem that airlines hiring internationally face: Lacking an indigenous population, Mauritius is a truly multi-cultural nation. In addition to the Indo-Mauritian majority there are three prevalent ethnicities (Creole, Chinese, and European Mauritians) and a variety of practiced religions. How, then, can one make sure that all members of multi-cultural crews work towards creating a productive work environment by creating a shared ‘cockpit’-culture?

Consequently, in addition to employing internationally proven computerized knowledge and cognitive tests (e.g. Zierke, 2014; Maschke, Oubaid & Pecena, 2011) as well an aviation-specific personality test and a final interview, we decided to include an observed team task for the first time in over four decades of global assessment efforts. We would not have felt comfortable recommending applicants without having observed their behavior while working cooperatively in diverse teams.

We were, of course, aware of the possible calamities that might result when a translated task, originally developed and tried in a western European country, was used in a different cultural context. Culture cannot be experimentally varied and the construct culture itself is neither mono-causal nor does it have trivially discernible consequences. Yet, as others have argued, all tests that we used had been explicitly designed to measure inter-individual differences. As long as extraneous conditions are controlled and the methods are sufficiently sensitive and selective, employing such measures in varying (cross-)cultural settings appears warranted (Simon, 2006).

Therefore, we were confident that the task (distributing turns of duty with varying popularity among a group of First Officers) would consistently produce a wide range of relevant behavior and that our rating system

would be suited to measure this behavior adequately. Four dimensions of performance were assessed by independent raters that are significant for safety in aviation: Leadership, Cooperation, Communication and Stress Resistance.

Because of the heterogeneous ethnic composition of the Mauritian population, and the constant need to bridge these differences by means of cooperation and negotiation, and the fact that the original countries of Mauritian ethnicities mostly score high on Hofstede's (2011; Hofstede, Hofstede, & Minkov, 2010) collectivism dimension, we expected a more pronounced cooperative effort from Mauritians when compared to German applicants (Paul, Samarah, Seetharaman & Mykytyn Jr, 2004; McLeod, Lobel & Cox, 1996). In addition we expected some Mauritian applicants to display a tendency to behave more timidly when compared to Germans, resulting in a low average score on Communication.

To test these hypotheses, the results of 52 Mauritians (age M=24.73, SD=2.20, 92.3% male) were compared with those of an analogous group of ab initio candidates from Germany (n=860, age M=21.16, SD=2.04; 90.5% male). The results are shown in Table 2. In contrast to our expectation, the Mauritian applicants were just as communicative as the Germans, even though there was a slightly higher variance among the applicants. Yet, the major hypothesis – that Mauritians behave more cooperatively than Germans – was supported by our data.

Table 2.
Mean scores and standard deviations of Mauritian and German applicants in Cooperation and Communication

	Germans *n*=860		Mauritians *n*=52	
	M	SD	M	SD
Cooperation	3.32	.72	3.77	.91
Communication	3.01	.87	3.06	1.18

Notes. Mean differences: Cooperation: t=3.53 p<.001, g=.62; Communication: t=0.28 p=.783, g=.06

For practitioners, it is important to note that our team task was very well suited for use in a diverse cultural setting. Both the task itself and the range of its rating system provided markedly selective results and , thus, a good basis for decisions and recommendations. In this specific example, we simply had to shift our focus *within* the existing methodology – from a given cooperation towards an emphasis on leadership skills.

Judging from this experience, we would encourage the application of work-related team tasks to directly observe the behavior in diverse teams as a measure for intercultural cooperation and suitability for a multi-cultural cockpit environment. Of course, every effort must be made to create tasks which are culturally fair and unequivocal for all participants. In our opinion, the inclusion of a team task extends the validity of the whole assessment process, e.g., by contributing hypotheses for subsequent interviews.

Conclusion

Because of pilot shortages in many parts of the world, airlines are forced either to recruit experienced pilots internationally or to develop their own cadet programs. Screening foreign experienced pilots is difficult because of privacy issues of some countries and the cost of searching numerous databases. Selecting pilots from multiple countries is problematic because it requires valid selection instruments that can be used cross-culturally. Currently, cognitive, personality, and motivational-assessment instruments show some cross-cultural validity but need further development. Team performance assessments may provide a valuable tool for determining how well an individual can function in a multi-cultural environment.

References

Campbell, J. S., Castaneda, M., & Pulos, S. (2009). Meta-analysis of personality assessments as predictors of military aviation training success. *The International Journal of Aviation Psychology, 20*, 92-109. doi:0.1080/10508410903415872

Carretta, T. R., Teachout, M. S., Ree, M. J., Barto, E. L., King, R. E., & Michaels, C. F. (2014). Consistency of the relations of cognitive ability and personality traits to pilot training performance. *The International Journal of Aviation Psychology, 24*, 247-264. doi: 10.1080/10508414.2014.949200

Ganesh, A., & Joseph, *C.* (2005). Personality studies in aircrew: An overview. *Indian Journal of Aerospace Medicine, 49*, 54-62.

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture, 2*(1). Retrieved from http://dx.doi.org/10.9707/2307-0919.1014

Hofstede, G., Hofstede, G.J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind*, (3rd ed.), NY: McGraw Hill.

International Test Commission (2001). *International Test Commission guidelines for test adaptation*. London: Author.

Maschke, P., Oubaid, V., & Pecena, Y. (2011). How do astronaut candidate profiles differ from airline pilot profiles? Results from the 2008/2009 ESA astronaut selection. *Aviation Psychology and Applied Human Factors, 1*, 38-44.

McCrae, R. R. (2002). Cross-cultural research on the five-factor model of personality. *Online Readings in Psychology and Culture, 4*(4). Retrieved from http://dx.doi.org/10.9707/2307-0919.1038

McLeod, P. L., Lobel, S. A., & Cox, T. H. (1996). Ethnic diversity and creativity in small groups. *Small Group Research, 27*, 248-264.

Meško, M., Karpljuk, D., Štok, Z. M.,Videmšek, M., Bertoncel, T., Bertoncelj, A., & Podbregar, I. (2013). Motor abilities and psychological characteristics of Slovene military pilots. *The International Journal of Aviation Psychology, 23*, 306-318. doi: 10.1080/10508414.2013.833750

Paul, S., Samarah, I. M., Seetharaman, P., & Mykytyn Jr, P. P. (2004). An empirical investigation of collaborative conflict management style in group support system-based global virtual teams. *Journal of Management Information Systems, 21*, 185-222.

Ryan, A.M. & Tippins, N. (2009). Designing and implementing global selection systems. Malden, MA: Wiley.

Sandal, G. M., van de Vijver, F., Bye, H. H., Sam, D.L., Amponsah B., Cakar, N.,,..Tien-Lun Sun, C. (2014). Intended self-presentation tactics in job interviews: A 10-country study. *Journal of Cross-Cultural Psychology*, 45, 939-958. doi: 10.1177/0022022114532353

Simon, P. (2006). The solution of fundamental methodological problems in cross-cultural psychology by guaranteeing the equivalence of measurements. In J. Straub, D. Weidemann, C. Kölbl & B. Zielke (Eds.), *Pursuit of meaning: Advances in cultural and cross-cultural psychology* (pp. 269-292). Bielefeld: Transcript.

Zierke, O. (2014). Predictive validity of knowledge tests for pilot training outcome. *Aviation Psychology and Applied Human Factors*, *4*, 98-105. doi: 10.1027/2192-0923/a000061

INVERTING THE HUMAN/AUTOMATION EQUATION TO SUPPORT SITUATION AWARENESS AND PREVENT LOSS OF CONTROL

Alex Kirlik, Kasey Ackerman, Ben Seefeldt, Enric Xargay, Donald Talleur,
Ronald Carbonari, Naira Hovakimyan and Lui Sha
University of Illinois at Urbana-Champaign
Urbana, IL

Anna Trujillo and Irene Gregory
NASA Langley Research Center
Hampton, VA

Despite the contributions of automation to aviation safety and efficiency, the problems associated with technology-centered rather than human-centered automation are well known: decreased pilot situation awareness, deterioration of manual piloting skills, difficulties pilots experience when trying to jump into the loop when needed, and so forth. We present a prototype architecture for human-automation interaction that reverses their traditional roles: in our design, the automation "looks over the shoulder" of the pilot and jumps into the loop when needed rather than the other way around to prevent aircraft loss-of-control (LoC). The architecture exploits the LoC prevention algorithm proposed by Wilborn and Foster (2004). This quantitative definition uses a set of five two-dimensional envelopes relating to critical flight parameters that account for aircraft flight dynamics, aerodynamics, structural integrity, and flight control use. The LoC algorithm is used to both present the pilot with a graphical cockpit display depicting aircraft state in relation to these safety envelopes in passive mode (i.e., depicting behavior-shaping constraints), and also to compensate for ineffective pilot inputs that would cause aircraft LoC in active mode. The prototype system has been implemented in our flight simulation lab and the details underlying the design will be presented. We conclude by describing the design of an experiment we are using to evaluate this human-automation interaction design concept and its implementation.

Introduction

The presence of automated control systems in aircraft is ubiquitous. As demands for aircraft safety and efficiency have increased, so too have levels of complexity found in these systems. While this automation has resulted in significant safety benefits, increased incidents and accidents due to a lack of pilot engagement, variously described as the “out-of-the-loop” (OOTL) problem (Endsley & Kiris, 1995) or “out-of-the-loop unfamiliarity” (OOTLUF) problem (Wickens & Hollands, 2000), have prompted much recent research, including a recent study on automation-induced task-unrelated thoughts or “mind wandering” by pilots (Casner & Schooler, 2014). Ironically, the increased reliability of automation brings with it an increased level of safety coupled with the fact that in many cases, automation behavior becomes transparent only at the point of failure. When this occurs, pilots are thrown ‘back into the loop’ to attempt recovery. When pilots are required to reenter the control loop unexpectedly, their ability to do so effectively is often compromised, a phenomenon known as “automation surprise” (Billings & Woods, 1994) or the “return-to-manual-control deficit” problem (Hadley et al., 1999). Various attempts have been made to cope with related problems such as “mode confusion” and “mode error” (Sarter & Woods, 1995; Degani & Heymann, 2002) which result from pilots having an inadquate understanding of automation due at least in part to the fact that automation behavior is insufficiently revealed or presented in cockpit interfaces.

We believe that in large part, these difficulties can be attributed to the fact that there is a significant loss of situational awareness surrounding automation state. While a pilot may be fully aware of various flight variables (such as heading, altitude or airspeed), that appear on the primary flight display, they have few indications of automation state apart from an active/inactive marker. This setup disregards the fact that highly complex automation contains huge amounts of system information regarding not only current aircraft state, but also the control corrections necessary to maintain that state and potential future states. These automated systems can be seen as ‘silent co-pilots’, controlling the plane but offering no insights into their process. D. A. Norman’s (1990) paper on automation sets up a thought experiment where the reader is asked to compare flying with the aid of an automated system with a human flight crew. At the point of failure, Norman describes how the “informal chatter” in the human-only cockpit facilitates early detection of flight problems, while the automated system silently compensates

until a more dramatic failure occurs. Overcoming problems of decreased awareness of automation then becomes a problem of reintroducing this informal chatter into the automated cockpit. Pilots should be given a steady stream of non-intrusive information to allow a continuous monitoring of automation's contribution to achieving safe flight.

In this paper, we present both a concept for coupling pilots and control automation and display designs that integrate information from an automated system into the traditional flight display. Additionally, we provide a novel display placed to the right of the primary flight display dedicated solely to surfacing automation information. Our work centers around a dynamic Flight Envelope Protection (FEP) system augmented with logic for loss-of-control (LoC) prediction and prevention. Previous work addressing technological solutions to LoC prevention, especially in off-nominal conditions, appears in Belcastro & Jacobson (2010), Belcastro (2011) and Belcastro (2012), using both visual and aural methods for notification and cueing, and adjustable autonomy (Kaber, 2012) in the way authority is partitioned between pilots and automation. Relatedly, Connor et al. (2012) present an approach to cockpit display design using perceptual cueing to indicate corrective control actions that should be taken to avoid aircraft LOC.

Our own approach toward reducting LoC events is part of a larger set of efforts to develop technologies to prevent incidents and accidents based on a combination of the AIRSAFE concept described in Belcastro's research (op. cit.), technologies for fault-tolerant flight control (Hovakimyan & Cao, 2010; Hovakimyan et al, 2011), fault detection and isolation (Lee et al., 2014), safe flight envelope estimation and detection (Tekles et al., 2014) and LoC prediction and prevention (Chongsvisal et al., 2015). The ovararching concept is to reduce LoC through the novel use of a set of safety envelopes defined by a set of flight parameters. This vocabulary of envelopes and safety limits is extended to our display enhancements, and to a logic by which flight envelope protection automation selectively engages to compensate for combinations of pilot commands and environmental disturbances to maintain stability to prevent LoC events when detected. We believe that this form of joint, compensatory architecture for coupling humans and automation is much in the spirit of the "horse and rider" or "H-Metaphor" guideline for vehicle automation and interaction (Flemisch et al., 2003). Additionally, we present our experimental design for a series of pilot-in-the-loop flight tests intended to gauge the efficacy of our system in scenarios requiring pilots and automation to maintain control in the presence of significant wind shear. We predict that the increased situational awareness provided by our augmented displays, coupled with an ability for automation to adaptively augment pilot inputs will result in improved overall performance, seen primarily through a decrease in the number of LoC events, the number of envelope exceedences and in the time needed to recover safe aircraft state.

Loss of Control Envelopes

The system described in this paper is based on a quantitative definition of loss of control as described in Wilbourn and Foster (2014). For a more comprehensive description of the system, refer to Chongvisal et. al. (2014). The displays described in future sections utilize information taken from the FEP system. This system takes advantage of five predefined safety-envelopes. These envelopes describe overall loss-of-control (LoC) as a relationship between dynamic flight parameters. Flight protection then, becomes a task of ensuring these flight parameters remain within the defined safety envelopes. Each envelope defines two boundaries, 'soft' LoC limits which are more restrictive but less critical, and 'hard' limits that are maintained by the protection scheme. The protection scheme computes these dynamic envelopes and determines an ideal control solution.

Display Design

In this section we present an overview of our display design. The material in this section is adapted from Ackerman et. al. (2015), which contains a more detailed account of automation and display design. The primary flight display (PFD), shown in Figure 1, is based on the standard flight display design, with the addition of three non-standard displays (Angle of Attack, Angle of Sideslip, and Load-factor).

Primary Flight Display

The elements of this modified PFD are further enhanced by the addition of FEP derived limits. As discussed in the previous section, the framework for LoC used by our system uses sets of hard and soft limits to define envelopes around critical flight features. A set of indicators (airspeed, pitch/roll, AoA, AoS, and load-factor) are modified to display not only current status, but current envelope position. By providing salient cues concerning boundaries used by automation, pilots will become more aware of the reasons for automation engagement.

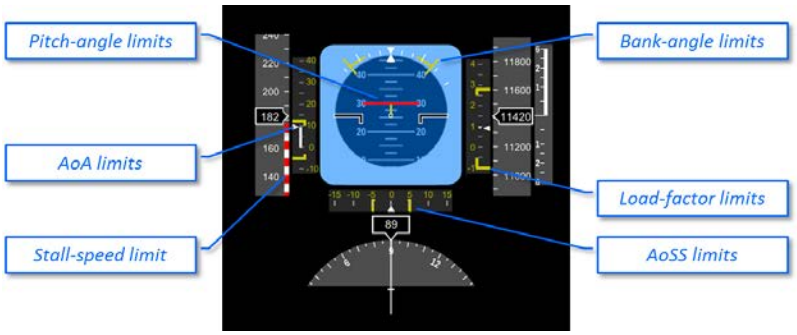


Figure 1. Primary flight display with FEP limit augmentations.

The general design for a limit indicator shows both hard and soft limits. For any given measurement, a yellow line is drawn parallel with the indicator movement. This line represents the range of values that are between the soft and hard limits. Moving into this region is an exceedance of the soft limits, and proper care should be taken that hard limits are not reached. The hard limit is marked at the end of the soft limit by a perpendicular yellow line. In the case of a soft limit excursion, this hard limit line turns red, drawing the pilot's attention. In Figure 1, we see that the pilot is within the safety envelope defined for bank, but has pitched upward at too extreme an angle. The soft limit has been crossed, and the hard limit line indicates this change.

This general design is replicated across other PFD parameters. However, the altimeter and heading indicator, which display aircraft position, and not aircraft movement are not augmented, as they have no limits defined in the FEP system. Additionally, the lower limit for airspeed does not show the traditional limit indicator, but rather is marked by a red and white bar that displays stall speed. This is in keeping with current design practices.

Envelope Protection Display

The primary flight display described above is primarily concerned with indicating current status and the presence of limits in relation to these measurements. It is necessary to distinguish between a purely descriptive interface, and one which provides feedback in relation to control input. The additional envelope indicators in the PFD do not provide the pilot with directly actionable information. When the FEP system is active, it directly limits pilot control input in response to potential envelope excursions. To communicate saftey envelopes in direct relation to pilot control inputs, we add the Envelope Protection Display (EPD) to the right of the PFD.

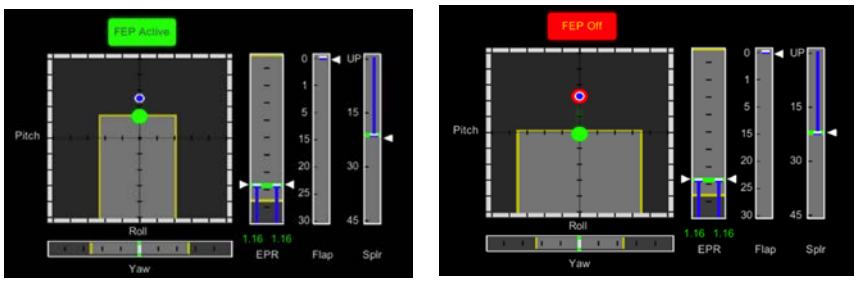


Figure 2a/b. The EPD in FEP On mode and Active state (left) and in FEP Off mode (right).

There are two main elements to the Pilot Input Display (PID): a square pitch/roll command box and a horizontal yaw command bar below the box. The box and the bar are marked by axis marks at regular intervals. The pitch/roll box and yaw bar depict the entire range of movement of the control yoke and rudder pedals respectively.

Within both display areas is a light gray rectangle bordered by yellow showing safe control inputs. Constraining control input to these rectangles guarantees hard FEP derived limits are not exceeded. These rectangles move in response to changing flight status. Any area beyond the yellow boarder is considered 'unsafe' operation. Both displays are marked by two control input indicators. The first, a blue circle or bar outlined in white, represents

the directed pilot input. This always corresponds to the position of the yoke or rudder pedals. The second, a larger green indicator, represents the ideal FEP derived control position. This marker will always remain inside the light gray box. As seen in Figure 2a, the green marker tracks the blue one indicating the safe position that is closest to the directed pilot input. When the pilot is directing input that is inside the light gray box, the two markers will overlap. At the top of the PID is an annunciator indicating the status of the FEP system. The behavior of this annunciator is dependent on the specifics of mode operation as described below.

Display Operations

FEP On. One motivation for our design was the observation that during previous FEP flight trials, pilots were unaware of the impact FEP automation was having on aircraft control. When control input was modified during a difficult flight scenario, they perceived these modifications as a loss of (their) control. Despite the fact that the control system was actually maintaining aircraft stability effectively, pilots felt hindered by the system. The design described above aims to provide pilots with a window into the automation, the so called "informal chatter" mentioned earlier. This allows pilots to construct a more veridical model of overall system state, and easily accounts for situations where pilot control input requires active compensation from FEP-based automation.

While the FEP system is in the "On" mode, there are two states that are reflected on the FEP state annunciator. When the pilot is current maintaining aircraft state and the directed input is within safety envelopes, the system is "Armed". This is noted by the FEP state annunciator being colored yellow and containing the text "FEP Armed". While in this mode, both blue and green indicator marks are overlapped, showing that the pilot is in complete control of the aircraft. The second possible state is "Active". This state occurs when the pilot directs a control position that is outside the displayed envelopes of safety. At this point, the FEP modifies pilot input, directing the aircraft using a control position that is inside the safety envelopes. When the excursion takes place, the state annunciator changes from yellow to green, and displays the text "FEP Active". Additionally, the blue indicator will move beyond the yellow boarder, while the green indicator remains within the envelope. In this situation, the pilot is no longer in direct control of the aircraft, rather the input displayed by the green indicator is being used.

FEP Off. While the primary configuration is intended to inform pilots of system state by making salient the impact of automation, it is possible to fly the aircraft without the use of the FEP system. Turning the FEP off allows the pilot to be in direct control of the aircraft, even in situations where the plane is in a potential loss of control event. In this mode, the FEP continues to compute safety envelopes and ideal control positions without modifying pilot input while continuing to display these envelopes. However, when the pilot directs the plane outside an envelope of safety, when the blue dot exceeds the yellow border, it is surrounded by a red halo, highlighting the excursion. Additionally, the state annunciator turns red, communicating to the pilot that an envelope has been breeched. The green marker remains inside the envelope of safety, marking an ideal position. This can be seen in Figure 2b, where the pilot has pitched up higher than is safely allowed. A 5-minute explanatory video depicting our design prototype in operation is available at: https://www.youtube.com/watch?v=gLZpFfXwGVQ#t=282.

Experimental Evaluation

As part of our continuing research, we are beginning experimental trials evaluating our design. We will be performing pilot-in-the-loop testing at the flight simulator at the Illinois Simulator Laboratory. Our simulator is a Frasca 142 cockpit, with the primary flight display panel replaced with a digital display panel. Surrounding the cockpit, are three projectors providing a 140° view of the outside world generated by X-Plane, while our physics model and FEP system are implemented in Matlab/Simulink. Our study participants are drawn from the Parkland College Institute of Aviation at UIUC. We are recruiting both students and instructors, with the minimum requirement for participation being Private Pilot certification.

We will be conducting a within-subject design comparing three aircraft configurations. Our control condition is the basic aircraft with no additional modifications or enhancements. Pilots will fly using the standard PFD with no envelope display. FEP will be disengaged entirely and the pilots will be in full control of the aircraft. The two additional experimental conditions are intended to provide a general 'stepping-up' of support in terms of display and automation aid. In both of these conditions, we will provide the pilot with the full extent of display modification. Pilots will fly this configuration both in "FEP Off" and "FEP On" modes.

In all configurations, pilots will complete ten separate scenarios designed to emulate challenging wind shear. Each scenario will be approximately two minutes long and consist of an initial level flight period, followed by a unique wind shear profile, in which the pilot will be instructed to regain control of the aircraft and return to initial flight conditions. These scenarios are designed to prompt the excursion of at least one safety envelope. For each of the three conditions, these ten scenarios will be run in a unique order, but the order will remain the same for all participants completing the given condition. To eliminate ordering effects, we counterbalance condition ordering.

Each participant will be evaluated on four separate days. The first of these will be exclusively used for participant training. On each subsequent day, the participant will fly the set of ten scenarios under different configurations. There will be a short training period for pilots to become familiar with the interface design before evaluation begins. Between scenarios the pilots will be prompted to complete an evaluation using a modified Cooper-Harper scoring system to determine a performance score. We will also evaluate pilot performance by recording live flight variables, specifically those related to envelope exceedance and ideal flight path deviation. In order to assess the impact of our novel interface panel we will also be gathering eye tracking data to determine the manner in which the various experimenal conditions affect participants' visual attention allocation patterns.

We anticipate testing will reveal the benefit of additional display information regarding automation state. The increase in situational awareness should allow pilots to apply expertise to the scenarios in a more accurate way. Rather than relying on secondary information regarding automation (that is, 'testing' the automation response to certain control movements), pilots will be able to directly observe the impact of automation. Even without automation control directly affecting flight, understanding flight dynamics in terms of safety envelopes and observing the impact certain control movements have on their parameters should increase performance, thereby reducing the number of times pilots either lose control of the aircraft or exceed LOC safety envelope barriers. We expect best perforrmance in the final condition when automation selectively provides active control compensation.

Conclusions

Our work aims to surface information provided by advanced automation systems to pilots in salient ways that promote overall situational awareness. We believe certain failures are caused not by a failure in automation, nor by a lack of pilot training or skill, but rather by insufficient communication between the two. We believe the insights gathered in this work apply not only to the domain of aircraft automation and control, but to a variety of domains which necessitate automation systems and human operators working in conjunction. In particular, the language of safety envelopes seems particularly effective for a subset of these problems. Future work will be informed by the results of our initial simulator study and focus on developing generalizable methods for use in other disciplines.

Acknowledgements

This research was supported by the National Aeronautics and Space Administration and by the National Science Foundation.

References

Ackerman, K., Xargay, E., Talleur, D. A., Carbonari, R. S., Kirlik, A., Hovakimyan, N., Gegory, I. M, Belcastro, C. M., Trujillo, A., & Seefeldt, B. D. (2015, January). Flight envelope information-augmented display for enhanced pilot situational awareness. *AIAA Infotech @ Aerospace.*

Belcastro, C. M. (2011). Aircraft loss of control: Analysis and requirements for future safety-critical systems and their validation. *Control Conference (ASCC), 2011 8th Asian*, 399-406.

Belcastro, C. M. (2012). Loss of control prevention and recovery: Onboard guidance, control, and system technologies. *AIAA Guidance, Navigation, and Control Conference*, AIAA-2012-4762, Minneapolis, MN.

Belcastro, C. M. & Jacobson, S. R. (2010). Future integrated systems concept for preventing loss-of-control accidents. *AIAA Guidance, Navigation, and Control Conference*, AIAA-2010-8142, Toronto, Canada.

Billings , C. E. , & Woods , D. D. (1994). Concerns about adaptive automation in aviation systems. In M. Mouloua

& R. Parasuraman (Eds.), *Human Performance in Automated Systems: Current Research and Trends* (pp. 264–269). Hillsdale, NJ: Erlbaum.

Casner, S. M. & J. Schooler (2014). Thoughts in flight: Automation use and pilots' task-related and task-unrelated thought. *Human Factors, 56*(3), 433-442.

Chongvisal, N. T., Tekles, N., Xargay, E., Talleur, D. A., Kirlik, A., & Hovakimyan, N. (2014). Loss-of-control prediction and prevention for NASA's Transport Class Model. *AIAA Guidance, Navigation and Control Conference,* National Harbor, MD.

Conner, K.J., Feyereisen, J., Morgan, J. & D. Bateman (2012). Cockpit displays and annunciation to help reduce loss of control (LOC) or lack of control (LAC) accident risks. *AIAA Guidance, Navigation and Control Conference*, Minneapolis, MN.

Degani, A. & Heymann, M. (2002). Formal verification of human-automation interaction. *Human Factors, 44*(1), 28-43.

Endsley, M. & Kiris, E. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*, 381–394.

Flemisch, F. O., Adams, C. A., Conway, S. R., Goodrich, K. H., Palmer & M. T. Schutte (2003). The H-Metaphor as a Guideline for Vehicle Automation and Interaction, NASA/TM-2003-212672, NASA Langley R. C.

Hadley, G. A., Prinzel , L. J., Freeman, F. G., & Mikulka, P. J. (1999). Behavioral, subjective and psychophysiological correlates of various schedules of short-cycle automation. In M. W. Scerbo & M. Mouloua (Eds.), *Automation Technology & Human Performance* (pp. 139–143). Mahwah, NJ : Erlbaum.

Hovakimyan, N. & Cao, C. (2010*). L_1 Adaptive Control Theory*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Hovakimyn, N., Cao, C., Kharisov, E., Xargay, E. & I. M. Gregory (2011). *).* L_1 adaptive control for safety-critical systems. *IEEE Control Systems Magazine, 31*(5), 54-104.

Kaber, D. B. (2012). Adaptive automation. In J.D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (pp. 594-609). New York: Oxford University Press.

Lee, H., Snyder, S. & N. Hovakimyan (2014). An adaptive unknown input observer for fault detection and isolation of aircraft actuator faults. *AIAA Guidance, Navigation and Control Conference*, AIAA-2014-0026, National Harbor, MD.

Norman, D. A. (1990). The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *327*(1241), 585-593.

Sarter, N. B. , & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors, 37*(1), 5–19.

Tekles, N., Xargay, E., Choe, R., Hovakimyan, N., Gregory, I. & F. Holzapfel (2014). Flight envelope protection for NASA's Transport Class Model. *AIAA Guidance, Navigation, and Control Conference*, AIAA-2014-0269, National Harbor, MD.

Wickens, C. D. & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Wilborn, J. E., & Foster, J. V. (2004). Defining commercial transport loss-of-control: A quantitative approach. In *AIAA Atmospheric Flight Mechanics Conference and Exhibit,* AIAA, Providence, RI.

FUNCTIONAL COMPLEXITY FAILURES AND AUTOMATION SURPRISES: THE MYSTERIOUS CASE OF CONTROLLED FLIGHT INTO STALL (CFIS)

Lance Sherry
Center for Air Transportation Systems Research at George Mason University
Fairfax, Virginia
Robert Mauro
Decision Research & University of Oregon
Eugene, Oregon

Nineteen modern airliner Loss of Control (LOC) accidents resulting in aerodynamic stalls were analyzed. These accidents involved structurally and mechanically sound aircraft decelerating through the $1.3V_{Stall}$ buffer to the stall airspeed - i.e. a Controlled Flight into Stall (CFIS). The analysis produced three main observations: First, the accidents were "functional complexity" failures -- the result of a complex sequence of behaviors of the automation functions. There were no consistent: failures that triggered the events (e.g. sensor failures), effects of triggering events on the automation (e.g. mode change), or commands issued by the automation (e.g. thrust setting). Second, the pilots were unable to intervene effectively due to the absence on the flight deck of relevant information and salient cues to monitor these rare events or their effects. Third, there was no single intervention that could mitigate all of these accidents. Implications for flight deck procedures, training and automation design are discussed.

Modern commercial aviation has achieved a remarkable safety record. The 10^{-13} accident rate for modern airliners (IATA, 2013) is four orders of magnitude smaller than the regulatory 10^{-9} target level of safety (ICAO, 2013). This is a testament to the operators, designers and regulators of this transportation system. The low accident rate is a product of the meticulous aviation "safety system" that is the envy of other domains (Bayuk, 2008).

As aviation safety has improved, the character of aviation accidents has changed. Today, commercial aviation accidents are unlikely to be the direct result of a major engine or other equipment malfunction. Instead, they are likely to be the result of a complex combination of technological, environmental, and human factors. In many cases, the accident aircraft itself is mechanically and structurally sound when the accident occurs. This paper describes an analysis of 19 loss of control (LOC) accidents which culminated in an aerodynamic stall. The sequences of events that led to these accidents were examined in an attempt to understand the complex interplay between the factors that precipitated the accidents, rather than determine *the* failure that may have caused each accident.

Methods

Official accident reports and industry airline safety data-bases (e.g. Aviation Safety Network) were searched to identify airline operations (Part 121) and air taxi operations (Part 135) in which aircraft transitioned from safe energy-states within the safe operating speed envelope to an energy-state outside of the safe operating speed envelope by decelerating through $1.3V_{Stall}$ and V_{Stall}. Accidents in the takeoff phase in which the aircraft did not accelerate enough to achieve a safe lift generating airspeed were not considered because a safe flying speed was never achieved. The accident scenarios were identified directly from the accident reports.

Results & Discussion

Nineteen accidents were reviewed. A list and summary of the accidents and their characteristics is included in Sherry et al (2014). All 19 accidents and incidents can be described by the sequence of events presented in Figure 1: A *triggering event* (e.g. sensor failure) has an *effect on the automation* (e.g. mode change). This leads to an *inappropriate command* for pitch or thrust. The inappropriate command occurs while the aircraft is experiencing a relative *deceleration* to the minimum safe operating speed. The relative deceleration is the result of either the aircraft decelerating to a fixed minimum safe operating speed or a change in the minimum safe operating speed (e.g. due to ice contamination on the wing). Finally, the flight crew *fails to intervene* to arrest the deceleration through the minimum safe operating speed and into controlled flight into a stall (CFIS).

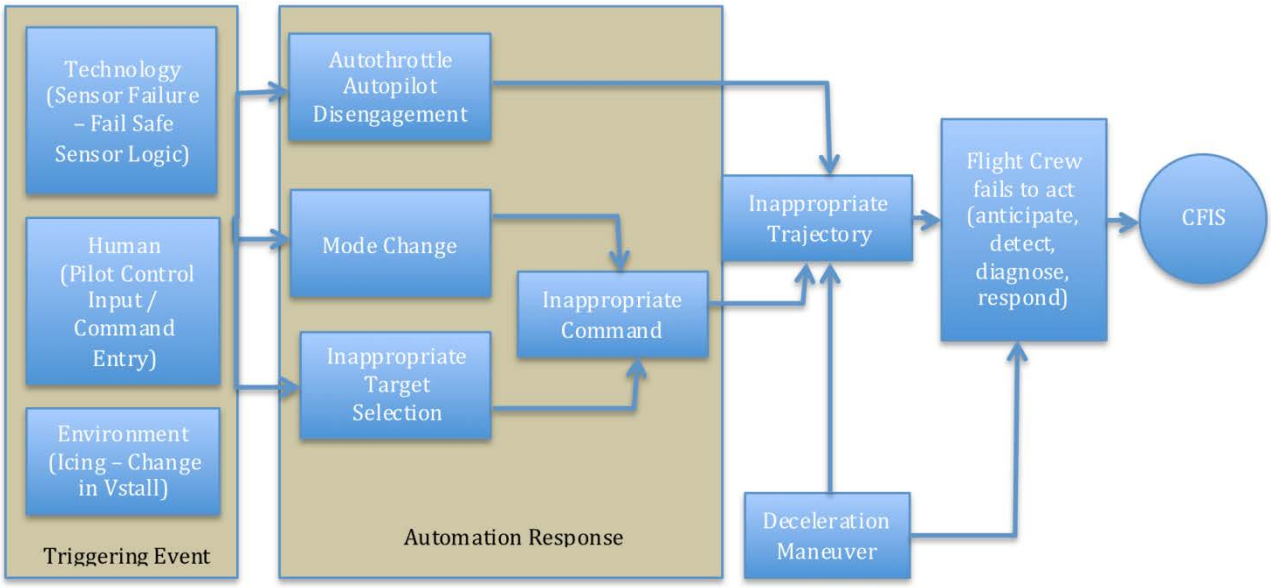


Figure 1. Sequence of Events Leading to Controlled Flight Into Stall (CFIS)

Phase of Flight

The CFIS events occurred in all phases of flight. Six accidents occurred while the aircraft were climbing to the cleared altitude with maximum thrust. In one case (Midwest 490), the flight crew had selected a fixed rate-of-climb at an airspeed that could not be maintained even at maximum thrust. In all the other CFIS cases that occurred during climb, errors in airspeed resulted in inappropriate commands. One accident occurred during cruise. Two accidents occurred during descent and 9 occurred during approach. The accidents that occurred during cruise involved sensor failures that could not be handled by the automation. The automation disengaged, transferring control of the flight trajectory to the flight crew. However, the flight crew was faced with the same inaccurate sensor data that caused the transfer of control and failed to regain control of the aircraft. All the descent and approach cases involved level flight or fixed rate of descent in which the thrust setting was too low to maintain the desired airspeed (e.g. XL Germany 888T, Colgan Air 3407) or the autothrottle was disengaged and no longer controlling airspeed (e.g. AAL 903) or a mode change occurred to a mode that no longer actively controlled to the target airspeed (Asiana Air 214, TA 1951).

Triggering Events

Sensor failures and failures in the associated sensor fail-safe logic were the most common triggering events (see Table 1). However, there was no single sensor-type or class of failure that was common to all of these cases. Sensors that experienced failures included: angle-of-attack sensors (XL Germany), pitot tubes (AF 447, Midwest 490, BirgenAir 301, NWA 6231), and radio altimeters (TA 1951). Further, the fail-safe sensor logic included both voting mechanisms that select the perceived non-failed sensor, as well as averaging mechanisms that average sensor inputs. Changes in aerodynamic characteristics of the aircraft due to icing conditions contributed to three accidents (Colgan Air EWR/BTV, Midwest 490, American Eagle 3008). Flight crew errors contributed to six accidents (Colgan Air EWR/BTV, Colgan Air 3407, United Express 629, KingAir Evereth, Asiana Air 214, Provincial Airlines). For example, the flight crew of Asiana Air Flight 214 inappropriately selected Flight Level Change (FLCH) mode during the approach (NTSB, 2013b). The selection of FLCH resulted in a change to a “dormant” autothrottle mode in which the autothrottle no longer controlled to the airspeed target. In some cases, the triggering event was not identified (AAL 903, ThomsonFly Bournemouth, ThompsonFly Belfast).

Effects of Triggering Events on the Automation

The triggering events had a variety of effects on the automation. When the triggering event was a sensor failure, there were four types of effects on the automation: 1) the automation disengaged (e.g. Air France 447), the

automation mode changed (e.g. Turkish Airlines 1951), 3) the target used for control was calculated incorrectly (e.g. XL Germany T888), or 4) the generated command for pitch or thrust was inappropriate for the current maneuver (e.g. BirgenAir 301).

Table 1.

CFIS Events by Category of Triggering Event and Effects of Triggering Events on the Automation.

Category of Triggering Events	**Effects of Triggering Events on Automation**	**CFIS Accidents/Incidents**
Sensor failures and associated fail-safe sensor failure logic	Disengagement	AF 447, Provincial Airlines, Midwest 490, AAL 903, Air France 447
	Mode change (A/T) Moded-Input Device mode change (Throttle Levers)	Asiana Air 214, TA 1951
	Target error	XL Germany
	Command error	Iceland Air 662, Midwest 490, Provincial Airlines, BirgenAir 301, NWA 6231
Changes in the aerodynamics of the aircraft	Stall speed calculation	Colgan Air/EWR-BTV, American Eagle 3008
Flight crew entry		Colgan Air EWR/BTV, Colgan Air 3407, United Express 629, King Air Eveleth
Unknown events		AAL 903, ThomsonFly Bournemouth, ThompsonFly Belfast, ThomsonFly/no location specified

Automation Response

The effects of the automation changes were to generate three different types of commands: (1) inappropriate thrust, (2) inappropriate pitch, or (3) autopilot disengagement. In seven cases, the autothrottle did not acquire the desired airspeed target (e.g. TA 1951). In four cases, the automation commanded an unexpected pitch-up that led to airspeed decay (e.g. BirgenAir). In one case, the automation terminated engagement and transferred control of the aircraft to the flight crew (AF 447).

Response Decision

There was no single flight crew response to the events that led to the CFIS accidents that would have been appropriate in all cases. There were a variety of correct responses. These responses can be categorized by the required change in flight crew actions and the appropriate degree of automation to use in the maneuver. In several accidents (XL Germany, Midwest 490, Provincial, ThomsonFly-Bournemouth, ThompsonFly - Belfast, and BirgenAir) the accident reports identify interventions in which the recommended procedure would have been to abort the maneuver being executed and transition to an alternate safe procedure. These interventions include aborting approaches by performing a Go Around (ThomsonFly - Bournemouth, ThomsonFly-Belfast) and terminating a climb or descent and level off (XL Germany, Midwest 490, Provincial, BirgenAir 301). In other cases, the accident reports suggest that it would have been possible to continue with the maneuver being executed with a manual over-ride of the auto-flight system (Colgan Air-Burlington, Colgan Air-Buffalo , Turkish Airlines 1951, United Express 629). In other cases (AAL 903), a manual mode/target selection followed by auto-flight system re-engagement would have been appropriate. In hindsight, the lack of airspeed information in two accidents (AF 447 and NWA 6231) predicated aborting the existing procedure and resorting to a “pitch-and-power” maneuver (i.e. wings level, pitch - 5 degrees up, power - 75% thrust).

The decision making required to identify the appropriate response frequently was not supported by the available automation cues. This decision must be based in part on the confidence that the flight crew has in the status of: 1) the aircraft structure and airfoils, 2) the aircraft sensors, 3) the control surface and propulsion systems, and 4) the automation. As the events leading to the CFIS accidents unfolded, the degree to which the automation was functioning, the status of the sensors, and the degree to which other aircraft systems may have been degraded would not have been obvious to the flight crew.

The execution of the appropriate intervention response was hindered in some cases by *moded input devices* that behaved differently under different circumstances. For example, in the aircraft involved in the TA 1951 accident, the throttle levers operate with two modes of operation. In the "airborne mode," the throttle setting can be manually over-ridden and will hold the manually set thrust setting. In the "land mode," the throttle setting can be manually over-ridden, but the thrust setting will automatically retard to idle unless a pilot holds the throttles in position. In the Asiana 214 accident, the pilot overrode the throttle setting by manually repositioning the throttles, but expected the autothrottle to advance to maintain airspeed. However, the act of repositioning the throttles resulted in the autothrottles entering a dormant mode. In these aircraft, the state of the autothrottle mode is not clearly annunciated on the flight deck.

General Discussion

Normal Accidents and Functional Complexity Failures

Studying a series of accidents across domains (i.e. nuclear power plants, aircraft, ships, petrochemical processing plants), Charles B. Perrow (1984), identified a phenomenon he labeled "Normal Accidents." These accidents were characterized by a failure that was the result of the interaction of functions with complex behaviors within a tightly coupled complex system. Perrow argues that in complex systems it is inevitable that the system occasionally will yield behavior that is inappropriate in certain circumstances and surprising to operators. For example, the system designers may not have considered particular combinations of input conditions that may occur in unusual circumstances. Alternately, the system itself may generate rare combinations of intermediate states that are not covered by the design. In these "functional complexity failures" there is no single point of failure. The automation behaves as it was designed, but the functional complexity results in a failure.

The CFIS accidents described here fit the characteristics of the "Normal Accident." In all of these cases, a structurally and mechanically sound aircraft was flown into an aerodynamic stall. Although a deceleration through $1.3V_{Stall}$ and then through V_{Stall} occurred in each accident, there is no pattern or consistent failure in the types of triggering events, in the effects of the triggering events on the automation, or in the inappropriate commands generated by the automation. The source of the failure is a complex interaction between factors.

To address functional complexity failures, one must examine the human-automation system as a whole. The concept of operations for the "flight deck system" is for the flight crew to delegate tasks to the automation and to supervise its performance. In the event that the automation generates an inappropriate command (e.g. throttles maintain idle thrust when the crew expects them to advance), the flight crew is expected to intervene. However, the ability of the flight crew to detect these rare events and to act appropriately is severely compromised by two different factors: 1) the knowledge required may not be present in the system and 2) the knowledge in the system is not properly communicated among the components.

Gaps in Flight Deck Knowledge to Respond to Functional Complexity Failures

Modern aircraft automation is inherently complex. It must be to deal with the complexity of the technology and operational environment. It is not feasible to train pilots to understand the *complete* behavior of the automation. The system itself is constructed by teams of engineers across a geographically dispersed supply chain, none of whom can be completely conversant with the behavior of the entire system. Hence, it is not surprising that pilots frequently do not fully understand their automation. Indeed, it would be impossible to provide pilots with a detailed knowledge of the automation. However, it may be possible to provide pilots with a functional knowledge of the automation. Current automation training is focused largely on learning procedures and not on developing a broad understanding of the automation. This leaves pilots with many gaps in their knowledge which they may plug with simplified behavioral models or misconceptions. In some cases, pilots may not even understand how commonly used procedures work and why potential alternative procedures would not.

But "automation education" by itself could not have prevented all of the CFIS incidents studied here. It is simply not reasonable to expect pilots to be able to learn all of the potentially useful information about their aircraft automation and to be able to recall it when needed. For example, in the Turkish Airlines 1951 accident, the pilots would have had to remember (without any cues) that one of the automation sub-systems (the auto throttle) relies on the Captain's radio altimeter alone to determine altitude and select the active mode. Typically, auto-flight functions

may be shifted from position to position (e.g. Captain's side to First Officer (F/O) side) so that when the FO is flying, most auto-flight systems rely on equipment on the FO's side -- but not in this case. Furthermore, the model of aircraft involved in the accident has two substantially different "RETARD" modes. While at altitude, "RETARD" will allow pilots to override the autothrottle by manually repositioning the throttle levers. However, when the aircraft is in the landing flare, "RETARD" will automatically reposition the throttles to idle. In addition, the Captain's radio altimeter on this particular aircraft had a history of maintenance issues, and hence might need to be monitored carefully.

In this case, the Captain's radio altimeter malfunctioned generating a value that showed that the aircraft had landed while it was still at 2000' above ground level. With the FO in command, the flight deck configured with the automation on the F/O's side, and the FO's radio altimeter functioning properly, the aircraft decelerated on the approach and the automation entered a "RETARD" mode as expected. However, because of the Captain's malfunctioning radio altimeter, the auto-flight system behaved as if it were in the *landing flare*, retarding the throttles to the idle position. When the FO pushed the throttles forward to arrest the deceleration, the automation returned them to idle. There is little doubt that had the pilots been briefed about these issues immediately prior to the flight, they would have remembered all of the relevant information during the approach. However, pilots are inundated by information from the Flight Crew Operating Manual (FCOM), Federal Aviation Regulations (FARs), flight training manuals, manufacturer's bulletins, "read-before-flight" memos, dispatch briefings, etc. In this context, it is unlikely that pilots would be able to retrieve the information relevant to a particular rare event months or years after they encountered it. The required knowledge is effectively not in the system.

Mitigation with Decision Support Tools

It is possible to provide some support. Sufficient automation education could be provided so that pilots would have substantial foundational knowledge and would know enough to ask the right questions. Then, computerized memory support tools could be used to provide the required information when needed. For example, when the aircraft serial number is entered before a flight, the on-board computer support tool could retrieve information about the aircraft maintenance history and provide implications of that history for the operation of the flight.

However, access to this knowledge is not sufficient. To determine what to do in any particular situation, pilots need to understand the current state of the aircraft. In the situations studied, this understanding was frequently lacking. Often, this occurred because the automation did not provide the necessary information. For example, in several cases, the automation disengaged when sensor input became unreliable. However, the same unreliable information was provided to the pilots who are in no better position than the automation to use this information.

In these situations, in very short order, pilots must determine why the automation disengaged. If it disengaged because sensor information became unreliable, the pilots must determine what information is reliable and what to do to regain control of the aircraft. Sometimes it is not clear what information is reliable and what information is not – this is especially difficult with sporadic failures and multiple displays that appear to reference independent sources but do not. In other cases, the information provided was ambiguous. Particularly problematic is information about the functioning of the auto-flight system. For example, the "RETARD" mode label described above refers to two substantially different functions. In general, the Flight Mode Annunciator (FMA) does not provide the information needed to properly supervise the aircraft automation and evidence suggests that pilots generally spend little time looking at FMA (Sarter et al., 2007; Björklund et al., 2006).

To properly supervise automation, pilots need to know: 1) what is controlled by each automation mode, 2) where each mode obtains data about the current state of the aircraft, 3) where each mode obtains targets, and 4) what actions each mode will take when the target is achieved. But having this knowledge is not sufficient. It merely provides a general framework. At every point during the flight, the framework must be populated with current information about the state of the aircraft and how it relates to the intended flight path. This requires that pilots: 1) know where to find the relevant information, 2) attend to these sources, 3) interpret the information correctly, and 4) integrate this information with their stored knowledge of the automated flight system and intended flight path. The requisite information is generally available somewhere on the flight deck, but it is not necessarily easy to find; it may be scattered between the PFD, MCP, ND, CDU, ECAM, and sometimes stand-alone thrust displays.

The cockpit of the modern airliner is a hybrid design. It incorporates the shell of an automated vehicle on top of that of an older manual aircraft design. Neither this design nor the training that accompanies it fully embraces the concept of the pilot as the supervisor of an integrated system. Instead, the pilots are alternately treated as passengers, data entry units, and barnstormers from an earlier era. Providing pilots with fundamental knowledge of the aircraft automation and clear information about the integrity of the sensor information and current status of the aircraft automation and its intentions would have averted the great majority of the CFIS accidents studied. However, this requires an appreciation of the role of the pilot in the modern commercial airliner and a willingness to make fundamental changes in our concepts of how aircraft automation should interact with pilots in this complex system.

Acknowledgements

This work was funded in part by NASA NRA NNX12AP14A and internal GMU Research Foundation funds. Thank you for technical suggestions from Immanuel Barshi, Michael Feary, Randy Bailey, Paul Krasa, Steve Jacklin, Houda Kourdali, Julia Trippe, George Dononhue, Akshay Belle, John Shortle, Mike Hieb, Paulo Costa, and Yong Tian.

References

Bayuk, A.J. (2008) Aviation Safety Management Systems as a Template For Aligning Safety with Business Strategy in Other Industries. American Society of Safety Engineers - The Business of Safety: A Matter of Success Symposium. Baltimore, Maryland, march 13-14, 2008.

Björklund, C., Alfredson, J., & Dekker, S. (2006). Mode Monitoring and Call-Outs: An Eye-Tracking Study of Two-Crew Automated Flight Deck Operations. International Journal of Aviation Psychology, 16, 257-269.

IATA (2013). IATA Safety Report. 49th Edition issued in April 2013. 800 Place Victoria, PO Box 113, Montréal, Quebec, H4Z 1M1.

ICAO (2013). 2013 Safety Report. International Civil Aviation Organization, 999 University Street, Montréal, Quebec, Canada, H3C 5H7.

Perrow, C. (1984). Normal Accidents. Basic Books, New York.

Sarter, N., Mumaw, R., & Wickens, D. (2007). Empirical Study Combining Behavioral and Eye-Tracking Data. Human Factors: The Journal of the Human Factors and Ergonomics Society, 49, 347.

Sherry, L. R. Mauro, I. Barshi & M. Feary (2014) Mitigating Controlled Flight in Stall Accidents. Internal Report, Center for Air Transportation Systems Research, George Mason University (CATSR-007-2013)

UNDERSTANDING AUTOMATION SURPRISE: ANALYSIS OF ASRS REPORTS

Julia Trippe & Robert Mauro
Decision Research
Eugene, Oregon

Pilots are frequently surprised by aircraft automation. These include cases in which the automation: 1) produces alerts to anomalies, 2) commands unexpected control manipulations (that may result in flight path deviations), or 3) simply disconnects. Aviation Safety Reporting System (ASRS) reports in which pilots indicated that automation produced unexpected actions were analyzed. Three general conclusions were drawn. First, many factors precipitate automation surprises. These include problems in: the auto-flight system and associated displays and interfaces, other aircraft sensors and systems, and interactions with weather and ATC. Second, inappropriate pilot actions are involved in a large proportion of these events. Third, recovery need not require reversion to manual control. There is no single general intervention that can prevent automation surprise or completely mitigate its effects. However, several different tacks (including improved training, displays, and coordination with ATC) taken together may be effective.

The capabilities of automated flight systems increased rapidly following the introduction of the electronic autopilot in the 1940's. In normal operations, the automated flight system of the modern airliner can now control nearly all functions required for flight. The effect of increased automation has been largely positive, greatly reducing errors due to pilot fatigue and allowing consistent precise navigation and performance. However, automation has given rise to new problems caused by faulty interactions between the pilot and the auto-flight system (AFS). This class of problems has been variously termed lack of mode awareness (Javaux and De Keyser, 1998), mode confusion (Degani, Shafto, & Kirlik, 1999), and automation surprise (Winter & Curry 1989; Woods, Sarter, and Billings, 1997, Burki-Cohen, 2010). In these cases, the flight crew expects the automation to command one behavior and is surprised when it commands another. When they do not jeopardize flight safety, automation surprises are a nuisance. But when the automation commands an aircraft trajectory that violates airspace or operational limitations, automation surprise becomes a critical problem (Reveley et al, 2010).

Automation surprise may result from undetected failures in aircraft sensors or other systems. Automation surprise also may result from pilots having an inadequate or mistaken "mental model" of the machine's behavior in the operational environment (Sarter and Woods, 1995). In addition, automation surprise may result from a problematic interface that does not provide adequate information about the status of the machine (Feary et al 1998; Norman, 1990, Degani, Shafto, and Kirlik, 1999).

Pilot lore is replete with complaints of flight management systems misbehaving. Flight management computers can appear to be pernicious allies that on occasion unilaterally decide to "drop" fixes, void altitude restrictions, or change modes of operation. For the most part the result of these events are relatively benign. No metal is bent; no one is injured; no lives are lost. But this is not always the case. Unexpected behaviors of the auto-flight system have been implicated in a number of recent fatal accidents (Sherry & Mauro, 2014). Furthermore, these events increase pilot workload, setting the stage for other errors. They create inefficiencies for the aircraft directly involved and may disrupt the flow of air traffic as controllers vector other aircraft to accommodate the aircraft whose crew is dealing with the unexpected behavior. To prevent or mitigate the effects of these "automation surprises" one must first understand why they occur.

People are surprised when they expect one event but another occurs. So, to understand automation surprise, one must ask why the behavior of the auto-flight system was not expected. Based on their training and experience, pilots build an understanding (a "mental model") of how their automation functions. Selected information about the current status of the aircraft, including its automation, is interpreted in the context of this mental model to build a mental representation of what the aircraft is currently doing and what it will do next. Hence, to be surprised, either the information fed into the mental model is inaccurate or the model itself is wrong. Pilots' expectations of what their automation will do may be in error when they attend to the wrong data, misinterpret data, or the data is in error. Alternately, their expectations may be wrong when their understanding of what the automation will do under the encountered conditions is wrong.

In this paper we examine pilot reports of unexpected automation behavior chronicled in the Aviation Safety Reporting System (ASRS) database in an effort to characterize the nature of these problems. Based on this understanding, technological and training strategies can be developed to prevent automation surprises and mitigate their effects.

Methods

The ASRS database was searched for automation-related event reports from 2012 by crews operating under Part 121. The initial search criteria were broadly specified to minimize the likelihood that relevant reports would be missed. Reports that mentioned automation, autopilot, auto throttle, flight management system, flight management computer, flight data computer, mode control panel, LNAV, VNAV, or any of the common abbreviations for these devices were retrieved. Of the 558 reports obtained, 234 described an event in which the pilots were surprised by unexpected actions of the auto-flight system.

The events that transpired before, during, and after the surprise were coded. Distinctions were made between five categories of events (actions or circumstances): precipitating, contributing, problem, detection, and response. *Precipitating* actions were those that preceded and led directly to the automation surprise. Within this category, we distinguished between primary precipitating or "catalytic" actions and secondary precipitating actions that occurred in response to the catalytic events. For example, in a number of cases, Air Traffic Control (ATC) instructions directed pilots to alter their previously programmed flight path. In entering flight path alterations into the flight management system (FMS), the pilots made an error that later resulted in a surprising aircraft behavior. We coded the ATC instructions as the primary or "catalytic" precipitating event and the pilot programming of the FMS as the secondary precipitating action. *Contributing* circumstances were those that did not directly precipitate the automation surprise but that may have contributed to the problem. For example, pilots may have reported being rushed or fatigued during the operation. *Problem* actions were those that produced the surprise. For example, in the prototypical ATC precipitated event described above, the pilots were often surprised by the aircraft veering away from the intended course. In this case, the problem event was coded as a course deviation. *Detection* actions were those that led to the discovery of the problem. These actions could involve direct observation of the aircraft behavior (as in the example above) by the pilots or ATC or observation of messages (e.g., Electronic Caution Alert Module (ECAM)), alerts (e.g., autopilot disconnect), or control movements. *Response* actions were those taken to resolve or recover from the surprise. These included actions such as taking manual control of the aircraft, switching to a lower level of automation (e.g., from VNAV to Mode Control Panel (MCP) control), or notifying ATC of a deviation. For each of these "action" categories, we coded the nature of the event, when the action occurred (phase of flight), who performed the action (e.g., ATC, crew, AFS), and the level of automation in use.

Results & Discussion

What Was Surprising?

A variety of different automation-related events surprised crews (see Table 1). In 15% (35) of the cases, the crew was surprised by changes in auto-flight system operation, including shutdown or freezing of various AFS components. In 11 of these cases, the autopilot disconnected. In three cases, the auto throttle disengaged or otherwise behaved unexpectedly. In 10 cases an unexpected mode change occurred and in 11 cases some component of the auto-flight system froze or failed.

In 12% (28) of the cases, the crew was surprised by the auto-flight system interface. Half (14) of these cases occurred when FMS data disappeared unexpectedly. In 73% (170) of the cases, the crew was surprised by aircraft behavior, including 21 cases in which the crew detected unexpected changes in aircraft control prior to a substantial change in aircraft position or velocity. However, in 64% (149) of all cases, the aircraft's velocity or position was altered substantially without the crew noticing. Twenty-seven of these resulted in airspeed changes, 35 in course alterations, 33 in altitude deviations and in 48 cases the aircraft's vertical path was affected unexpectedly.

Table 1.

Surprising Automation-Related Events

Event Type	Percent (n)	Event Type	Percent (n)
AFS operation only	15 (35)	**AFS problem affects aircraft control**	9 (21)
AFS Component Failure	4.7 (11)	**AFS problem affects aircraft behavior**	64 (149)
Auto Throttle	1.3 (3)	Airspeed	11.5 (27)
AP Disconnect	4.7 (11)	Altitude	14.1 (33)
Unexpected Mode Change	4.3 (10)	Course	15.0 (35)
AFS interface only	12 (28)	Localizer	2.6 (6)
Display	6.0 (14)	Vertical Path	20.5 (48)
FMS Drop	6.0 (14)	**Other**	0 (1)

Note. N=234; number of cases in parentheses.

When were crews surprised?

Overall, 55% of the automation surprises occurred during the arrival and approach phases of flight (see Table 2). By contrast, only 13% of the events occurred during cruise. However, this pattern differed according to the type of event. Failures or freezing of auto-flight system components, categorized as "AFS Component Failure" in Table 2, were evenly divided between climb, cruise, and approach. Two of the three auto-throttle events occurred in cruise. Although most of the autopilot disconnects, unexpected mode changes, and control anomalies occurred during arrival and approach, a substantial proportion occurred during climb and cruise. Display faults occurred equally during climb, cruise, and arrival. Waypoints dropping from FMS flight plans as well as lateral course deviations occurred mainly during climbs below 10,000 feet and arrival and approach phases.

Table 2.

Surprising Events by Phase of Flight

	Problem Phase of Flight										
Surprising Event	**Before Push**	**Take Off**	**Climb Below 10K**	**Climb Above 10K**	**Cruise**	**Descent Above 10K**	**Arrival**	**Approach**	**Go Around**	**Unknown**	**Total N**
AFS Component	0%	9%	18%	9%	27%	9%	0%	27%	0%	0%	11
Auto Throttle	0%	0%	0%	0%	67%	0%	0%	0%	33%	0%	3
AP Disconnect	0%	18%	0%	0%	27%	0%	36%	18%	0%	0%	11
Mode Change	0%	0%	30%	0%	20%	0%	40%	10%	0%	0%	10
Display	0%	14%	29%	0%	21%	0%	29%	7%	0%	0%	14
FMS Drop	7%	7%	29%	7%	0%	0%	7%	36%	0%	7%	14
Control	0%	0%	19%	10%	29%	5%	14%	19%	5%	0%	21
Airspeed	0%	0%	15%	15%	15%	0%	11%	37%	7%	0%	27
Altitude	0%	0%	0%	12%	3%	9%	36%	39%	0%	0%	33
Course	0%	0%	37%	6%	14%	0%	20%	17%	0%	6%	35
LOC	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	6
Vertical Path	0%	0%	4%	0%	2%	15%	69%	10%	0%	0%	48
Other	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	1
Total	0%	3%	15%	6%	13%	5%	31%	24%	2%	1%	234

There are several possible explanations for the disproportionate number of surprising events reported during the approach and arrival phases of flight. During these transitional phases crews are preparing for landing,

traffic is increasing, and ATC often places additional demands on pilots. To comply with these demands, pilots make heavy use of their automation, resulting in discovery of automation problems that lay dormant during previous phases of flight. Furthermore, pilots may make errors while changing modes and programming flight plans. The ASRS reports provide indirect evidence for these explanations. Problems that are likely to be caused by issues with electronic components (e.g., AFS and display problems) were more likely than other problems (e.g., course or altitude deviations) to occur during cruise. This follows, given that electrical failures are likely to be dependent on the amount of time spent in operation, whereas flight path deviations are likely to occur when the aircraft is in proximity to other aircraft, and thus likely to be directed by ATC to change course or altitude.

Proximal Precursors of Surprising Events

Narratives of the ASRS reports were analyzed to determine the sequence of events that preceded the automation surprises. Sometimes, the pilots described probable causal sequences. This was particularly likely when the pilots determined that their actions had led to the surprise. In other cases, analysts could reasonably infer an event sequence based on pilots' descriptions of their actions and the automation response in combination with knowledge of AFS operations. Problems attributed to malfunctioning automation components rarely contained sufficient information to verify this conclusion. In many of these cases, pilots reported contacting maintenance, but rarely reported ensuing findings.

A precipitating event could not be determined with sufficient confidence in 21 of the examined cases. Of the remaining 213 cases, 66% (140) involved human errors in auto-flight system operation (see Table 3). Pilot actions engendered the majority of airspeed (64%), altitude (84%), course (71%), localizer (60%), and vertical path (73%) surprises. Pilot actions also led to the majority of surprises resulting from auto-flight system problems, control manipulations, display problems, unexpected mode changes, and dropped waypoints. However, in 24% of cases the AFS or another technological system was apparently responsible for triggering the surprise. This includes two out of three auto throttle changes, the majority (80%) of the autopilot disconnect events, and a large proportion of the auto-flight problems (44%), control manipulations (44%), display problems (38%), and mode changes (33%).

Table 3.
Surprising Event by Source of Event

Surprising Event	Source of Event					
	Pilots	**Environment**	**AFS**	**Other System**	**Other**	**Total N**
AFS Problem	44%	0%	33%	11%	11%	9
Airspeed	64%	28%	8%	0%	0%	25
Altitude	84%	3%	6%	3%	3%	31
Auto Throttle	33%	0%	67%	0%	0%	3
Control	44%	11%	22%	22%	0%	18
Course	71%	0%	17%	6%	6%	35
AP Disconnect	20%	0%	20%	60%	0%	10
Display	54%	0%	23%	15%	8%	13
FMS Drop	100%	0%	0%	0%	0%	11
LOC	60%	40%	0%	0%	0%	5
Mode Change	56%	11%	33%	0%	0%	9
Vertical Path	73%	2%	18%	2%	5%	44
Total	66%	7%	16%	8%	3%	213

As noted above, 66% of the surprising events were precipitated by pilot actions. However, these actions were unabetted in only a small proportion (28%) of these cases. In the remaining 72% of these cases, pilot actions were triggered by external events. In 52% of the cases (73), pilots were attempting to comply with ATC instructions when they inadvertently triggered the unexpected automation response. In the remaining 20% of cases, pilots were attempting to cope with equipment issues when they inadvertently triggered an unexpected automation action.

Detection

In general (78% of the time), pilots were the first to detect the surprising events. However, ATC detected the problems simultaneously or before the pilots in 20% of the cases. In 36% of altitude deviation cases, ATC detected the deviation simultaneously (12%) or before (24%) the pilots. In 69% of course deviation cases, ATC detected the deviation simultaneously (6%) or before (63%) the pilots. In 12% of vertical path deviations, ATC detected the deviation before the pilots.

Resolution

The automation level at which the aircraft was being operated at the time of the surprising event was compared to the automation level during resolution of the event. In 34 cases the automation during one period or the other could not be determined with reasonable certainty. In 48% of the cases (95), the same level of automation was maintained throughout the reported event. In 45% of the cases (90) in which the aircraft was being flown under some level of automation, pilots resorted to manual control following the surprising event. In the remaining 55% of the cases (110) automation was used in the recovery. When VNAV was in use at the time of the surprising event (72 cases), pilots resolved the issues and continued under VNAV 32% of the time. In the remaining VNAV cases, 22% of the crews used MCP inputs to control the aircraft and 42% resorted to manual control. When LNAV was in use at the time of the surprising event (36 cases), the pilots continued to fly using LNAV 56% of the time. In 14% of the cases they relied on the MCP. In 22% of LNAV cases, pilots resorted to manual control. When the aircraft was being controlled using the MCP at the time of the event (28 cases), pilots continued to fly using the MCP in 64% of the cases and resorted to manual control 32% of the time.

General Discussion

Three important conclusions can be drawn from the results discussed above. First, many different factors may precipitate automation surprises. These include problems in the auto-flight system, problems in the displays and interface with the automation, problems in other aircraft sensors and systems, interactions with weather and other aspects of the external environment, and inappropriate actions taken by the pilots. Second, inappropriate actions by the pilots are involved in a large proportion of the automation surprise events. Third, recovery from automation surprises need not require reversion to manual control. In many cases, pilots continued to fly successfully using the same level of automation used prior to the automation surprise. Based on these observations, it is clear that there is no single general intervention that can prevent automation surprise or completely mitigate its effects. However, several different tacks taken together may be particularly effective.

First, new methods for pilot automation training need to be developed and tested. A large portion of the reported automation surprises can be traced to inappropriate pilot actions. In some cases, pilots understood what had happened after the fact. In other cases they did not, but probable causes of the surprises were apparent in their reports. Providing pilots with a better understanding of their automation would likely decrease the number of surprises. Producers of automated systems have long touted their ability to simplify pilots' tasks and improve precision and efficiency. However, researchers have repeatedly noted that while aviation automation has improved the efficiency and precision of operations, it has not reduced complexity. Indeed, automation has increased the complexity of the pilots' job. Training has not kept pace. Methods for automation education need to be developed which can help pilots develop an understanding of their automation that allows them to anticipate automation actions and not simply respond with a small set of canned procedures.[1] For pilots to construct adequate mental models of automation, they do not need to know the intricacies of the underlying engineering, but they must know how the system interacts with the environment – how it obtains information, what it controls, and what targets it is trying to achieve. Hence, pilots must be trained to understand: 1) what is controlled by each automation mode, 2) where each mode obtains data about the current state of the aircraft, 3) where each mode obtains targets, and 4) what actions each mode will take when the target is achieved. But having this knowledge is not sufficient. It merely provides the framework for the model. At every point during a flight, the model must be populated with current information about the state of the aircraft and how it relates to the intended flight path. This requires that pilots: 1)

[1] One area of particular difficulty appears to be the interaction between the auto-flight functions controlled through the mode control panel and those controlled by the flight control computer through a display unit.

know where to find the relevant information, 2) attend to these sources, 3) interpret the information correctly, and 4) integrate this information with their stored knowledge of the automated flight system and intended flight path.

Second, improved displays need to be developed that provide pilots with predictive indications. In many ASRS cases, the automation performed as it had been programmed to perform. However, errors in the pilots' programming or other inappropriate actions led to a discrepancy between what the pilots thought the system was programmed to do and what it was actually programmed to do. Typical automation interfaces do not provide clear displays of the programmed and predicted flight paths. Without this support, errors that humans inevitably make may go unnoticed until the aircraft is substantially off course, altitude, or airspeed.

Third, a large portion of pilot precipitated automation surprise events were themselves caused by instructions from ATC that proved problematic for pilots. A substantial decrease in the number of automation surprise events likely could be attained through restructuring ATC arrival and approach procedures. Decreasing the number of unnecessary "mission surprises" with which pilots must cope is likely to decrease the number of automation surprises. For example, new RNAV arrival procedures may be so complex that they cannot be reliably flown manually. However, ATC procedures allow controllers to vector aircraft into these procedures and to alter their components. Because these approaches effectively must be flown by the automation, modifications force pilots to program the FMS while flying the procedure. In this process, errors may be made that surprise the pilots and disrupt the flow of traffic. Modifying ATC procedures could substantially decrease the number of these problems.

The results reported here also underscore the importance of understanding and developing strategies for addressing the problem of automation surprise before NextGen becomes fully operational. In a large proportion of the cases examined, ATC called pilots' attention to a deviation from the planned course or altitude. Frequently, ATC handled the problem by providing a new clearance. Under NextGen, aircraft may fly in close proximity along defined 4D paths. Deviations such as those observed here would bring aircraft dangerously close to one another. At best, these events would cause substantial disruption to the traffic flow. At worst, they could result in collisions.

Acknowledgements

This work was funded by NASA NRA NNX12AP14A. Special thanks to Lance Sherry, Immanuel Barshi, and Michael Feary for technical suggestions.

References

Burki-Cohen, J. (2010). Technical Challenges of Upset Recovery Training: Simulating the Element of Surprise. In *Proceedings of the AIAA Guidance, Navigation, & Control Conference*: Toronto, CA.

Degani, A., Shafto, M., & Kirlik, A. (1999). Modes in human-machine Systems: Constructs, representation, and classification. *International Journal of Aviation Psychology*, *9*(2), 125-138.

Feary, M., McCrobie, D., Alkin, M., Sherry, L., Polson, P., Palmer, E., & McQuinn, N. (1998). Aiding Vertical Guidance Understanding. In *NASA Technical Memorandum NASA/TM- 1998-112217*, Ames Research Center, Moffett Field, CA.

Javaux, D., & De Keyser, V. (1998). The Cognitive Complexity of Pilot-Mode Interaction: A Possible Explanation of Sarter and Woods' Classical Result. In *Proceeding of the International Conference on Human-Computer Interaction in Aeronautics* (pp. 49-54). Montreal, Quebec: Ecole Polytechnique de Montreal.

Norman, D. A. (1990). The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation.' *Philosophical Transactions of the Royal Society B: Biological Sciences*, *327*(1241), 585-593.

Reveley, M., Briggs, J., Evans, J., Sandifer, C., & Jones, S. (2010). Causal Factors and Adverse Conditions of Aviation Accidents and Incidents Related to Integrated Resilient Aircraft Control. In NASA TM-2010-216261.

Sarter, N. B., & Woods, D. D. (1995). How in the world did I ever get into that mode? Mode error and awareness in supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5-19.

Sherry, L., & Mauro, R. (2014). Controlled Flight Into Stall (CFIS): Functional Complexity Failures and Automation Surprises. In *Integrated Communications Navigation and Surveillance Conference*.

Woods, D. & Sarter, N. (2000). Learning from Automation Surprises and "Going Sour" Accidents. In Sarter, N. & Amalberti, R. (Eds.) *Cognitive Engineering in the Aviation Domain*. LEA: Mahwah, NJ.

FLIGHT DECK INTERVAL MANAGEMENT AVIONICS: EYE-TRACKING ANALYSIS

Kara Latorella
NASA Langley Research Center
Hampton, VA 23681
John W. Harden
Old Dominion University
Norfolk, VA 23508

Interval Management (IM) is one NexGen method for achieving airspace efficiencies. In order to initiate IM procedures, Air Traffic Control provides an IM clearance to the IM aircraft's pilots that indicates an intended spacing from another aircraft (the target to follow – or TTF) and the point at which this should be achieved. Pilots enter the clearance in the flight deck IM (FIM) system; and once the TTF's Automatic Dependent Surveillance-Broadcast signal is available, the FIM algorithm generates target speeds to meet that IM goal. This study examined four Avionics Conditions (defined by the instrumentation and location presenting FIM information) and three Notification Methods (defined by the visual and aural alerts that notified pilots to IM-related events). Current commercial pilots flew descents into Dallas/Fort-Worth in a high-fidelity commercial flight deck simulation environment with realistic traffic and communications. All 12 crews experienced each Avionics Condition, where order was counterbalanced over crews. Each crew used only one of the three Notification Methods. This paper presents results from eye tracking data collected from both pilots, including: normalized number of samples falling within FIM displays, normalized heads-up time, noticing time, dwell time on first FIM display look after a new speed, a workload-related metric, and a measure comparing the scan paths of pilot flying and pilot monitoring; and discusses these in the context of other objective (vertical and speed profile deviations, response time to dial in commanded speeds, out-of-speed-conformance and reminder indications) and subjective measures (workload, situation awareness, usability, and operational acceptability).

Background

Interval Management (IM) is one NexGen method for achieving airspace efficiencies. In order to initiate IM procedures, Air Traffic Control provides an IM clearance to the IM aircraft's pilots that indicates an intended spacing from another aircraft (the target to follow – or TTF) and the point at which this should be achieved. Pilots enter the clearance in the flight deck IM (FIM) system; and once the TTF's Automatic Dependent Surveillance-Broadcast (ADS-B) signal is available, the FIM algorithm generates target speeds to meet that IM goal. The algorithm generating these speeds [1] is based on the standard terminal arrival route (STAR) in use, conforms to standard speed constraints in the terminal environment, and is adaptive to forecasted winds. When conducting FIM operations, in accordance with the concept of operations for NASA's technology demonstration efforts [2], the crew operates with autothrottles on, with autopilot engaged, and the auto-flight system in Vertical Navigation (VNAV) and Lateral Navigation (LNAV). In the tested concept of operations, the IM speeds are presented in the flightdeck, and the crew is responsible for selecting the new speed in the Speed Window of the Mode Control Panel (MCP), instructing the aircraft to achieve this new speed. To support FIM, the crew is responsible for safely flying the aircraft while maintaining situation awareness of their ability to follow FIM speed commands and to achieve the FIM spacing goal.

The objective of this investigation was to assess different FIM Avionics configurations based on objective data (flightpath and speed profile deviations, and response times), subjective assessments and ratings, and eye-tracking data. This paper discusses other results, but focuses on the eye-tracking metrics of performance used to characterize crew performance.

Methods

Participants

Twelve crews participated in the study, each with two experienced (between 19 and 40 years of flying, mean of 28.9 years) commercial pilots who were type rated in the same class as the simulated aircraft. Eleven of these subjects reported demographic data.

Apparatus & Scenarios

The study was conducted in NASA Langley's Integration Flightdeck (IFD) simulator, which approximated a Boeing 757 aircraft. The standard flightdeck was augmented with two Electronic Flight Bags (EFBs), and two ADS-B guidance displays (AGDs). Figure 1 shows these displays (in the aft position for the EFB) for the left side; where the position was mirrored for the right side. Scenarios required crews to fly from approximately 25,000 feet to land at Dallas-Fort Worth International Airport (KDFW). Scenarios began in level flight prior to top of descent, in VNAV Path autoflight mode engaged. The aircraft were in an unconstrained vertical path descent from at or near Top of Descent until reaching the first altitude constraint at 11,000 feet. The aircraft operated in VNAV Speed with the MCP speed window open, until the flaps were extended, and the autoflight mode reverted to VNAV Path. The last speed target given was the reference speed for flaps at 30, plus five knots to enable stabilized approach by 1000 feet above ground level. Scenarios concluded typically after roll out on touchdown, but occasionally in advance of that in order to save time (always after aircraft configuration for a stabilized approach was complete). Subjects were instructed to fly as they typically would, as though they had passengers in the back of the airplane, to respond to speed targets in a timely manner, to try to maintain speed conformance within seven knots, and to remain within 400 feet of the VNAV path. Confederate Air Traffic Controllers provided realistic communications to both the IFD and to roughly 20 other simulated aircraft in the environment. Prerecorded Automatic Terminal Information Service (ATIS) messages were available on the appropriate frequency.

Figure 1. The Integration Flightdeck Simulator, showing the EFB in the Aft position, and the AGD.

Experimental Conditions and Design

The Avionics Configurations tested were defined by an Avionics Condition (display devices and locations) and a Notification Method (whether events were indicated only visually, or were augmented with aural indications). Each crew evaluated four Avionics Conditions: (1) Integrated –FIM target speeds were presented in the upper left corner of the primary flight display (PFD) and speed profile deviation information was implicitly indicated as the deviation between current speed and an instantaneous speed profile bug on the PFD speed tape. The FIM page in the MCDU displayed numeric speed profile deviation (in knots). Significant deviations from the speed profile triggered a message on the EICAS system. (2) EFB-Aft –speed targets, speed deviation information and messages, and all elements of the IM clearance were presented on an EFB in the position shown in Figure 1. (3) EFB-Fore – all information was presented on the EFB, but this display was located in a more forward location, just under the outboard window. (4) EFB-Aft-AGD – in which the EFB-Aft condition was augmented with the ADS-B Guidance Display (AGD). The AGD repeats the same FIM target speed and speed deviation information given on the EFB.

Crews received notifications when conditions required their attention, *i.e.*, when a new FIM target speed occurred (*target speed onset*), if the current aircraft speed significantly deviated from the FIM target speed (*conformance deviation*), and if they failed to enter a new FIM target speed within a reasonable time period (*reminder*). A conformance deviation indicator was provided when the aircraft current speed was more than seven knots different from the instantaneous speed on the FIM speed profile, the speed changed more than five seconds ago, and aircraft current speed was not converging to the FIM target speed. A reminder was provided if the crew did not dial in the correct FIM target speed within 10 seconds. If the speed was still not dialed in, the reminder indication was repeated at most two more times at 10 second intervals. This study evaluated three notification methods defined by the modality (V for visual, A for aural) associated with the triplet of implementations: *target speed onset*, *conformance deviation*, and *reminder* events. The VVV method provided only visual (V) cues for all three events. The AAA method augmented these visual indications with an aural (A) tone, again for all three events.

The VAV method included visual indications for all three events, and presented the tone only if pilots significantly deviated from the speed profile.

Each crew member had the opportunity to fly an arrival and approach with each of the Avionics Conditions twice, once as pilot flying (PF) and once as pilot monitoring (PM). Avionics condition and Crew Role were within-crew variables. Notification Method was a between-crew variable. Order of Avionics conditions were counterbalanced over crews, and the assignment of scenarios to Avionics conditions was also counterbalanced.

Data Collection & Analysis

The study collected objective (vertical and speed profile deviations, response time to dial in commanded speeds, out-of-speed-conformance, and reminder indications) and subjective ratings (workload, situation awareness, usability, and operational acceptability) for each run. Post-experiment questionnaire items asked subjects to consider pairwise preference comparisons and to also rate (using 9-point scales with anchoring cues) the operational acceptability of the Avionics conditions in the context of the notification method they received, the utility of aural indications, and factors associated with operational acceptability (workload, situation awareness, and crew coordination).

Oculometer data was collected using two 6-camera Smarteye (SE) eye-trackers (SE Pro software, version 5.8) and recorded in Smarteye logfiles at 60Hz, corresponding to a nominal frame rate of about 17 msec. Oculometer data was also sent to simulation files, which were recorded at 5Hz. This experiment resulted in 192 eye-tracker logfiles (12 crews x 2 pilots/crew x 8 runs/crew). Complete logfile data was available for 168 (87.5%) of the data. The majority of missing data pertained to the first speed target, after pruning these from consideration in all datafiles, only seven logfiles (approximately 3.6%) remained affected by significant datafile loss.

In addition to incomplete data files, recorded data may be of questionable quality. SE software reports a head and gaze quality value for each reported point of gaze (POG), defined by the system's confidence in head and eye position assessment, normalized over the data previously acquired in that session. SE's Gaze Direction Quality metric ranges from 0.0 to 1.0; where 0 corresponds to the 1st percentile of all quality values experienced to that point, and 1.0 corresponds to the 99th percentile. As such, this value is individual-dependent and only useful as a general guide to the degree to which the eye-tracker has sufficient information upon which to base a POG determination. SE recommends that the system be given some time to "fill up" the buffer for this measure so that its reported values stabilize. Therefore, removing data associated with the beginning of the runs had the added benefit of stabilizing the quality measures. Unless specified otherwise, the following analyses were conducted on only those POG data that were associated with a gaze direction quality of 0.7 or greater. Regretably, data loss and insufficient data quality can not be considered random errors. Situations in which pilots gaze was extreme (downward or to the side) was more likely to result in lost or poor quality data. As such, data from the EFB_Aft condition was disproportionally affected.

Generalized linear models, with compound symmetry covariance structures (assuming heterogeneous variances and constant correlations among repeated measures) were used to model this mixed factor study with repeated measures. These models employed robust estimation of variances (to handle violations of model assumptions) and Satterthwaite adjusted degrees of freedom (to mitigate issues associated with missing data). Statistics were calculated with respect to Gamma distributions using a log link function, as most data were defined by non-negative values, and all distributions were positively skewed. Models included terms for main effects associated with Avionics Condition (EFB_Aft, EFB_Forward, EFB_Aft+AGD, Integrated) and Notification Method (VVV, VAV, AAA); and the two-way interactions of these main effects. For some measures, each pilot provided data (e.g., Noticing Time); whereas for others, the crew served as the experimental unit (e.g., Minimum Noticing Time). When the experimental unit was a pilot, the Role (Pilot Flying (PF) or Pilot Monitoring (PM)) and interactions of Role with Avionics Condition and with Notification Method were included in analyses. Significant fixed effects were further investigated with Sidak-adjusted sequential pairwise comparisons; which protect for inflated alpha, and are more powerful than Bonferroni-adjusted tests. Results were interpreted at alpha=0.10, but p-values are provided for the reader who choses to consider more stringent criteria.

Oculometer Results

Oculometer data was taken to help characterize the attentional sampling pilots used in response to the different Avionics Configurations for presenting FIM information. These metrics included those that addressed: the frequency with which pilots sampled the FIM display(s); the degree to which pilots' points-of-gaze were "Heads-Up," that is, looking out the window; and the time for pilots to notice IM events on the FIM displays. In addition, eye-tracker data was used to analyze metrics related to workload: the length of time pilots dwelled on the FIM

display on first regard following an IM event, and an entropy-based measure that has been proposed to be related to workload. The selection of data appropriate for consideration of each measure is presented per section.

Sampling the FIM Display

FIM Display Sampling analyses were conducted on the portion of each scenario from 99 seconds into the run, until 19 seconds following the eighth speed target encountered. This period was determined by attempting to maximize data used, minimize disproportionate lost data, include an equivalent number of speed target changes, and attempt to have roughly equivalent task durations. While the number of data points taken in these windows differed only slightly, counts were normalized by data frames per scenario. Results show only the Avionics Condition significantly predicted differences in counts of POGs on FIM displays ($p<0.001$). On average, pilots were most likely to sample the Integrated FIM Display; of the retrofit conditions, more likely to sample the FIM Display(s) associated with the EFB_Fore, then EFB_Aft+AGD, and least likely to sample the FIM Display associated with the EFB_Aft condition (all pairwise comparisons, $p<0.032$). Notification Method was not significant.

Heads-Up Sampling

The set of data used in this assessment was defined in the same manner as for the FIM Display Sampling analysis. However, whereas the FIM Display analysis used only data in which gaze direction quality was sufficient, this analysis employs a technique developed at NASA Langley [3] to define Heads-Up gazes from head pitch data when gaze quality is questionable. Avionics Conditions significantly affected Heads-Up POGs ($p=0.005$). Pairwise tests show only one significant comparison; this indicating that pilots experienced significantly more Heads-Up POGs in the Integrated Condition than for the EFB_Aft Condition ($p=0.016$), where all other comparisons were not significant (all $p>0.106$). The Avionics Condition and Notification Method interaction term was significant ($p=0.011$), but pairwise comparisons did not reach significance (all $p>0.438$).

Noticing Times

This analysis addresses the time (Noticing Time) for the PF and PM, separately, to first attend to the display containing information about FIM speeds. Appropriate display(s) are defined by Avionics Condition, as previously described. For the EFB_Aft+AGD condition, the first Noticing Time was identified as a POG on either the EFB or the AGD. Noticing times were identified in logfile data, and conducted on periods following each of eight speed targets per run. These were defined as the first POG that landed on the appropriate display(s) for which the data quality was 0.7 or greater, and for which there was a second such subsequent gaze, with fewer than five frames (nominally 88msec of data) of intervening missing or poor quality data in the same display. Noticing Times were significantly affected by Avionics Conditions ($p < 0.001$), the interaction of Avionics Condition and Notification Method ($p=0.001$), and Role (PF v. PM) ($p=0.077$). Noticing Times, and the variability in these, tended to decrease across conditions in this order: EFB_Aft, EFB_Fore, EFB_Aft+AGD, Integrated. Noticing time with the Integrated condition was, on average, over five times faster than with the EFB_Aft condition. The Integrated condition was significantly faster than all other conditions, and the EFB_Aft+AGD condition was significantly faster than both the EFB_Fore and the EFB_Aft conditions (all pairwise, $p< 0.013$). This main effect contains a significant interaction of the Avionics Condition and the Notification Method which shows that, for the EFB_Fore condition, Noticing time for the AAA condition was significantly longer than for the VAV condition ($p=0.084$). For the other three Avionics Conditions, pairwise comparisons of Notification Methods did not significantly differ (all $p\geq0.250$), but means suggest that pilots with the VVV method were slowest to notice new speed targets. PFs were faster to notice commanded speed changes than PMs, by about 200msec.

The same data set was similarly analyzed to investigate how the crews' Minimum Noticing Time, and the absolute difference of pilots' Noticing Times were affected by experimental conditions. Factors of significance for these variables are similar to findings observed for each pilot's Noticing Times. Avionics Condition ($p<0.001$), and the interaction of Avionics Condition and Notification Method ($p<0.001$) significantly affected the crews' first notice of a new speed target. With regard to the main effect, means followed the same order as for Noticing Times, but were more sensitive to differences in conditions. Pairwise comparisons showed only significantly faster Minimum Noticing Times for the Integrated condition than other conditions (all $p<0.065$). Pairwise tests of interaction terms show that the AAA Method was associated with significantly faster Minimum Noticing Times than the VAV Method (with the EFB_Aft Condition, $p=0.086$) and the VVV Method (with the Integrated Condition, $p=0.006$).

For the same data periods used to assess pilot and crew Noticing Times above, when a Notice was detected, a count was kept for how many times the PM was the pilot to notice first. Analysis as a Poisson distribution with a

log link function shows only a significant effect of Notification Method, whereby PMs were more likely to be first to notice new commanded speeds than PFs when using the VVV method than the AAA method.

Indicators of Pilot Workload

Two measures postulated to reflect workload are examined: Dwell Time and a measure related to scan path entropy. Initial Dwell Time on a display has been associated with the difficulty of processing visual stimuli and therefore extracting meaning from it [4], and has been associated with pilot workload [5]. Dwell Time was calculated from the data frame of first Notice (a POG on an appropriate FIM display) until either the frame before a POG on another display was reported, or five frames of missing or poor quality data occurred; then this number of frames was multiplied by the nominal frame rate. Data was framed by the eight speed targets encountered starting with the second of these and, as for Noticing Times, used logfile data. Dwell times were significantly affected by the Avionics Condition factor ($p<0.001$). The Integrated and EFB_Aft+AGD conditions did not significantly differ from each other, but they both supported significantly shorter Dwell Times than either the EFB_Fore or EFB_Aft conditions (and these last two did not significantly differ from each other) (all pairwise, $p <0.002$).

In theory, more information-dense, confusing, or unintitive presentations should require longer dwell times to detect and extract pertinent information. Based on initial work by [6], this measure was applied to characterize distribution of POGs [7]. These authors and others [8] found that as pilots' workload increased, visual sampling became more systematic and entropy decreased. The Nearest-Neighbor Index (NNI) measure of entropy has been found to be consistent with both objective (p300 EEG responses) and subjective (NASA-TLX score) measures of workload [9]. The NNI metric investigated here, is the ratio of the average observed minimum distances among POGs, and the mean distance expected if the distribution were random. The NNI is therefore equal to one when the distribution is completely random, and higher values suggest more systematic search – presumably induced by higher workload conditions [10]. Total entropy measures were calculated from software developed for this purpose at NASA Langley [11], based on Di Nocera's publications [9,10]. The 5 Hz eye-tracker data was used for this analysis due to processing complexity. Data included in this analysis was from the eight speed targets beginning with the second speed target occurring in logfiles for each run. NNIs were calculated for good quality data following the occurrence of a new speed target, and for the following 19 seconds. Notification Method ($p=0.099$) and Role ($p=0.024$) significantly affected NNIs. While pairwise comparisons on Notification Methods failed to reach significance (all $p\geq0.132$), observation of means shows clearly higher entropy (higher workload) when pilots had VVV notifications than other Notification Methods. NNIs were higher for pilots when in the PM role.

Discussion

The Integrated condition supported better heads up time than the EFB_Aft condition, fastest noticing times than all other conditions, shorter first dwell times than both the EFB_Aft and EFB_Fore conditions, and was sampled most frequently. The finding that this Avionics Condition was sampled more frequently than others is not surprising. The Integrated condition presented FIM information on the PFD, and obviously other information on this display is crucial to flight operations. Regrettably, the eye-tracker data did not provide sufficient resolution to distinguish between POGs to FIM information vs. other PFD content. However, in concert with other findings, this result indicates this condition most effectively supports FIM operations with minimal disruption to scan. Subjective ratings of Situation Awareness, distraction, and pairwise preference comparisons as reported elsewhere [12] are consistent with this finding.

Subjective commentary and ratings were least complimentary of the EFB_Aft condition, and eye-tracker findings are again consistent. When in this position, FIM information was sampled least frequently (based on means, though not significantly different from the other retrofit solutions), was slowest to notice (not significantly different than the EFB_Fore condition – but this was hampered inordinately when paired with the AAA Notification Method), and was one of the conditions that caused longer initial looks on the FIM display to extract information (where the EFB_Fore condition did not statistically differ). While the deleterious impacts of the EFB_Aft condition may not be surprising to this community, this study assessed it because the EFB has been implemented in this position in some cases. The aforementioned results, and those that show that both the EFB_Aft+AGD had faster noticing times and shorter dwell times than the EFB_Fore, seem to indicate superiority of the EFB_Aft+AGD over the EFB_Fore condition. However, other results show the reverse order – FIM information was sampled more frequently in the EFB_Fore condition, and other results based on pilots' awarness of new speeds and overall acceptability ratings were higher.

Most of the significant results associated with these analyses pertained to differences among the Avionics Conditions – that is, the placement and type of display used to present the FIM information, rather than the type of

aural/visual Notification Method used. It is, however, important to consider this finding in light of the experimental design: whereas Avionics Condition was considered as a within-subject/crew variable, Notification Method was a between-subject/crew variable – and therefore was subject to greater noise in the data from individual differences across levels. Notification Method did not statistically affect differences in Sampling Frequency, Heads-Up Sampling, or Dwell Times; and had only interaction effects with Avionics Condition for pilot's Noticing Times (where the AAA method seemed to significantly delay noticing in the EFB_Fore condition), minimum crew Noticing Times (showing superiority of the AAA method over the VAV method for the EFB_Aft condition and over the VVV method for the Integrated condition), and weak effects on the NNI (where means indicate higher workload for the VVV condition).

When in the PF role, pilots were generally faster in regarding the FIM display after a speed change, and had more systematic scan patterns (higher NNIs). However, when aural indications were available for all FIM events (the AAA method), PMs were more likely to be the first to notice speed changes than PFs. While decreasing responsiveness to FIM events by some small degree may advantage FIM operations, integration of new technologies and procedures must consider the full context of performance and cohesive job design – and the disruption of PF scan may be more costly to overall operations than the benefit to FIM operations.

Acknowledgements

This work was conducted under collaborative sponsorship by the NASA Airspace Systems Program and the Federal Aviation Administration, Human Factors Division (ANG-C1). The authors gratefully acknowledge the support of Ms. Jan Spangler and Mr. John S. Barry of Northrup Grumman, under the AMA-TEAMS2 contract to NASA Langley for assistance with data reduction and extraction.

References

[1] Abbott, T.S. (2012). *An Overview of a Trajectory-Based Solution for En Route and Terminal Area Self-Spacing: Third Revision*. NASA/CR-2012-217786. NASA Langley Research Center: Hampton, VA.

[2] Baxley, B. T., Swenson, H. N., Prevot, T., & Callantine, T. J., (2012). NASA's ATM Technology Demo-1: Integrated Concept of Arrival Operations, *31st Digital Avionics Systems Conference*, Williamsburg,VA, Oct.

[3] Ellis, K.K.E., Arthur III, J.J., Latorella, K. A., Kramer, L., Shelton, K J., Norman, R.M. & Prinzel, L.J. (2012). Quantifying Pilot Visual Attention in Low Visibility Terminal Operations. In *Proceedings of the AHFE International Conference on Applied Human Factors and Ergonomics*: San Francisco, CA.

[4] Callan, D. J. (1998). Eye movement relationships to excessive performance error in aviation. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*: Santa Monica, CA.

[5] Becker, S.I. (2011). Determinants of Dwell Time in Visual Search: Similarity or Perceptual Difficulty? *PLoS ONE 6*(3): e17740. doi:10.1371/ journal.pone.0017740.

[6] Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Tech Journal 27(3): 379–423.

[7] Tole, J., Stephens, A., Vivaudou, M., Harris, R. & Ephrath, A. (1982). entropy, Instrument Scan, and Pilot Workload. In *Proceedings of the IEEE Conference on Systems, Man & Cybernetics*: Seattle, WA.

[8] Harris, Sr., R.L., Glover, B.J. & Spady, Jr. A. (1986). *Analytical Techniques of Pilot Scanning Behavior and Their Application.* NASA-TP-2525 19860018448. NASA Langley Research Center: Hampton, VA.

[9] Camilli, M., Terenzi, M., & DiNocera, F. (2007). Concurrent validity of an ocular measure of mental workload. In D. deWaard, G.R.J. Hockey, P.Nickel, and K.A. Brookhuis (Eds.), *Human Factors Issues in Complex System Performa*nce. Shaker Publishing: Maastricht, the Netherlands.

[10] Di Nocera, F. (2007). Cognitive Aspects and Behavioral Effects of Transitions Between Levels of Automation. *European Office of Aerospace Research & Development Report* (Contract Number FA8655-05-1-3021).

[11] Harden, J.W. & Latorella, K.A. (2014). *Software for the analysis of Smarteye Oculometer Data for Workload and Scan Path Comparions.* Internal Report.

[12] Latorella, K. (2015) *Avionics Configuration Assessment for Flightdeck Interval Management: A Comparison of Avionics and Notification Methods.* NASA-TM-in press.

PUPILLARY RESPONSE AS AN INDICATOR OF PROCESSING DEMANDS WITHIN A SUPERVISORY CONTROL SIMULATION ENVIRONMENT

Ciara Sibley
Joseph Coyne
Naval Research Laboratory
Washington, DC
Akshith Doddi
Georgia Institute of Technology
Atlanta, GA
Phillip Jasper
Clemson University
Clemson, SC

Current Unmanned Aerial Vehicle (UAV) operator task demands are highly variable and unbalanced across team members, resulting in sub-optimal operator utilization which leads to mishaps. This has driven the Department of Defense's desire for more flexible team structures and task allocation tools. Unobtrusive and continuous measures of operator state are needed to effectively allocate tasking to operators and prevent errors. Twenty participants completed two twenty minute supervisory control sessions where task load was manipulated by varying event frequency (e.g., information requests) and eye tracking data was collected. Pupillometry data revealed increased mean and maximum pupil sizes with increased task load and larger pupil size standard deviation in participants who performed poorly, compared to those who performed well. These results suggest that increased pupil size is indicative of increased processing demands and could be predictive of task performance within a complex environment where performance measures can be challenging to obtain.

An increasing number of military aviation missions are being performed by unmanned systems, reducing the risk to Warfighters while increasing mission capabilities. Ironically though, unmanned systems have high manpower costs associated with conducting operations, partially due to specialized operator roles and highly variable levels of tasking throughout missions. Numerous Department of Defense (DoD) roadmaps have been promoting reductions in UAV manning via increased automation and more effective tasking tools (DoD, 2013); in particular, the 2015 Navy S&T plan highlights the need for improvements in "task allocation/assignment, planning, and coordination and control for heterogeneous systems" (ONR, 2015).

Enhanced planning and task assignment tools necessitate knowledge of the unmanned system's state and availability as well as the operator's state and availability. Ensuring that an operator's workload is balanced is critical, given that extremes in workload have been demonstrated to cause reductions in performance and human error (Yerkes & Dodson, 1908; Kahneman, 1973). This paper presents how pupillometry data collected from remote eye tracking systems can help provide insight into an operator's mental state and possibly aid in future task allocation.

Research conducted over the last several decades has established that pupil size varies as a function of cognitive processing and that the magnitude of the dilation correlates with the amount of mental effort exerted (Kahneman, 1973; Beatty & Lucero-Wagoner, 2000; Andreassi, 2007). The vast majority of these studies, however have been conducted within highly controlled and simple environments where participants are asked to perform tasks such as digit sequence recall, mental arithmetic, or verbal processing (Klingner, Tversky, & Hanrahan, 2011; Johnson, Miller Singley, Peckham, Johnson, & Bunge, 2014). For the purposes of this initial study, the authors were interested in investigating whether pupillometry data collected in a realistic UAV supervisory control environment could serve as a continuous metric of user state and be predictive of task performance.

Method

Supervisory Control Operations User Testbed

Human subject data collection was conducted using the Supervisory Control Operations User Testbed (SCOUT) which enables the investigation of the impact of scenarios of varying levels of difficulty on UAV operator task performance. SCOUT was developed by the Naval Research Laboratory to represent the tasking that a future UAV supervisory controller will likely perform assuming advancements in automation. The tasking within SCOUT was developed through interaction with current UAV operators and involved an iterative design and feedback process. SCOUT contains a pre-mission planning phase as well as mission execution phase.

During planning, the operator must determine the best initial route for sending their three heterogeneous vehicles, with different speed and sensor range capabilities, to search for seven stationary targets with varying priority levels, deadlines, and location uncertainties. Once a vehicle arrives at a target search area, the search is automated such that a target is found once the vehicle's sensor comes within range of the target. After the planning phase, mission execution begins and the operator is responsible for responding to incoming information requests (via chat), performing flight parameter updates (i.e. changing altitude and speed), managing airspace (requesting access to restricted operating zones) and re-planning vehicle routes as either new targets or new intelligence on existing targets (i.e. more precise location information) becomes available. All events within SCOUT are pre-scripted to occur at specific times.

In order to promote participant motivation, SCOUT was designed to be game-like, where the user receives points for responding to chat messages and finding targets. Additionally, the user loses points if a vehicle enters a restricted operating zone without receiving authorization. SCOUT was designed to be used on two thirty inch monitors. Figure 1 shows SCOUT's left and right screens, where the left screen's primary functions involve communication, planning, and airspace management, and the right screen functions include monitoring and updating vehicle parameters. Lastly, SCOUT is integrated with the SmartEye Pro 6.1 eye tracking system, such that all simulation events, behavioral data and eye tracking data are synchronized and logged together.

Figure 1. SCOUT's left and right screens depicting three heterogeneous vehicles conducting a search mission.

Experimental Design

Twenty-three individuals voluntarily participated in this experiment. Three participants were excused from the study; two due to a lack of comprehension and one due to a simulation error. Analyses were conducted on the twenty remaining participants (7 women and 13 men) ranging in age from 18 to 48 years (M = 30, SD = 9.6).

Prior to data collection, participants received approximately thirty minutes of training, which included videos demonstrating how to perform tasks within SCOUT and interactive assessments to ensure comprehension. Following training, each participant engaged in two experimental sessions. The order in which participants conducted sessions was randomized to help prevent any order effects, where half the participants received Session A first, while the other half received Session B first. Table 1 illustrates the difficulty associated with each block, demonstrating that Block 1 was always of medium difficulty, followed by either an easy or hard block.

Each SCOUT session was pre-scripted and included four segments that always occurred in the following order: Planning, Block 1, Block 2, Block 3. Participants were given ten minutes to create an initial plan before the mission execution phase. Each mission execution block took approximately six minutes, for a total of approximately 18 minutes of mission execution per session. Each block had a different level of difficulty which was manipulated via the frequency of chat tasking and the frequency of new targets added (see Table 2). During the easy, medium and hard blocks, events were presented approximately every 75, 45 and 15 seconds, respectively. In particular, participants were tasked via chat message to update flight parameters, i.e. speed and altitude, on specific vehicles. This task required the operator to increase or decrease the current speed or altitude by a specific amount (e.g. “Decrease altitude of UH-28 by 117”). Performance on each of these requests was analyzed in terms of completion, reaction time and accuracy.

Table 1.
Sessions and difficulty blocks

	Difficulty Level		
	Block 1	**Block 2**	**Block 3**
Session A	Medium	Hard	Easy
Session B	Medium	Easy	Hard

Table 2.
Difficulty manipulations by block

Block Difficulty Level	Chat Task Frequency	New Targets Added
Easy	75 seconds	1
Medium	45 seconds	3
Hard	15 seconds	4

Results

Performance Results

The primary performance metric within difficulty blocks was success on flight parameter updates (i.e., altitude and speed updates). Success was measured by percent of tasks completed, reaction time, and percent error for the completed tasks within each block. A two-way (Session X Difficulty) repeated measures ANOVAs was conducted for each of the flight parameter update metrics (percent completed, reaction time, percent error). Table 3 shows the p-values associated with each two-way ANOVA test.

As shown in Table 3, there was a significant and large main effect of Difficulty $F(2,38) = 37.422$, $p = .000$, $\square_p^{\square} = 0.663$ for the percent of tasks completed. Additionally, there was a significant and medium main effect of Difficulty $F(2,38) = 4.095$, $p = 0.025$, $\square_p^{\square} = 0.177$ for the percent of error on completed tasks. See Figure 2 for a visual depiction of these effects.

No main effects were found for Session, however there was a significant and medium interaction effect between Difficulty and Session $F(2,38) = 3.288$, $p = .048$, $\square_p^{\square} = 0.148$ for the percent error, due to the fact that error slightly increased in the session two easy block and decreased in the hard and medium blocks. No differences were found in reaction times, indicating that participants responded at the same rate, both across sessions and difficulty blocks.

Table 3.
Significance values from two-way (Session X Difficulty) repeated measures ANOVA

	Performance Data ANOVA Significance Results		
	Percent Completed	**Reaction Time**	**Percent Error**
Difficulty	p = 0.000*	p = 0.860	p = 0.025*
Session	p = 0.335	p = 0.113	p = 0.883
Difficulty X Session	p = 0.074	p = 0.636	p = 0.048*
** Denotes significance at p < 0.05*			

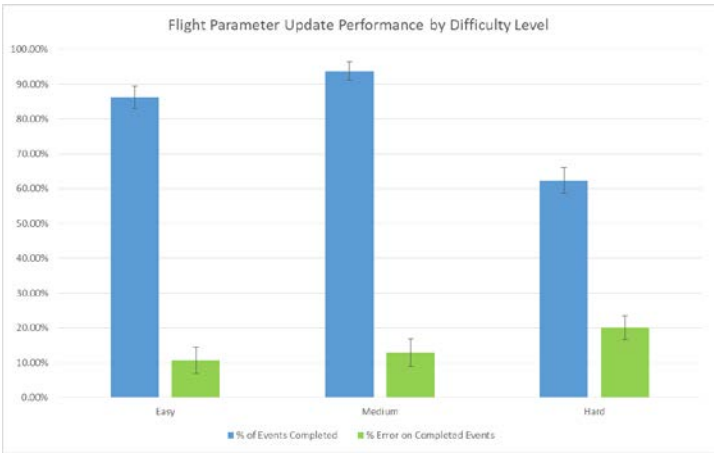


Figure 2. Performance on the Flight Parameter update task shown by block difficulty level

Pupillometry Results

Pre-processing was conducted prior to analyzing the pupillometry data. The first step involved filtering out all dropped and poor quality data, defined as having a value of zero, using SmartEye Pro's built-in quality measure. This initial step reduced the amount of data by 13%. Next, outliers were removed, which were defined as the top and bottom 1% of the entire twenty subject dataset. This ensured that erroneous data which were not caught by the quality measure (e.g. pupil measurements of 12 mm) were removed. Pupillometry data from both the left and right eye were consistent, but only data from the left eye are presented here.

Three pupillometry metrics were considered within this analysis: pupil size mean, pupil size standard deviation, and pupil size maximum. Each of these metrics was calculated for each individual, during each block, and within each session. A two-way (Session X Difficulty) repeated measures ANOVA was conducted on each of the three pupil size metrics (see Table 4). Analysis revealed a significant main effect of Difficulty for pupil mean $F(2,38) = 8.985$, $p = .001$, $\square_p^{\square} = 0.321$ and pupil maximum $F(2,38) = 3.577$, $p = .038$, $\square_p^{\square} = 0.158$.

Post hoc comparisons using a Least Significant Difference test indicated that the mean pupil size in the hard block (M = 3.60, SD = 0.55) was significantly greater than the mean pupil size in the easy block (M = 3.52, SD = 0.55). Additionally the maximum pupil size in the hard block (M = 5.84, SD = 0.25) was also greater than the maximum pupil size in the easy block (M = 5.73, SD = 0.41). Lastly, there was a significant main effect of Session $F(1,19) = 15.026$, $p = 0.001$, $\square_p^{\square} = 0.442$ for the pupil mean metric, where mean pupil size was significantly smaller in the second session than the first, which may reveal fatigue or learning. No interaction effects were found.

Table 4.
Significance values from two-way (Session X Difficulty) repeated measures ANOVA

	Pupillary Metrics ANOVA Significance Results		
	Pupil Mean	**Pupil Standard Deviation**	**Pupil Maximum**
Difficulty	p = 0.001*	p = 0.062	p = 0.038*
Session	p = 0.001*	p = 0.801	p = 0.586
Difficulty X Session	p = 0.301	p = 0.319	p = 0.301
** Denotes significance at p < 0.05*			

Pupillometry data were also collected immediately following a Situation Awareness (SA) probe that occurred within each block. During this time, known as the Regain SA phase, the screen and mission execution was paused so that the user could take as much time as needed to relax, regain situation awareness, and not have to attend or respond to any events before the block tasking resumed. Each of the three pupil metrics were computed during the easy, medium and hard block's Regain SA phases. A two-way (Session X Difficulty) ANOVA was conducted for each pupil metric (pupil mean, standard deviation and maximum) and revealed no significant main effects and no significant interaction effects; meaning no differences existed among the different Regain SA blocks.

Operator Performance and Pupillometry Data

Out of the twenty participants, the top four and bottom four performer's pupillometry data were analyzed to investigate whether those who performed well had different pupillary signatures than those who perform poorly. Performance was considered based upon a combination of overall mission score at the end of each session, and flight parameter update performance (i.e. high performance was defined as high percentage of events answered, low reaction times, and low error). No statistical analyses were conducted, given the low number of participants; however Figure 3 shows a comparison of pupil size standard deviations for the top and bottom performers in the easy, medium and hard blocks.

Visual inspection suggests that the bottom performers had higher variability in their pupil sizes, compared to the top performers. The minimum and maximum pupil sizes were assessed for each performer group to ensure there wasn't an inherent pupil size difference between the two groups. No difference existed: the maximum pupil size for the low performers was 5.966 and 5.961 for the high performers; the minimum pupil size was 1.997 for the low performers and 1.997 for the high performers. Mean pupil size and maximum pupil size were also compared between the two performance groups, but did not reveal anything conclusive.

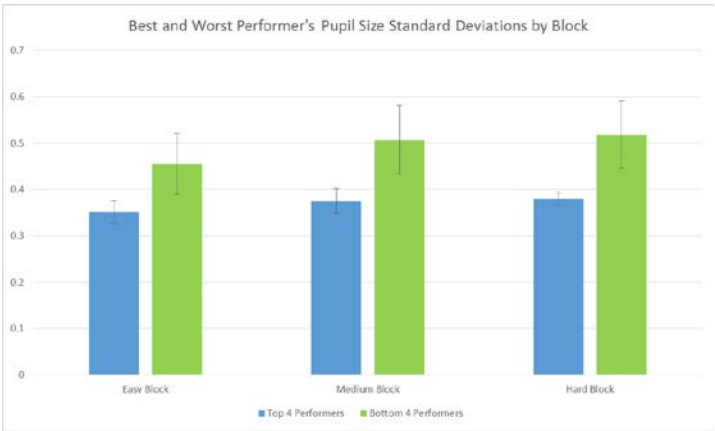


Figure 3. Pupil Size Standard Deviations for top and bottom four performers during easy, medium and hard difficulty blocks

Discussion

The results of this initial study are consistent with previously mentioned research conducted within simple task environments, and demonstrates that pupillometry data is also effective at discriminating between extremes in task load levels within a complex supervisory control environment. In particular, the mean and maximum pupil size metrics were significantly larger in high task load blocks of time compared to low task load blocks. Furthermore, the lack of difference among the metrics within the RegainSA phases, which took place in the middle of each difficulty block, provides additional evidence that pupillometry metrics are in fact indicative of a user's mental state (i.e. cognitive processing) and not some other factor such as fatigue.

While pupil size standard deviation was not statistically different among block levels at the aggregate level, it did appear to be trending towards significance at p=0.062. More interestingly, though, is the data in figure 3 which suggests that individuals who are struggling with tasking have greater variability in their pupil size. This finding requires further investigation but could be promising for potentially identifying operators who are overloaded and susceptible to making an error. None of the twenty participants were able to complete all tasking during either of the difficult blocks without making an error. Future analysis will involve a finer grained investigation into this performance data to assess whether it is possible to identify a signature in the pupillometry metrics indicating a user has become overloaded and is susceptible to making errors.

These results suggest that pupillometry data could be a useful metric for determining how to allocate tasking within an actual supervisory control environment, where operator state and task performance data is very rarely captured or available. Pupillometry data provides a continuous stream of information about an individual's mental state, which could be used to reveal when a user is overloaded. This information could be used to shed tasking either to another team member or automation before an error occurs. Future research will focus on building individual performance models and investigate the utility of embedding pupillometry data within a task allocation tool.

References

Andreassi, J.L. (2007). "Pupillary response and behavior," in *Psychophysiology: Human behavior and physiological response.* 5th ed (Mahwah, NJ: Lawrence Erlbaum Associates), 289-307.

Beatty, J., and Lucero-Wagoner, B. (2000). "The pupillary system," in Handbook of Psychophysiology, eds. J.T. Cacioppo, L.G. Tassinary & G.G. Berntson. 2 ed (Cambridge, UK: Cambridge University Press), 142-162.

Department of Defense (2013). "Unmanned Systems Integrated Roadmap FY2013-2038". (Washington, DC: Department of Defense). Retrieved from:http://www.defense.gov/pubs/DOD-USRM-2013.pdf

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Developmental Psychology*, *5*, 218. http://doi.org/10.3389/fpsyg.2014.00218

Kahneman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, 48(3), 323-332.

Office of Naval Research. (2015). Naval S&T Strategic Plan. Arlington, VA. Retrieved from: http://www.onr.navy.mil/About-ONR/~/media/Files/About-ONR/2015-Naval-Strategy-final-web.ashx

Yerkes, R.M., and Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. Journal of comparative neurology and psychology 18, 459-482.

THE ELECTROOCULOGRAM AND A NEW BLINK DETECTION ALGORITHM

SAMANTHA L. EPLING, Ball Aerospace and Technologies Corporation
MATT MIDDENDORF, Middendorf Scientific Services
MICHAEL HOEPF, Oak Ridge Institute for Science and Education
CHRISTINA GRUENWALD, Southwestern Ohio Council for Higher Education
LUCAS STORK, Southwestern Ohio Council for Higher Education
SCOTT GALSTER, Air Force Research Laboratory

Accurate and efficient real-time cognitive workload assessment has many important applications, and physiological monitoring has proven quite helpful with this assessment. One such physiological signal, the electrooculogram (EOG), can provide blink rate and blink duration measures. In a recent study, we developed and validated a robust blink detection algorithm based on the vertical EOG (VEOG). This algorithm does not require baseline data and is adaptive in the sense that it works for a wide variety of individuals without any experimenter adjustments. The performance of the algorithm is quantified using truth data based on video recordings. The algorithm produced blink rate and blink duration data for participants in a simulated remotely piloted aircraft experiment. Although this paper focuses on the blink detection algorithm, some results from the study will be included. Specifically, it was found that participants blinked fewer times and with a shorter duration in the more difficult experimental conditions.

A warfighter's dynamic cognitive workload is a ubiquitous concern, due in part to the close relationship between workload and performance (Cain, 2007). Methods of high workload detection and mitigation are therefore important areas of research. Cognitive workload is a complex and dynamic construct - it is dependent on many variables such as the operator's level of training, expertise, experience, motivation, time on task, and the difficulty of the task itself. The physiological measures used to assess workload are also affected by various factors such as fatigue, stress, engagement, and the environment (Wang & Zhou, 2013).

Physiological measures have proven quite useful in real-time cognitive workload assessment, in both laboratory and real-world settings. Wilson and Russell (2007) demonstrated that physiologically-determined adaptive automation could be used to significantly improve performance. Blink measures based on EOG can shed light on mental workload, and because of their noninvasive nature, they are well suited for real-time use.

Currently, despite the widespread use of eye blinks in research, a singular method of detection does not exist. Counting blinks from videotaped recordings is traditionally rejected due to the inordinate amount of time it takes. The use of cameras with complex eye tracking packages are an improvement technologically, but complexity and practicality remain a concern for field use. Blink detection using EOG has advantages over both of the approaches, and there is always a need to explore new detection algorithms.

The purpose of the current research is to present a blink detection algorithm that has a high accuracy level, is adaptive, dynamic, works in real-time, and does not require experimenter adjustments or calibration. The details of this algorithm and its validation using truth data will be presented. Results from a recent study employing this algorithm will be shown.

Background

Various eye metrics, including blinks, have been shown to be useful indicators of cognitive state (Wang & Zhou, 2013). Blinks are classified in three ways. Voluntary blinks are those that occur with a conscious decision to shortly close the eye. Involuntary blinks involve both reflexive (startle) blinks, which occur to protect the eye in reaction to an external impetus, and spontaneous (endogenous) blinks, which are also reflexive but serve to maintain corneal moisture (Andreassi, 2007). Any blink mentioned henceforth refers to a spontaneous blink.

According to Andreassi (2007), blinks in a relaxed state occur at an average of 15-20 times per minute, have average amplitude of 380μV, and average duration of 120ms. Blink rate varies widely within an individual depending on the type of task, environment in which it is performed, and information processing demands. There are also large differences between individuals. For example, Sforza and colleagues (2008) showed that women

spontaneously blink more frequently than men, and that younger people blink with more eyelid displacement than older people. Kong and Wilson (1998) claimed that such variability substantiates the need for blink detection algorithms, using the EOG signal, that are robust to noise, artifacts, and intra- and inter- individual variations.

Because blink rate and duration have been shown to relate to workload in environments with visual task demands (Wang & Zhou, 2013), the accurate detection of blinks holds great promise for continuous workload monitoring during many common human-computer interactions.

Blink Detection Algorithms

While a blink in the VEOG is visually distinct to the human expert, the varied parameters and noisy signal make blink detecting algorithms quite difficult to create. Kong and Wilson (1998) filtered the signal four different ways before processing it with their algorithm. Blinks were then determined by finding a negative peak followed by a positive peak within a specified time window, along with other features used to give each potential blink a composite score. Another method, the workload assessment monitoring (WAM) system, also detects blinks using an algorithm that finds consecutive negative and positive slopes within a specified range (Wilson, 1994). However, unlike Kong and Wilson's approach, WAM criteria must be adjusted for every participant.

In order to detect blinks in the VEOG, a reliable algorithm is essential. Without accurately detecting the blinks, the extracted features (blink rate and durations) are less useful for assessing cognitive workload. Many researchers in the human factors, psychophysiology, ophthalmology, and human computer interaction domains have attempted blink detection in different ways, for different purposes. However, most literature available on these various approaches lack specific detail. Therefore, the present research fills that gap by presenting an algorithm that performs with a high level of accuracy and has a well-documented methodology.

The New Blink Detection Algorithm

Blink characteristics. The basic shape of a blink in the VEOG signal has distinctive features. Andreassi (2007) describes the waveform as a sharp rise immediately followed by a sharp fall. The duration is short, the peak is rounded, and there is a noticeable overshoot before the signal returns to zero (Figure 1). Each time the VEOG signal goes above and below a threshold value is referred to as a bump. Data extraction software was written to compute a simple threshold, and use it to extract features from all bumps in the VEOG signal. The bumps in the signal include blinks and non-blinks (i.e., eye movements and noise). The features extracted by the software include the slope up at the midpoint, the slope down at the midpoint, the peak amplitude, and the duration at the midpoint.

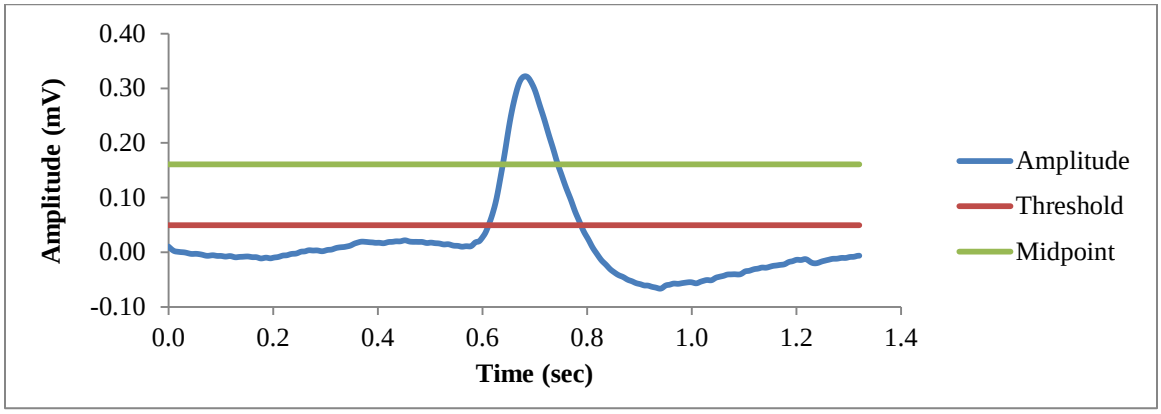


Figure 1. The basic shape of a blink.

Primary criteria. Not all excursions in the VEOG signal that go above and below threshold (i.e., bumps) are blinks, so criteria values needed to be established to distinguish blinks from non-blinks. To accomplish this, the data extraction software was used on a large database of existing VEOG data to extract the four features described in the above paragraph. Raters were trained to recognize the basic shape of a blink. The raters then visually observed each VEOG signal and coded each bump as a blink or non-blink. This data was used to determine eight criteria

values needed to develop the initial blink detection algorithm. These eight criteria are the minimum and maximum values for slope up, slope down, peak amplitude, and duration at the midpoint. The extracted values for blink amplitude and duration were sorted and plotted for visual inspection (see Figure 2). Histograms were also created and the data was found to be normally distributed. The data for the two slope features were also examined and found to be normally distributed. The initial values for the primary criteria were determined using the 98th percentile of the distributions. The initial values were slightly tweaked following some testing. The final values of the primary criteria used in the detection algorithm are shown in Table 1.

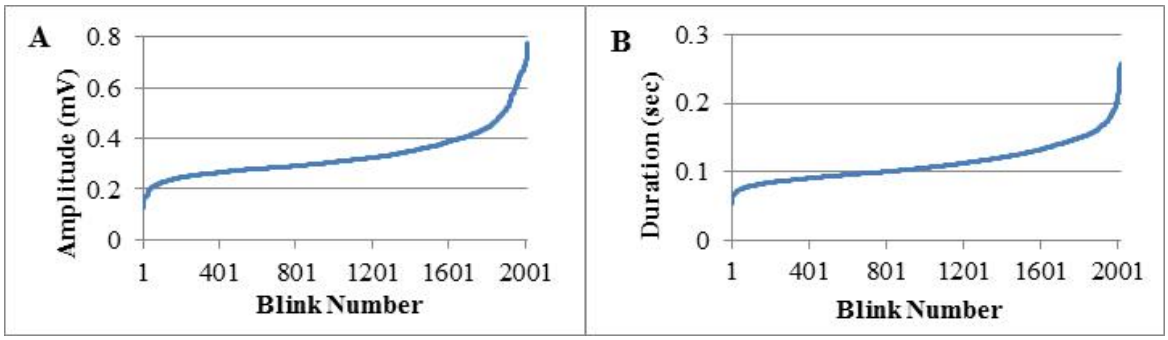


Figure 2. Blink amplitude (A) and duration (B) extracted from over 2000 blinks.

Table 1.
Criteria Values for the Primary Features

Blink Feature	Minimum Criteria Value	Maximum Criteria Value
Amplitude (mV)	0.1211	0.6483
Blink Duration at the Midpoint (s)	0.06	0.198
Slope Up at the Midpoint (mV/s)	1.5	13.41
Slope Down at the Midpoint (mV/s)	-10.0	-1.25

The existing database of VEOG data used above is from 12 participants performing four different tasks. This resulted in a total of 3102 bumps that went above and below threshold. The raters manually coded 2020 as blinks and 1082 as non-blinks. Note this is not absolute truth data because the raters were observing an electrical recording (VEOG) rather than a video recording. Therefore it is possible for a rater to occasionally miscode a bump. The use of video recording to generate actual truth data is discussed in the Algorithm Use and Validation section.

Secondary criteria. The initial blink detection algorithm was written to perform blink classification using the eight primary criteria identified above. New VEOG data was collected to test the classification logic. Each extracted feature from the new VEOG data had to fall within the range of the corresponding primary criteria values. For example, the amplitude must be between 0.1211 and 0.6483, otherwise the bump is not classified as a blink. The same logic is applied to the other three main features (slope up, slope down, and duration at the midpoint). For a bump to be classified as a blink, all four extracted features must be within their associated ranges. This classification logic was tested with additional new VEOG data and a few false positives were occurring.

To remedy this problem, five secondary features were extracted from each bump. In a manner similar to the primary features, criteria values were established for the secondary features. These features were used to provide a confidence assessment to refine classification accuracy. Specifically, all four of the primary features must fall within their associated ranges and three of five secondary features must meet their criteria values. Additional detail is provided in the scoring and classification section.

The five secondary features are the closure duration, the two R^2 values for linear fits at the midpoint and two additional duration measures (see Figure 3). The distance between the two linear fits at the peak is referred to as the closure duration. The distance between the two linear fits at the zero crossing is the blink duration at the zero crossing due to midpoint extrapolation. A similar duration is measured using linear fits about the threshold. The blink classification code was enhanced to incorporate the five secondary features. The number of false positives

produced by the algorithm was substantially reduced. The actual values of the secondary criteria used in the detection algorithm are shown in Table 2.

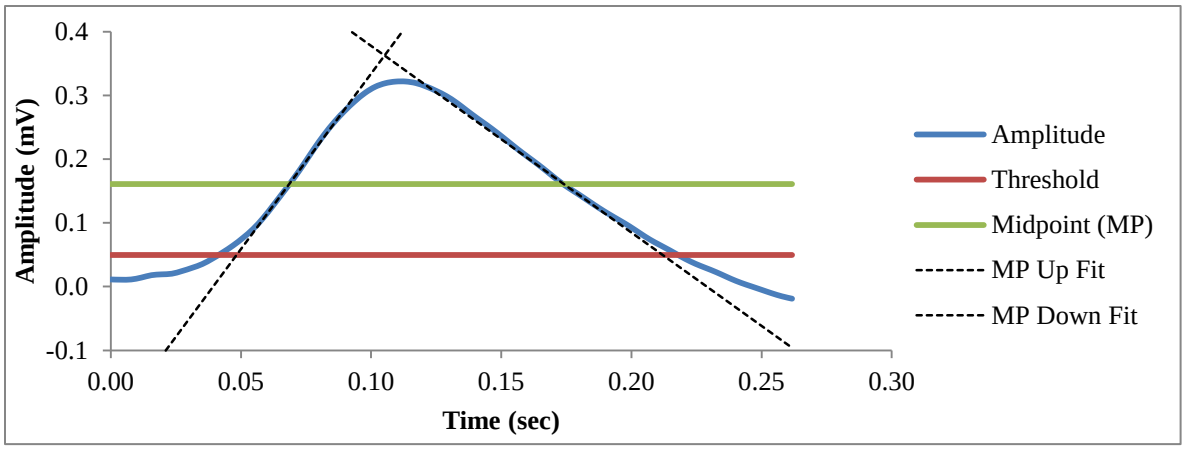


Figure 3. A typical blink with linear fits about the midpoints.

Table 2.
Criteria Values for the Secondary Features

Blink Feature	Minimum Criteria Value	Maximum Criteria Value
Closure Duration (s)	0.01	0.10
Slope Up at the Midpoint R^2	0.996	N/A
Slope Down at the Midpoint R^2	0.995	N/A
Blink Duration ZCMP (s)	0.1162	0.3
Blink Duration ZCT (s)	0.1	0.35

Note. ZCMP is duration at the zero crossing due to midpoint extrapolation. ZCT is duration at the zero crossing due to threshold extrapolation.

How the Algorithm Works

The major components of the blink detection algorithm are threshold generation, feature extraction state machine, scoring and classification, and blink save and false positive detection logic.

Threshold generation. The threshold generation approach uses a sliding five second window of raw VEOG data. To minimize the effects of blinks and eye movement on the threshold, the data is high pass filtered using a first order Butterworth filter with a break frequency of 10 Hz. This essentially leaves in the "noise" from which the threshold is calculated. The filtered signal is then rectified and the median is taken for the *raw* threshold value. The median is used because the data in the five second window can be highly skewed when there is a blink in the window.

The second stage of threshold generation imposes limits on the *raw* threshold and adds in a threshold reduction value to accommodate double and multiple blinks. Initially the threshold limits are static, but after ten blinks have been detected, the limits are dynamic based on the mean amplitude of the recorded blinks. The threshold reduction value is necessary due to the high pass filter in the signal acquisition hardware, which causes the signal to overshoot zero on the down slope of the blink. If the blink is immediately followed by another blink the subsequent blink starts below zero (Figure 4). If the threshold reduction value is not applied, the subsequent blink(s) may be easily missed. The threshold reduction value is based on the amount of overshoot of the previous blink. The threshold returns to its normal (non-reduced) value using a function that is the inverse of the high pass filter implemented in the signal acquisition hardware.

Feature extraction state machine. This state machine uses the threshold to monitor the VEOG signal. The state machine has four values (0, 1, 2, and 3). In state zero the logic waits for the signal to be below threshold. In state one it waits for the signal to go above threshold, at which time upward threshold crossing data is captured and

the threshold is frozen. In state two the logic is waiting for the signal to go back below threshold. During this time peak data and downward threshold crossing data are captured and the threshold is unfrozen. In state three the signal overshoot value is captured and the extracted features are scored to see if the signal excursion above and below threshold is a blink. The state machine then returns to state zero.

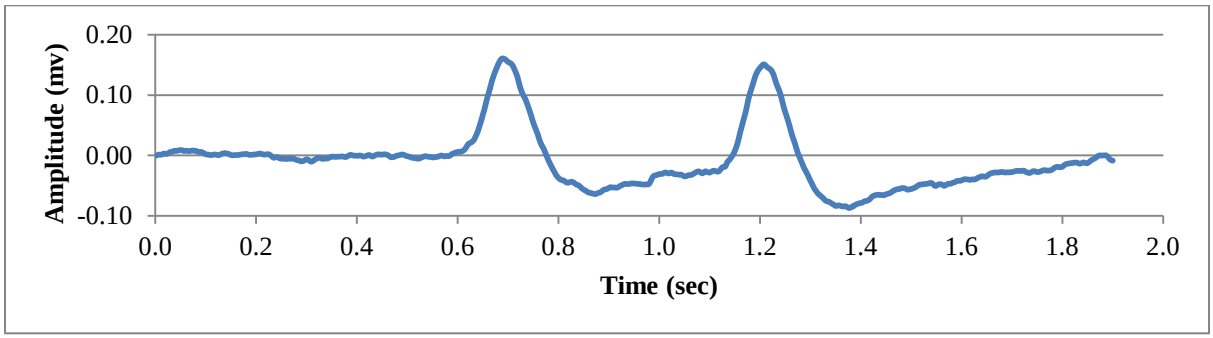


Figure 4. The threshold reduction logic is needed when multiple blinks occur in a short time frame. Because of the overshoot following a blink, the next blink starts from a lower value.

Scoring and classification. The VEOG bump that goes above and below threshold is scored using criteria values described in the previous sections. One point is awarded when each of the four primary criteria are met and one tenth of a point is awarded when each of the five secondary criteria are met. Therefore, the maximum score for a VEOG bump is 4.5 points. Bumps that score 4.3 points or higher are reliably classified as blinks. This requires that all four of the main features be met, and at least three of the secondary features be met. Requiring scores higher than 4.3 points results in some blinks being missed. Allowing scores lower than 4.3 results in some false positives.

Blink save and false detection logic. This logic applies to a very small number VEOG bumps. Bumps that fail only one of the four main criteria, but otherwise have a nearly perfect score (3.4 and 3.5), are given a second look. When a bump fails the maximum amplitude criterion, the criterion can be adjusted upward using amplitude data from previous blinks (minimum of 10 required). A likewise, adaptive test is applied when the minimum amplitude fails or the slope down at the midpoint fails. Currently only one false positive test is performed. Bumps that have two peaks are rejected.

Algorithm Use and Validation

Experimental results. In a recent study participants were asked to track targets using remotely piloted aircraft. Workload was experimentally manipulated and physiological measures were collected. VEOG data was processed using the blink detection algorithm discussed in this paper. For the sake of brevity, only the blink rate and duration results are discussed here. For a full discussion of the experiment, see Hoepf, Middendorf, Epling, and Galster, this volume. In the study, high workload had a statistically significant effect on blink rate and duration. Blink rate was slower and blink duration was shorter. It was encouraging that the algorithm was sensitive to small changes in the blink measures due to the workload manipulation.

Truth data validation. Video recordings were used to generate truth data to help validate and quantify the blink detection algorithm. Eight participants were video recorded while performing trials in a recent experiment. Two trials were recorded for each participant using a Basler high speed camera. The output of the blink detection algorithm was evaluated by two separate individuals using the video recordings. Both individuals watched the recording of each trial and noted each time the participant blinked. If a participant blinked and the algorithm did not detect the blink, a “miss” would be counted. If the algorithm detected a blink when there was not one observed, it was classified as a “false positive.” Only 2.5 percent of blinks were missed, whereas 1.0 percent of blinks were falsely detected. Overall, the blink detection algorithm had an accuracy rating of 96.7 percent

Discussion

In our recent study we collected eye activity data using both EOG and a camera-based eye tracking system. Two advantages of the camera-based system are its completely off-body and it produces position measures. For

example, its eye lid opening measure is in meters. The EOG signal is an electrical measure and cannot be directly related to position. Due to drift in the EOG signals, a high pass filter is commonly used in the signal acquisition hardware. Therefore EOG is good for detecting rapid eye movements, but is not good for measuring gaze angles.

An advantage of EOG is that it does not have restricted field-of-view. The camera-based system can lose its lock on the eye if the participant slouches, changes seating position, or turns their head. Conversely, the EOG signal remains continuous regardless of participant movement. In our experiment participants needed to occasionally look down at the keyboard to press a key. When they did this the camera-based system stopped producing data, whereas the EOG approach did not.

The blink detection algorithm discussed here is adaptive in the sense that it works well for a wide selection of individuals. In addition, after the algorithm has compiled statistics on a few blinks, it can adapt some on the criteria to improve its classification accuracy. The algorithm is also dynamic in the sense that the detection threshold will change in real-time in response to changes in the VEOG signal.

A positive aspect of the blink detection algorithm is that it does not require baseline data or calibration. There is no need (or mechanism) for experimenter adjustments. This algorithm produces measures in real-time, which is an advantage over *post hoc* approaches.

Conclusion

The new blink detection algorithm discussed in this paper works extremely well, in regard to both misses and false positives. It has served well as a tool to support the analysis of electroencephalogram (EEG) data using artifact separation (Credlebaugh, Middendorf, Hoepf, & Galster, this volume). The algorithm produced blink rate and blink duration measures that are sensitive to changes in cognitive workload.

Acknowledgements

The authors would like to thank Chelsey Credlebaugh and Jonathan Mead for their assistance in data reduction and Chuck Goodyear for his help with statistical analysis. We would also like to thank Kevin Durkee, Noah DePriest, and Mark Squire for their technical support. The views expressed in this report are solely those of the authors and do not necessarily reflect the views of the employers or granting organizations.

References

Andreassi. Psychophysiology: Human Behavior and Physiological Response 5th ed. NY, Taylor & Francis Group, LLC. 2007.

Cain, B. (2007). A review of the mental workload literature, Technical Report. Defense Research and Development Canada Toronto.

Credlebaugh, C., Middendorf, M., Hoepf, M., & Galster, S. (this volume). EEG data analysis using artifact separation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Hoepf, M., Middendorf, M., Epling, S. & Galster, S. (this volume). Physiological indicators of workload in a remotely piloted aircraft simulation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Kong, X. & Wilson, G. (1998). A new EOG-based eyeblink detection algorithm. *Behavior Research Methods, Instruments, & Computers. 30*(4), 713-719.

Sforza, C., Rango, M., Galante, D., Bresolin, N., & Ferrario, V. (2008). Spontaneous blinking in healthy persons: An optoelectronic study of eyelid motion. *Opthal. Physiol. Opt. 28*, 345-353.

Wang, Y. & Zhou, J. (2013). Literature review on physiological measures of cognitive workload. Machine learning research group – NICTA

Wilson, G. & Russell, C. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiological determined adaptive aiding. *Human Factors, 49(6)*, 1005-1018.

Wilson, G. F. (1994). Workload assessment monitor (WAM). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 38* (15), pp. 944-944). SAGE Publications.

SACCADE DETECTION USING POLAR COORDINATES – A NEW ALGORITHM

MATT MIDDENDORF, Middendorf Scientific Services
CHRISTINA GRUENWALD, Southwestern Ohio Council for Higher Education
LUCAS STORK, Southwestern Ohio Council for Higher Education
SAMANTHA L. EPLING, Ball Aerospace and Technologies Corporation
MICHAEL HOEPF, Oak Ridge Institute for Science and Education
SCOTT GALSTER, Air Force Research Laboratory

Over the past few decades substantial research has been conducted regarding saccades (rapid eye movements). There are two components of this research. First there is the detection of the saccades, and second how to interpret the saccades features (amplitude, length, and velocity) to inform specific areas of research. This involves both experimental research and clinical applications. The detection of saccades is typically accomplished using two approaches, including cameras and the electrooculogram (EOG). Both of these approaches require algorithms to process the raw data, detect saccades, and calculate the saccade features. The current effort focuses on detecting saccades in the EOG using a new algorithm based on polar coordinates. The details of this algorithm will be presented, as will a calibration procedure and validation of the algorithm's accuracy. This algorithm was used in a recent study in which operator workload was manipulated. The saccade features produced by the algorithm were analyzed with respect to the workload manipulations. These results will be discussed.

Saccades have been used in research for cognitive state assessment, investigation of drug effects, and clinical applications for neurological and psychiatric disorders (Romero, Mañanas, & Barbanoj, 2008). For cognitive state research, saccades have been used both directly and indirectly. Literature suggests that peak saccade velocity can be used directly for the evaluation of mental workload (Di Stasi, et al., 2010). An indirect use of saccades is to mediate ocular artifacts in the electroencephalogram (EEG), which is also often used for workload assessment.

In a recent study (Heopf, Middendorf, Epling & Galster, this volume) participants were asked to track targets using remotely piloted aircraft (RPA) while workload was experimentally manipulated. Participants were equipped with electrodes to measure EEG, vertical EOG (VEOG), and horizontal EOG (HEOG). The primary purpose of the EOG data was to support an ocular artifact mediation approach we refer to as artifact separation (Credlebaugh, Middendorf, Hoepf & Galster, this volume). Since both VEOG and HEOG data were collected, an opportunity was present to develop the polar saccade detection algorithm

The artifacts in the EEG resulting from vertical ocular activity (blinks and saccades) tend to propagate symmetrically from anterior sites to posterior sites (Romero, Mañanas, & Barbanoj, 2008). Horizontal eye movements (saccades) mainly affect lateral frontal electrodes (Lins, Picton, Berg, & Scherg, 1993). Polar coordinates are specified in angle and magnitude. Therefore, the polar coordinate detection approach allows for more specific propagation mapping using the saccade angle.

Background

In a prior study, VEOG was collected to support the development of a blink detection algorithm (Epling, et al., this volume). This was done to deal with blink artifacts when analyzing the EEG data. However, it was realized that there were also artifacts in the EEG due to saccades. A task-related effect was discovered that was a direct result of one of the experimental manipulations (Credlebaugh, Middendorf, Hoepf & Galster, this volume). Specifically, one of the manipulations introduced substantially more horizontal saccades. This made the analysis of EEG data difficult. This is the reason why the current study uses both channels of EOG data, and the polar saccade detection algorithm was developed. With the new algorithm EEG spectral results can be flagged as containing blink *and* saccade (regardless of angle) artifacts. This data can be separated from the artifact free data at the analysis stage.

One particular challenge in the development of the polar saccade detection algorithm was presented by blinks. Specifically, the up slope of a blink has very similar qualities as a saccade. Special queuing logic had to be implemented to prevent a blink from also being counted as a saccade. When an epoch of data contains a blink, saccades are not searched for. This essentially means that blinks "trump" saccades. This must be taken into account

when analyzing EEG data using the aforementioned artifact separation technique (Credlebaugh, Middendorf, Hoepf & Galster, this volume). When blinks are removed, a higher density of saccades will be found in the remaining data.

How the Algorithm Works

The major components of the algorithm are signal filtering, threshold generation, saccade endpoint detection, dynamic linear fit, mathematical calculations, classification, and saccade queuing.

Signal filtering. The raw EOG data contains saccades that are evident to the naked eye. The distinctive shape of a saccade, shown in Figure 1, contains the pre-saccadic spike (Thickbroom & Mastaglia, 1986) followed by a sharp monotonic increase (or decrease for look down and look left). Then there is a slow decay back to zero due to the high pass filter used in the signal acquisition hardware. The raw EOG also contains micro-saccades, which unlike the major saccades, are very small in amplitude but, occur very frequently. These micro-saccades can occur in the middle of a major saccade (Figure 2). When this happens, the micro-saccades can cause the dynamic linear fit portion of the algorithm to make mistakes. Specifically, the full amplitude of the major saccade may not be reported. To prevent this problem, the raw EOG data is filtered using a first order Butterworth low pass filter with a break frequency of 50 Hz.

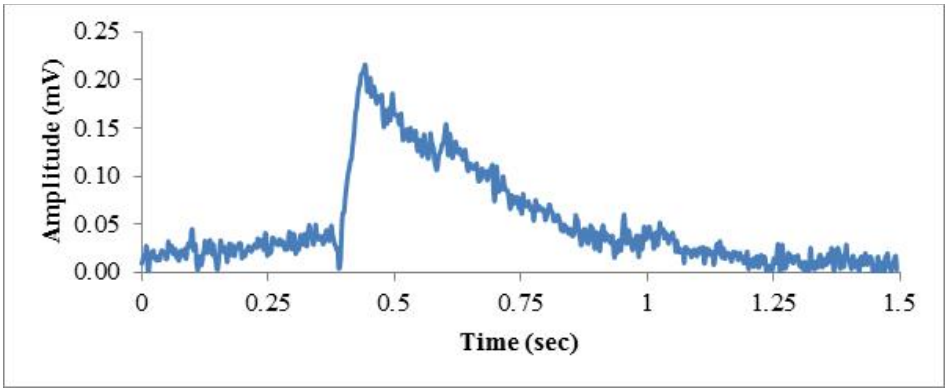


Figure 1. The typical shape of a saccade. This is a horizontal saccade to the right.

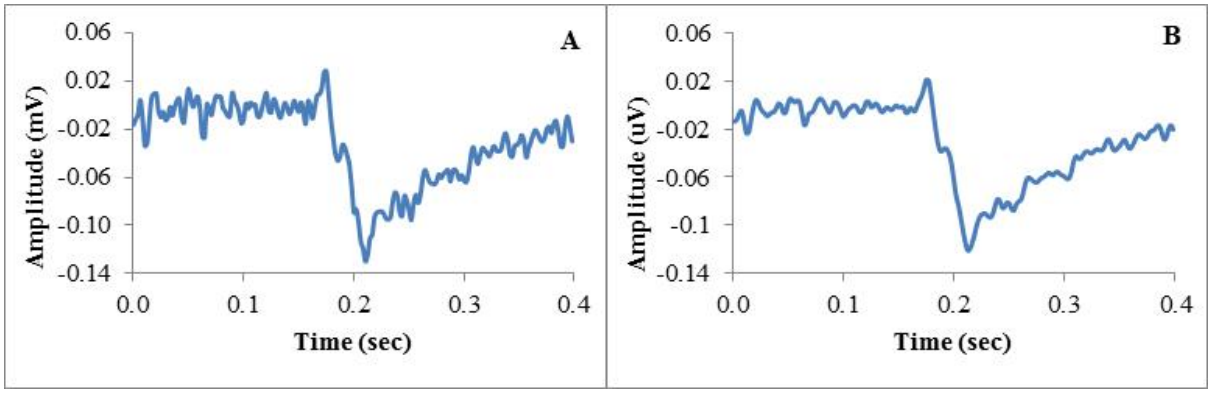


Figure 2. This horizontal saccade is from a leftward eye movement. It has a micro saccade in the middle of it (A) that causes the linear fit to fall short of the full saccade amplitude. After filtering (B) the micro saccade is reduced enough so that it does not interfere with the full linear fit.

Threshold generation. A robust threshold generation approach was developed for the blink detection algorithm (Epling, et al., this volume). A scaled value of this threshold value is used for the polar saccade detection algorithm. The threshold generation approach uses a sliding five second window of raw VEOG data. To minimize the effects of blinks and eye movement on the threshold, the data is high pass filtered using a first order Butterworth filter with a break frequency of 10 Hz. This essentially leaves in the "noise" from which the threshold is calculated. The filtered signal is then rectified and the median is taken for the raw threshold value. The median is used because

the data in the five second window can be highly skewed when there is a blink in the window. Limits are imposed on the raw threshold. Note that the threshold used in the polar saccade detection algorithm is circular (Figure 3).

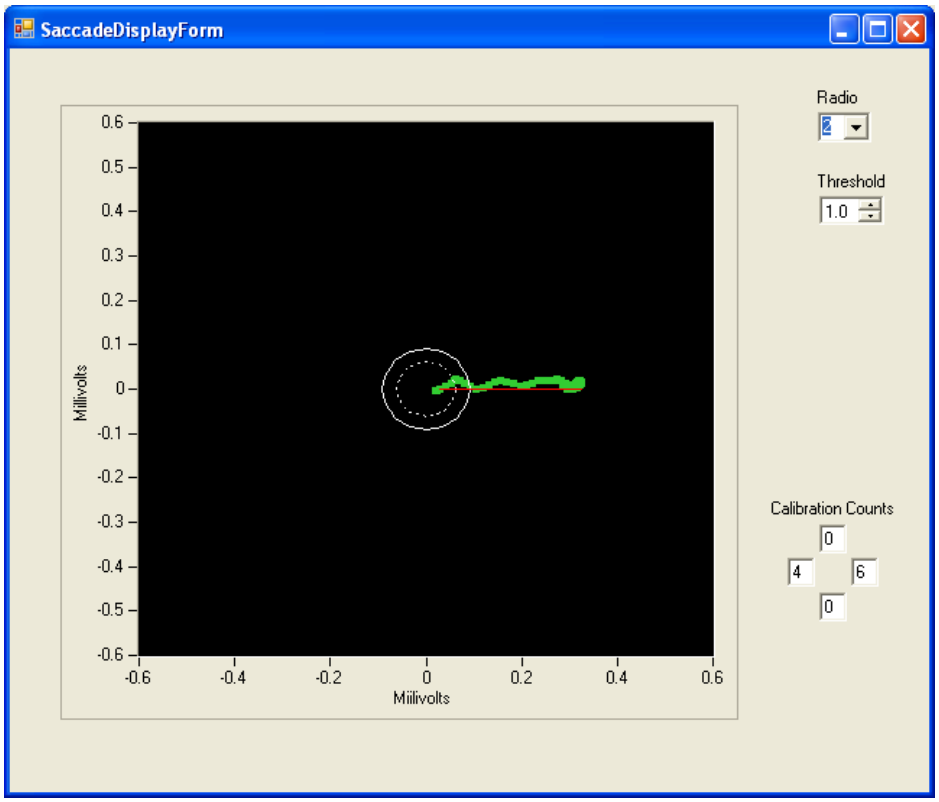


Figure 3. Screen shot of the polar saccade display. HEOG is displayed on the x-axis and VEOG on the y-axis. The solid circle is the threshold and the dashed one is the return circle. The green line is the last 100 EOG samples (VEOG & HEOG). The red line is the saccade detected by the algorithm.

Saccade endpoint detection. The filtered EOG data is evaluated in Cartesian coordinates with HEOG on the x-axis and VEOG on the y-axis. The circular threshold is centered about the x/y origin. Initial saccade detection follows three simple steps. First, the x/y position must start inside the circular threshold. Second, the x/y position must travel outside the circular threshold. Third, the x/y position is allowed to move away from the origin for as many samples as possible until it moves back toward the origin for two samples in a row. The last sample that is moving away from the origin is the end point of the saccade.

The above approach had to be enhanced to prevent many small false positives from being reported due to noise in the EOG signals. This happened when the signal was returning from outside the threshold to inside of it. When the returning signal was near the threshold, the noise in the signal could cause the signal to jump back and forth across the threshold, thus triggering the false positives. To prevent this, a second circle was added (Figure 3) called the return circle. This circle is centered about the origin and its radius is equal to two-thirds of the threshold circle. The saccade endpoint detection logic was modified so that the signal must return to inside the return circle before it can be tested for traveling outside the threshold. The saccade endpoint detection is accomplished using a state machine.

Dynamic linear fit. The linear fits are performed in rectangular coordinates. That is, the VEOG and HEOG are processed separately based on the saccade endpoint. Two vectors are used to find the saccade starting point for each signal (VEOG & HEOG). These two vectors are referred to as the small vector and the big vector. The initial length of the two vectors is the same, which is 20 milliseconds (Chen & Wise, 1996). The heads of the vectors are set to the saccade endpoint and the tails are 20 milliseconds backwards in time.

The length of the small vector remains constant. The big vector grows in length backwards in time. The tail of the small vector is anchored to the tail of the big vector. The small vector is used to terminate the growth of the big vector. Specifically, when the slope of the small vector differs substantially the slope of the big vector, the saccade starting point has been found. After the dynamic linear fit has been performed on both axes, the x and y coordinates of the saccade starting point and ending point are known. Note that these coordinates represent a potential saccade, which must be subjected to classification criteria to determine if the coordinates represent an actual saccade. This is an important distinction because the EOG signals can cross the threshold due to other reasons (e.g., noise, blinks, and slow eye movements).

Mathematical calculations. Once the coordinates of a potential saccade are known several variables must be calculated. The rectangular coordinates need to be converted to polar coordinates (magnitude & angle). The amplitude, length, velocity, and peak velocity are computed for the potential saccade. The two R^2 values from the dynamic linear fits must be combined into a single R^2 value. Finally fixation duration is computed.

Classification. Three of the variables computed above (R^2, velocity, and length) are compared to criteria values to determine if the potential saccade is an actual saccade. All three of these criteria must be met for a positive classification to occur. The criteria values, shown in Table 1, were determined using data from a mini-study with four participants. All of the detected potential saccades were hand scored using EOG playback to generate truth data.

Table 1.
Criteria values used for saccade classification.

Criteria	Minimum Value	Maximum Value
Combined R^2	0.85	N/A
Velocity (mV/sec)	1.0	6.9
Length (sec)	0.28	0.125

Saccade queuing. Special queuing logic needed to be developed to ensure that the identified saccade is not actually the up slope of a blink. While waiting for the excursion of the EOG signal to reach its maximum distance from the origin, the blink detection algorithm is monitored. If the blink detector is active a flag gets set to indicate that the saccade must get queued. Following a positive classification the flag gets checked, and the saccade gets queued if needed. On future updates, if there is a saccade in the queue, the blink detector gets monitored. If a blink occurs the queued saccade is discarded. If the blink detector returns to its initial state and a blink was not detected, the queued saccade gets recorded.

Algorithm Use and Validation

Experimental results. The polar saccade detection algorithm was used in a recent study in which operator workload was manipulated. Although the focus of this paper is on the algorithm, a brief discussion of study results is presented. This is important to illustrate the sensitivity of the saccade measures (amplitude, length, velocity, and peak velocity) to the experimental manipulations. For the sake of brevity, a very condensed description of the experiment is given.

In the study, the participants performed two separate tasks using video feeds from remotely piloted aircraft (RPA). First there was a surveillance task (find the high value target (HVT) walking around in a compound), followed by a tracking task (follow HVTs travelling on motorcycles). In the surveillance task there were two experimental factors, sensor fuzz (on vs. off) and the number of distractors (high vs. low). When sensor fuzz was on (high workload), the video feed was degraded. A high number of distractors (high workload) means there were 48 other people walking around in the compound in addition to the HVT, as opposed to 16 in the low number of distractors condition. In the tracking task the two experimental manipulations were route type (city vs. country) and the number of targets (1 vs. 2). Tracking targets in the city is harder than the country and tracking two targets is harder than one.

The EOG data was processed by the polar saccade algorithm to generate the measures. The measures were statistically analyzed using a repeated measures ANOVA. For the sake of brevity, only the peak velocity plots are shown (Figure 4). For the surveillance task, the fuzz manipulation did not have a statistically significant effect on the saccade measures. The number of distractors manipulation was significant for all four saccade measures.

Although the mean differences were small, they are in the correct direction (Di Stasi, et al., 2010). Specifically, peak velocity was lower in the high workload condition (high distractors). In the tracking task, the route manipulation was not statistically significant. The number of targets manipulation had a significant effect on saccade amplitude, velocity, and peak velocity. These results make sense, even though they are not in the expected direction. In the one target condition participants focused on a single video feed, which did not require large gaze angle changes (i.e., saccades). In the two target condition, participants had to regularly shift their gaze between two video feeds, thus introducing several large saccades. It is reasonable to suggest that this task-related effect caused large mean differences that overwhelmed small differences (in the opposite direction) that may result from increased workload.

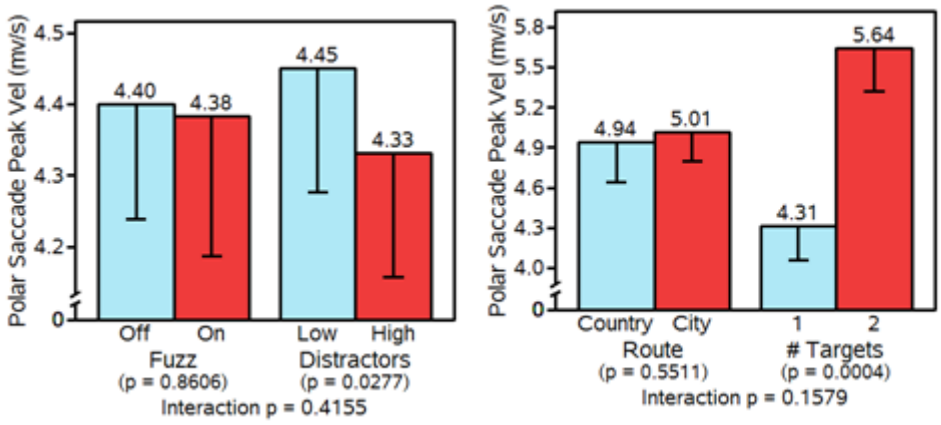


Figure 4. Peak saccade velocity was measured using the polar saccade detection algorithm. The left panel shows results for the surveillance task, and the right for the tracking task. The error bars are standard error.

Truth data validation. A study was conducted to test the performance of the polar saccade detection algorithm. In this study visual stimuli were presented at known angles and distances at regular intervals (1.5 seconds). Two researchers independently reviewed the raw VEOG and HEOG signals to verify that the saccades were present in the signal. This truth data was used to validate the algorithm. Results show that the algorithm had zero false positives, but did have occasional misses. Overall the algorithm had an accuracy rating of 92.6%.

Discussion

The polar coordinate saccade detection approach has an advantage over approaches that independently perform detection on each axis (VEOG & HEOG). The magnitude of a saccade in polar coordinates is almost always larger than the amplitude in the rectangular axes. This makes it easier to detect saccades and makes it possible to detect smaller saccades.

The present approach for saccade detection using a state machine for endpoint detection is computationally friendly. This is because linear fits are *only* performed when an endpoint has been detected. This is in contrast to approaches that perform sliding linear fits in the raw data, which is computationally intensive due to the high number fits that are performed

There is a scaling issue with the EOG data that had to be addressed. The electrical values measured for VEOG and HEOG are not equal for the same angular movement of the eyes. The VEOG is typically smaller than the HEOG, sometimes by as much as a factor of two. This causes the EOG measures to be elliptical rather than circular, thus distorting the calculation of saccade angle. The EOG data needed to be normalized. To accomplish this, a calibration procedure was developed. Visual stimuli are presented on a computer monitor in the vertical and horizontal axes at equal distances from center (Figure 3). The measured responses for each axis are then averaged and used to compute a scale factor. The scale factor is used to normalize the EOG data is real time. It is important to note that the computed scale factor is very stable from calibration to calibration, and from day to day. Additional testing is under way, but it appears that an individual only needs to be calibrated once.

The seating position of participants should be adjusted so that their eyes are near the center of the monitor. Otherwise some error will occur in the calculation of saccade angle. This is true for calibration and experimental

trials. A future enhancement is planned that will allow offsets to be entered for the participants position relative to the monitor.

For the purpose of artifact mediation in EEG, an algorithm was written to detect saccades directly in the EEG signal itself (Credlebaugh, Middendorf, Hoepf, & Galster, this volume). To improve the accuracy of the EEG-based saccade detection algorithm, it will be coupled with the polar saccade detection algorithm. Future work will be performed to corroborate the EEG-based results with the polar saccade results. Specifically, if an EEG-based saccade is detected, then there must be a corresponding polar saccade within the correct angle range.

Conclusion

The measures produced by the polar saccade detection algorithm have been shown to be sensitive to experimental manipulations. In the results reported above, the manipulations introduced a task-related effect that caused a systematic change in eye behavior. Therefore it cannot be concluded that the saccade measures are indicators of cognitive workload, but it is encouraging that the algorithm is capable of measuring small changes. The algorithm produces these measures in real time and in a computationally efficient manner.

A calibration approach was developed to allow the two axes of EOG to be normalized, thus improving the accuracy of the reported saccade angles. The calibration procedure allows researchers to account for individual differences. Early results indicate that participant calibration is stable from day to day. Another positive aspect of the current work is it will support artifact mediation approaches when performing EEG analysis.

This algorithm will be used in future research to determine its usefulness for assessing cognitive workload. Careful consideration will be taken to ensure that experimental manipulations do not systematically change eye activity. Enhancements need to be made to the algorithm to improve overall accuracy of 92.6% (see results section).

Acknowledgements

The authors would like to thank Chelsey Credlebaugh for her assistance with data collection and Chuck Goodyear for his help with statistical analysis. The views expressed in this report are solely those of the authors and do not necessarily reflect the views of the employers or granting organizations.

References

Chen, L. L. & Wise, S. P. (1996). Evolution of directional preferences in the supplementary eye field during acquisition of conditional oculomotor associtions. *The Journal of Neuroscience 16*(9), 3067-3081.

Credlebaugh, C., Middendorf, M., Hoepf, M., & Galster, S. (this volume). EEG data analysis using artifact separation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Di Stasi, L., Rennar, R., Staehr, P., Helmetr, J., VelichKovsky, B., Canas, J., Cantena, A. & Pannasch, S. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine 81*(4), 413-417.

Epling, S., Middendorf, M., Hoepf, M., Galster, S., Gruenwald, C., & Stork, L. (this volume). The electrooculogram and a new blink detection algorithm. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Hoepf, M., Middendorf, M., Epling, S. & Galster, S. (this volume). Physiological indicators of workload in a remotely piloted aircraft simulation. In *Proceedings of the Eighteenth International Symposium on Aviation Psychology*, Wright State University.

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in EEG and event-related potentials I: Scalp topography. *Brain Topography 6*(1), 51-63. doi: 10.1007/BF01234127

Romero, S., Mañanas, M. A., & Barbanoj, M. J. (2008). A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical variables: A simulation case. *Computers in Biology and Medicine 38*, 348-360. doi: 10.1016/j.compbiomed.2007.12.001

Thickbroom, G. W., Mastaglia F. L. (1986). Presaccadic spike potential. Relation to eye movement direction. *Electroencephalogr. Clin. Neurophysiol. 64*, 211–214 [PubMed]

Wang, Y. & Zhou, J. (2013). Literature review on physiological measures of cognitive workload. Machine learning research group – NICTA

ENROUTE ATC INDUSTRY PERCEPTIONS OF SIMULATION FIDELITY

Colin Dow
University of Waterloo
Waterloo, Canada
Jonathan Histon
University of Waterloo
Waterloo, Canada

Enroute air traffic control (ATC) relies heavily on simulation in training, research, and concept development applications. However, it has little domain-specific research on the effects of simulation fidelity and lacks a standardized definition of simulation fidelity in the literature. A survey of ATC industry professionals shows that simulation fidelity is not perceived to be well defined for the domain of enroute ATC, regardless of respondent nationality, experience, use of simulation or gender. Parts of the operational environment that survey respondents felt were important components in a definition of simulation fidelity are reported; Communications is the most important component regardless of nationality, experience, use of simulation or gender. Implications for the development of a standardized definition of simulation fidelity are discussed.

Simulation fidelity has been researched and investigated for over half a century, yet it remains a somewhat nebulous concept today. The high-level concept of simulation fidelity can be best understood from a definition posited by Hays and Singer: "Simulation fidelity is the degree of similarity between the training situation and the operational situation which is simulated (1989, p. 50)." While this definition is intuitive, more detail is needed for operational applications such as determining the most effective simulation environments for training. Hays and Singer have also provided a more comprehensive definition:

> "Simulation fidelity is the degree of similarity between the training situation and the operational situation which is simulated. It is a two dimensional measurement of this similarity in terms of: (1) the physical characteristics, for example, visual, spatial, kinesthetic, [auditory], etc.; and (2) the functional characteristics, for example the informational, and stimulus and response options of the training situation (1989, p.50)."

Specific components, and more generally a definition of simulation fidelity, are not found in the ATC literature. This is despite the widespread and frequent use of simulation in ATC for a variety of purposes that include training, testing new operational concepts or tools, and research into the future ATC environment. Simulation environments of varying degrees of fidelity are relied upon throughout these different areas, from a classroom-based scenario introducing new trainees to basic concepts to the complex, multi-user MACS simulation used for research at the NASA Airspace Operations Laboratory (e.g. Kraut et al, 2011). When fidelity is reported and discussed within the ATC literature, it is most often in the general terms of low, medium and high; however it is not clear that all the simulators reported in any one category are in fact of equivalent fidelity. For example, Loft et al. (2004) developed an enroute ATC simulation environment intended for research on various human factors related topics. While they discuss their simulator's usefulness as a medium fidelity research tool, the lack of definitions for 'low, medium and high' fidelity make it difficult and potentially ambiguous to compare with other simulation environments . Establishing a definition specific to enroute ATC would provide a formal reference point when discussing simulation fidelity, allowing for critical research to be conducted on the links between simulation fidelity and simulation use within the industry.

Other fields have developed a domain-specific definition of simulation fidelity. For example, Estock et al. (2006) identified and refined specific environmental components (Estock et al. refer to these as dimensions rather than components) that they believe affected the fidelity of a simulation of an F-16 cockpit. Some of the components Estock et al. identified were unique to their work environment, such as the "Visual scene simulation" or "Whole body motion", while others such as "Communications" are important in a variety of work environments. This demonstrates the contextual nature of simulation fidelity definitions, as the components specified by Estock et al. (2006) are appropriate for the simulation of an F-16 cockpit, but many of their components would not work for a simulation of an ATC workstation or an operating room.

An important aspect of the process specified by Estock et al. is that once they had identified their fidelity components, they were verified by consulting with flight simulation experts to determine their validity. As identified by Hays and Singer, receiving feedback from subject matter experts is an important step in defining simulation fidelity for a particular domain (1989). Since they are experts within the domain being studied, their experience with the operational environment will be able ensure that no components have been overlooked.

This process of narrowing the focus of a fidelity definition to be highly domain specific is necessary for researchers to be able to study how fidelity is perceived in a given work environment. More importantly, this allows for objective research into the links between fidelity and simulation use for training, testing new operational concepts and research within the given domain. Developing a clearer picture of what components affect fidelity for a particular operational environment opens up the potential for using a variety of different simulation environments to achieve outcomes in each of these areas in perhaps a more effective and cost-efficient manner.

As part of a project developing a simulation fidelity definition for enroute ATC (e.g. similar to the Estock method note above), an industry wide survey was conducted investigating the perceptions of simulation fidelity and how simulation of varying degrees of fidelity ought to be used. Considering that the process of defining simulation for a particular domain has seldom been done, and never for enroute ATC, there was a clear opportunity to develop a domain-specific definition of simulation fidelity for enroute ATC (Dow and Histon, 2014). Included in this survey were questions that sought to determine if simulation was already a well defined concept within ATC, and what environmental components individuals were currently considering when making a fidelity determination for a simulation environment.

As part of assessing the potential for a general consensus on the appropriate components in an ATC simulation fidelity definition, this paper compares the perceptions of different demographic sub-groups regarding the need for a standardized simulation fidelity definition for the ATC domain, as well as the particular environmental components that they believe ought to be considered when defining simulation fidelity for enroute ATC. The demographic sub-groups used for comparison are nationality, survey participant's primary use of simulation, survey participant's years of experience with simulation, and gender. What follows is a description of the methods used to gather the professionals' perceptions on simulation fidelity, presentation and analysis of the results, and finally a discussion of the findings and limitations of the study.

Method

As part of a larger effort investigating the concept of simulation fidelity within the ATC domain, professionals within the ATC industry were asked a series of questions about simulation fidelity through an online survey. This paper focuses on the responses to two questions in the survey. The first was a question on whether or not professionals believe simulation fidelity is well defined within the enroute ATC industry. The second question, asked individuals to provide a list of the environmental components that they believed affect the fidelity of a simulation environment.

The survey was first distributed to personal contacts within various ANSPs and researchers around the world who met the participant criteria of the survey. This was done to try to ensure that the survey participants were coming from as reliable a source as possible due to concerns about the lack of control and verifiability of those completing an online survey anonymously. The target population was anyone who had experience developing or using ATC simulations, which included the following examples of potential participants:

- Active air traffic controllers who have used simulation for training / participated in simulation studies
- Controller training designers / developers
- Air traffic control instructors
- Researchers who have used simulators for human-in-the-loop studies
- Operational concept developers and controller tool developers who have used the results of simulation studies

The survey was then made publicly available on aviation public domain websites (e.g. liveatc.net, pprune.org) and through air traffic control publications (e.g. ATC Network and Air Traffic Management) where the target population for this research typically frequent. Free response questions provided opportunities to carefully screen responses for appropriateness and consistency with the background and experience reported by the participants.

Survey questions consisted of a mix of Yes/No, Likert scale ratings, and short and long answer questions. Topics covered in the survey include participant perceptions regarding the concept of simulation fidelity in the domain of air traffic control, what level of simulation fidelity is required to train for a certain skill or test/evaluate a particular concept, and the acceptability and accuracy of a simulation fidelity definition and categorization system developed for the enroute ATC domain. Full results are presented and discussed in Dow (2015).

There were 86 completed responses. The key characteristics of participants were gender (Male=69, Female=16, Prefer not to respond=1), nationality (United States=34, Canada=29, International=22, Not specified=1), years of experience working with simulation (0-5 years=23, 6-10 years=11, 11-15 years=14, 16 years or more=38), and the survey participant's primary use of simulation (Training=58, Testing new operational concepts=17, Future ATC environment research=11).

Results

Participants were explicitly asked whether or not they believe simulation fidelity was a well-defined concept in the domain of ATC. Participants were given radio buttons and could chose either Yes or No and were asked to explain their choice in a free-response text box. The results are presented in Figure 1.

The 'Yes' and 'No' columns in Figure 1 represent responses where participants demonstrated a clear understanding of the question based on their follow-on explanation of why they answered Yes or No. Survey participant explanations that clearly indicated they did not understand the question were categorized as 'Non pertinent'. For example, a response of, "I have been working in ATC for 23 years, and simulation has been in use all of this time." was judged to indicate the participant had not understood the question. The final column, 'No explanation', represents the percentage of responses where survey participants provided no explanation to their Yes/No answer and therefore an assessment of their understanding of the question could not be established. The analysis below focuses only on the Yes and No response columns, referred to henceforth as the "analyzed responses" in the subsequent analysis.

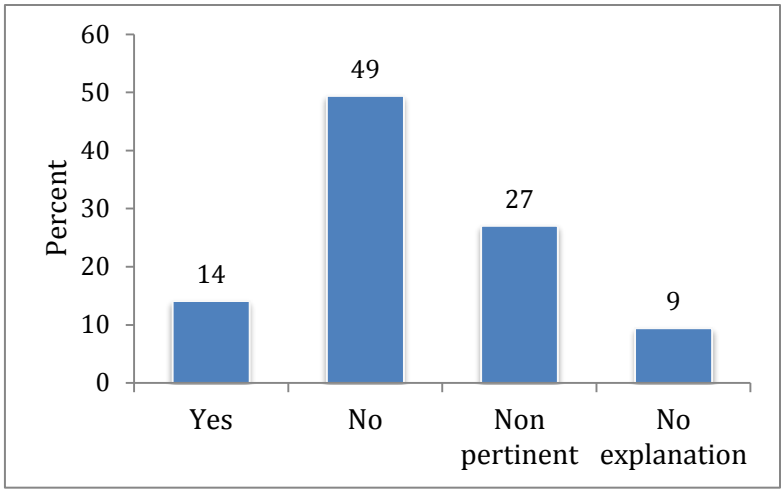


Figure 1. Survey participant responses to the question: "Do you believe that simulation fidelity is a well-defined concept in the ATC domain?" N=85.

Results of a chi-square goodness of fit test show that the observed Yes/No response frequencies are statistically significant. They are different from what would be expected if half of the ATC industry believed that simulation fidelity is well defined for the ATC domain ($\chi^2(1, N=54)=16.67$, $p<0.001$). Demographic data collected as part of the survey was used to investigate whether the perception that fidelity is not well defined is wide-spread across gender, nationality, experience and primary use of simulation. A comparison between the Yes/No response rates for these four demographics is presented in Figure 2. As seen in the figure, the proportion of Yes/No responses, while varied, shows a strong and consistent pattern of a belief that simulation fidelity is not well defined. A chi square analysis was performed to determine if there were any differences within the demographic groups. It was found that there were no differences with regards to the belief that simulation is **not** well defined for ATC when comparing within the demographic groups of gender, (χ^2 (1, N=54)=0.04, p=.851), nationality, (χ^2 (2, N=54)=2.06, p=.385), years working with simulation, (χ^2 (3, N=54)=3.78, p=.287), or survey participant's use of simulation, (χ^2 (2, N=54)=1.83, p=.400.

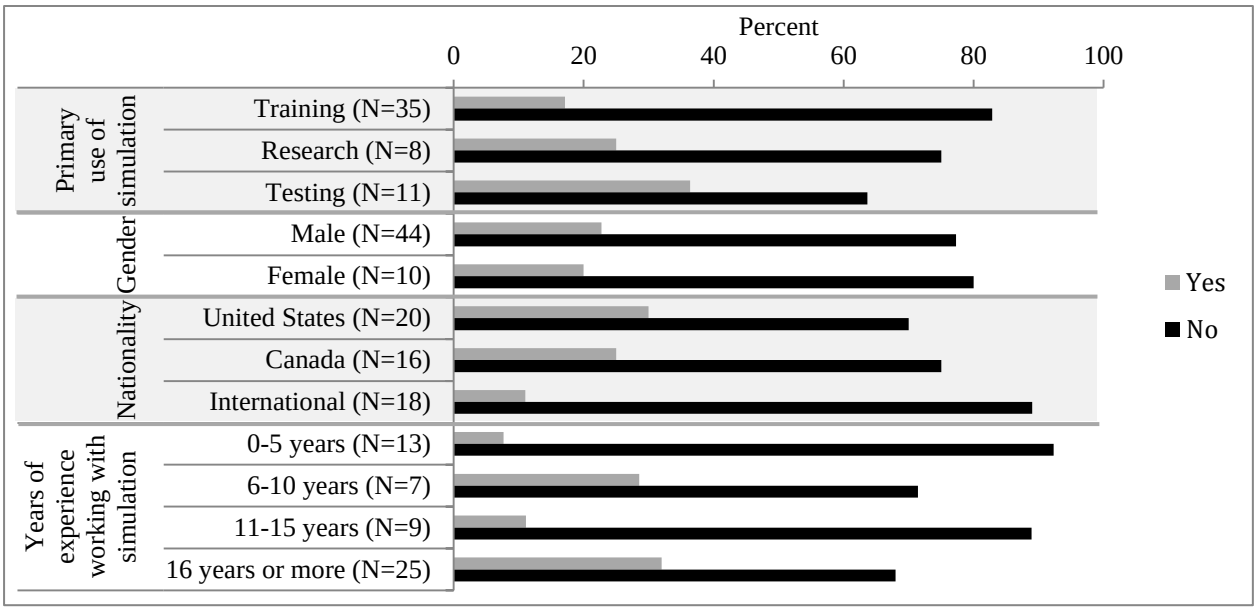


Figure 2. Demographic sub-group comparison of responses to the question: "Do you believe that simulation fidelity is a well-defined concept in the ATC domain?" N=54

In order to further explore why survey participants feel simulation fidelity is not well defined for ATC, Table 1 presents sample comments from both the pertinent "Yes" and "No" responses to the follow-up question asking if they could explain their answer in more detail.

Table 1.
Sample comments from survey participants' explanations of their responses to the question in Figure 1.

Sample Comments from 'Yes' Responses	Sample Comments from 'No' Responses
• I think that it is well defined, but in reality, it is under-utilized. • We all know what fidelity means. Realistic. Realistic aircraft, realistic routes, realistic responses. Responses that are dynamic in nature, changing depending on what the student is doing. • Though I'm not aware of a quantitative definition, fidelity is something researchers and trainers know when we see it, and it is easy to ordinally rank different simulators or simulations in terms of their fidelity. I have created and used an informal table that lists the different levels of fidelity and their characteristics. • I think it's defined and conceived just fine, but, in my opinion, it's not implemented very well.	• I think that "simulation fidelity" is one of those concepts that "everyone knows what it means" but that formal, valid definitions are lacking. • I believe simulation fidelity means different things to different people. I believe current controllers are not involved enough in validating the fidelity of a simulation before it is used in the field. • I have not come across such a concept definition so far. On the contrary, many times the term "high fidelity" is interpreted in various ways. • I've met a lot of people in my business who have a significantly different perception of what is high and what is low fidelity simulation. • My interpretation of high fidelity simulation is the recreation of the real live ATC environment in as much detail as possible. I don't believe this to be a universally shared interpretation and that there are varying degrees of separation from my idea.

The sample comments from those who responded "Yes", are representative of a recurring belief that simulation fidelity is a well-defined concept, but is not put into practice or referenced enough with regards to the many uses of simulation within the industry. However, what is clearly demonstrated in the sample comments from those who responded "No" is that the problem is not with an individual's definition in isolation, but rather when discussing the issue as a collective and not sharing the same definition with those they interact with.Comments such as "I believe simulation fidelity means different things to different people", "On the contrary, many times the term "high fidelity" is interpreted in various ways", or "I don't believe this [his/her interpretation of fidelity] to be a

universally shared interpretation and that there are varying degrees of separation from my idea", all indicate an awareness of the impact of a lack of standardization with regards to simulation fidelity in the ATC domain.

In addition to the question above, survey participants were asked to provide the environmental components they believed affected the fidelity of a simulation environment. Eight optional text boxes were provided to participants in order to receive as many different environmental components as possible. The responses were then coded by identifying a high level topic or theme in the response, and the top ten coded reponse frequencies from the overall responses as well as the nationality demographic groups are presented in Table 2.

Table 2.

Top ten coded response frequencies for all survey participants and nationality demographic groups for the question "In your opinion, what parts of the enroute ATC work environment affect the fidelity experienced by someone using an enroute ATC simulation?"

Fidelity Components	Response frequency (% of N)			
	Overall (N=73)	United States (N=31)	Canada (N=24)	International (N=18)
Communications	62	55	71	56
Equipment	42	35	46	44
Environment	32	32	42	17
Aircraft performance	30	16	46	28
System participants	29	23	38	22
Unpredictability	29	19	42	28
Traffic	23	19	25	28
Weather	21	10	42	11
Automation	19	16	17	28
Operational stress	11	6	13	11

Tables similar to Table 2 were also prepared for the demographic groups of gender, simulation use, and years of experience with simulation; however, they are not shown due to space considerations. Across all demographic groups, the "Communications" component received the highest response frequency for each sub-group, indicating its high overall rank was the result of a widespread and shared perception of its importance for a definition of fidelity for enroute ATC simulation environments. For a detailed description of the fidelity components, see Dow (2015). Not all components appear to be perceived equally across the different nationality groups, though statsitical tests of significance have not been completed. For instance, Canada had a much higher response frequency for "Weather," while the United States had lower response frequencies for "Aircraft performance", and the International group had lower response frequencies for the "Environment" component but higher response frequencies for "Automation". From the tables not shown, the researchers demographic group overwhelmingly identified Communications (73%) and Equipment (45%) components, while all other components were at less than 27%. The demographic group of testing new procedures had almost no (< 7%) mentions of Unpredictability, Weather, Automation, and Operational Stress. Table 2 also illustrates that there were differences in how many components each nationality was providing, with Canadian survey participants providing more components then the other two groups.

Discussion

The results presented in Figure 1 showed that simulation fidelity is viewed as not being well defined for enroute ATC; this indicates that there is an opportunity for developing a standardized definition of simulation fidelity for the enroute ATC domain. The consistency of this finding within the different demographic sub-groups in Figure 2 suggests that the notion of simulation not being well defined is wide spread across gender, nationality, survey participant's simulation use, and their years of experience with simulation.

The examination of what components participants felt contributed to simulation fidelity indicates some potential sources of this perception, as well as the basis for development of a standardized definition. It is clear that the components listed in Table 2 are not unanimously agreed as only one component (Communications) was identified by more than half of all participants. The response frequencies from the different nationality groups also showed that it appears individuals are considering different parts of the operatinal environment when making a determination regarding the fidelity of an enroute ATC simulation environment. While the overall response rates give confidence in drawing conclusions for the primary results, the small number of participants (minimum of 11)

within some of the demographic groups suggests caution in interpreting the findings for components for any one sub group. The presence of the differences, however, is consistent with the ambiguity and confusion around the concept of simulation fidelity and the difficulty in having objective discussions regarding the implications of simulation fidelity, let alone conduct research regarding the link between fidelity and training, for example.

However, the components listed in the first column of Table 2 also offer a reasonable consensus of the components that can affect the fidelity of an enroute ATC simulation environment. Identifying the range of components that are considered and capturing them within a proposed definition of simulation fidelity (see Dow and Histon, 2014) is an important step towards developing an operationally useful and widely accepted understanding of simulation fidelity in enroute ATC.

Summary

Given the significant amount of survey participants who believe simulation fidelity is not well defined for ATC, and the variation inherent with individuals' sets of fidelity components, there is an opportunity for increased standardization by developing a definition of simulation fidelity for the enroute ATC domain. One such definition has been developed by Dow and Histon (2014). This construct presents a set of environmental components that can affect the fidelity of an enroute ATC simulation simulation environment. Most important is the process it has also taken to include SMEs in both the development of the construct and the validation of the final product. As noted earlier, this process of validation by those who work closest with the operational environment being simulated is important in not only developing a definition that captures the relevant components, but one they will have more confidence in once they are using it.

Even if a definition were developed, this does not close the door on the topic of simulation fidelity. It is a first step to a clearer understanding of the concept, and more work is needed. What a definition will provide is the foundation of how simulation environments are compared and contrasted, essentially the points of comparison. The definition would then need to be operationalized in some form to be able to clearly delineate between simulation environments. The most likely form is that of a categorization system similar to that used by the FAA to classify flight simulation environments but for enroute ATC simulations. This preliminary work then creates the opportunity for the important research in to the link between fidelity and simulation use for training, testing new operational concepts and future ATC environment research.

Acknowledgements

We would like to thank the National Sciences and Engineering Research Council (NSERC) for their financial support of this project.

References

Dow, C. and Histon, J. (2014). An Air Traffic Control Simulation Fidelity Definition and Categorization System. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 58(1), 92-96. DOI: 10.1177/1541931214581020

Dow, C. (May, 2015). *Developing an Objective Definition of Simulation Fidelity for Enroute Air Traffic Control.* (Master's dissertation). University of Waterloo; Waterloo, ON, Canada.

Estock, J. L., Alexander, A., Gildea, K. M., Nash, M., and Blueggel, B. (2006). A Model-based Approach to Simulator Fidelity and Training Effectiveness. In *Proceedings of Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).* Orlando, FL: National Security Industrial Association. Retrieved from design-usability.aptima.com/publications/2006_Estock_Alexander_Gildea_Nash_Blueggel.pdf

Hays, R.T. and Singer, M.J. (1989). *Simulation Fidelity in Training System Design: Bridging the Gap Between Reality and Training.* New York, NY: Springer.

Kraut, J., Kiken, A., Billinghurst, S., Morgan, C., Strybel, T., Chiappe, D., Vu, K. (2011). Effects of Data Communications Failure on Air Traffic Controller Sector Management Effectiveness, Situation Awareness, and Workload. *Human Interface and the Management of Information. Interacting with Information Lecture Notes in Computer Science*. Volume 6772, pp 493-499. DOI: 10.1007/978-3-642-21669-5_58

Loft, S., Hill, A., Neal, A., Humphreys, M., & Yeo, G. (2004). ATC-lab: an air traffic control simulator for the laboratory. *Behavior Research Methods, Instruments, & Computers*, 36(2), 331–8. Retrieved from http://www.ncbi.nlm.nih.gov/plubmed/15354699

FOLLOW-UP EXAMINATION OF SIMULATOR-BASED TRAINING EFFECTIVENESS

Maxine Lubner, Ph.D.
Vaughn College of Aeronautics and Technology
New York, NY
Andrew R. Dattel, Ph.D.
Embry-Riddle Aeronautical University
Daytona Beach, FL
Deb Henneberry, M.A.
Vaughn College of Aeronautics and Technology
New York, NY
Sharon DeVivo, Ed.D.
Vaughn College of Aeronautics and Technology
New York, NY

A descriptive examination of the effectiveness of a simulator-based training program for pilots was conducted. Of 55 students of varying backgrounds, but mostly with limited flight experience, 13 enrolled in an intensive, simulator-based flight training program. Within two years the remainder had enrolled in conventional collegiate flight training, supplemented with some simulator training. The students in the intensive program completed their FAA Private Pilot certificates in an average of 5 weeks (not including simulator time). Moreover, the intensive program group earned their private pilot's certificate in statistically significantly fewer hours (M=46.03) than the conventional collegiate flight training group (M=76.06). The intensive group returned to conventional training and completed their Commercial certificates in an average of 20 weeks and CFI qualifications in an average of 40 weeks. The potentially useful aspects of the intensive program are discussed, including type of training such as intensive classroom, simulator and traditional in-aircraft instruction in addition to the psychosocial impacts of camaraderie and shared learning experiences.

Introduction

Aviation simulators have been a part of flight training since 1909, shortly after the Wright Brothers' first flights. The precursor to the modern aviation simulator, the Link Trainer, was developed as a cost effective and efficient form of flight training that could improve instrument flying skills from the early days of flying and during World War II (Wicks, 2003). When designed correctly, a training program that includes the appropriate use of simulators will provide facets of instruction that may not be otherwise possible (Harris, 2011).

Simulator centric training (SCT) offers several advantages. Firstly, depending on the equipment used and scenario being taught, costs can be significantly reduced when simulators instead of in-aircraft training are utilized. Capital investment in aviation simulators is becoming increasing affordable because high fidelity simulation is not required for positive transfer of training (Salas, Bowers, & Rhodenizer, 1998; Taylor et al., 1999). Secondly, overall training time can be used more efficiently because simulator training can take place when inclement

weather prohibits in-aircraft training. Thirdly, many effective training scenarios can be created in a simulator. Learning objectives can be implemented in a deliberate manner to ensure that all performance criteria are satisfied. Fourth, by freezing the simulator during performance evaluation, deficiencies can be discussed as they occur. Full attention can be given to the analysis without devoting the resources needed to fly the airplane.

Fifth, the simulator offers many opportunities for part-task training, where the instructor can break a complex task into smaller parts so that the student can concentrate on mastering those and then re-incorporate the components into the larger task (Dattel, Durso, & Bedard, 2009; Harris, 2011). By evaluating performance at the time of action, flight instructors can better assess students' conceptual understanding of situations when part-task training is implemented. A greater conceptual understanding is particularly important for complex aviation maneuvers, non-routine conditions, and situation awareness (Dattel, Durso, & Bedard, 2009). One example of part-task training is allowing students to control the aircraft's yoke while the instructor handles the task of using the throttle. Another less commonly employed example is to have the student use only the throttle while the instructor operates the other airplane controls. Performing these exercises in a simulator allows the additional and important opportunity to return to the smaller building blocks making up those tasks, while engaging the student's conceptual understanding of the procedure. In this example, the simulator records the student's actions, thereby allowing analysis and reflection of each task component by the student and the instructor.

Sixth, by incorporating scenario-based training (SBT), students are able to develop mental models that permit them to hone judgment and decision-making skills for a variety of situations (FAA, 2008). Other factors have been examined in relation to simulator based training. Complex skill sets, such as crew resource management, have been positively transferred in even the most commonplace desktop simulators (Johnston, McDonald, and Fuller in Harris, 2011).

Comprehensive instruction in a simulator must include conceptual and procedural methodologies, both of which are independent of simulator fidelity (Hawkins, 1997). Conceptual training is accomplished through the incorporation of scenario-based instruction as a part of the decision making process. This technique is also effective in the mastery of other skills, including traffic pattern operations. Simulator training can easily incorporate conceptual, procedural, scenario, collaborative and individual styles of training (Dattel, et al., 2009, Dattel, Kossuth, Sheehan, & Green, 2013).

While flight simulators are generally considered an enhancement to the training process, a multi-factorial, instructional model should be followed by instructors and training program designers to produce an optimal outcome. Simulator training should avoid excessive reliance on simulation-centric training (SCT). Certainly, individual instructor effectiveness is reported as necessary to ensure positive and satisfying pilot training (AOPA, 2010). Cognitive, and possibly psychosocial variables related to the students should also be included in a comprehensive flight training program. Several individual level variables have been found to influence training outcomes before and during training, including motivation, self-efficacy and attitudes (Alvarez, Salas, & Garofano, 2004). Scenario based training (SBT) is likely to enhance simulator centric training (SCT) because this approach includes the social and psychological components of

instruction, such as collaborative and individual techniques, cognitive advancement of decision making skills, ways to increase motivation, create useful attitudes, and uncover gaps in comprehension.

Vaughn College embarked on a simulator-centric flight training program in partnership with a company far from its New York campus, where three cohorts of students had to travel together for an intensive flight training schedule. Later, these students returned to a local New York flight school. There were also groups of students who followed conventional flight training at the local flight school. Conventional training included some simulator practice, lessons spaced over time, and reduced opportunity for interaction with fellow flight students. Although it was not possible to create a control group or even quasi-experimental design here, the brief intensive and the ongoing conventional flight training programs provided an opportunity to identify predictors and questions about ways to increase efficiencies in flight training: Would the intensive program help students to acquire FAA pilot qualifications in a timely manner, what depth and duration of knowledge and skills could be acquired, would time and costs for training be affected, what other aspects of flight training should be examined in a more constrained manner?

Method and Program Description

Three sets of cohorts from Vaughn College were sent to a simulator-centric flight training school in the southwest United States. Each cohort started with five to eight students. To qualify for the cohort, students had to have a G.P.A. of 3.0 or better, possess an FAA Class III Medical Certificate, take a demonstration flight, successfully pass the FAA private pilot knowledge exam, obtain financial counseling and agree to remain substance free during the training period.

The training period was designed to last approximately 4-6 weeks. During each training period, students stayed at a hotel and dined together. Students flew in aircraft and simulators six days a week. Students commented that the social time and cohesiveness they experienced when away from home was an important part of their experiences.

The program was designed for students to travel to the simulator-centric flight training school for the pilot private training, then return home for final completion that ended in successfully obtaining a private pilots certificate. After a few weeks of completing the private pilot certificate, students would return to the simulator-centric flight training for the instrument rating, return home for the final completion of the instrument rating, and then repeat the same format for the commercial pilot certificate.

Beginning in January 2012, students in the first cohort group travelled to the simulator-centric flight training school for private pilot training, instrument training, and commercial training. However, due to internal and external issues, the second cohort group only travelled to the simulator-centric flight training school twice – for private pilot and instrument training. The third cohort group only attended the simulator centric flight training school for their private pilot training.

The second cohort group continued their commercial flight training without the benefit of the simulator-centric flight training at a Part 141 flight school located about an hour's drive from Vaughn College. The third cohort group continued their instrument and commercial flight training at the same Part 141 flight school. Beginning in the Fall of 2013, all flight students attended the Part 141 flight school without the benefit of the simulator-centric flight training.

Results

An independent means t-test was conducted between the group that received a private pilot's license with intensive combined with simulator-based training and the group that received a private pilot's license with conventional training (no simulator training). The t-test showed $t(23) = 6.704$, $\eta^2 = .661$, $p < .001$ that simulator training (See Figure 1) at the private pilot stage (M=46.03, SD=10.21) significantly reduced the number of flight hours required to complete the training compared to the group that had conventional (no simulator) training (M=76.06, SD=11.76).

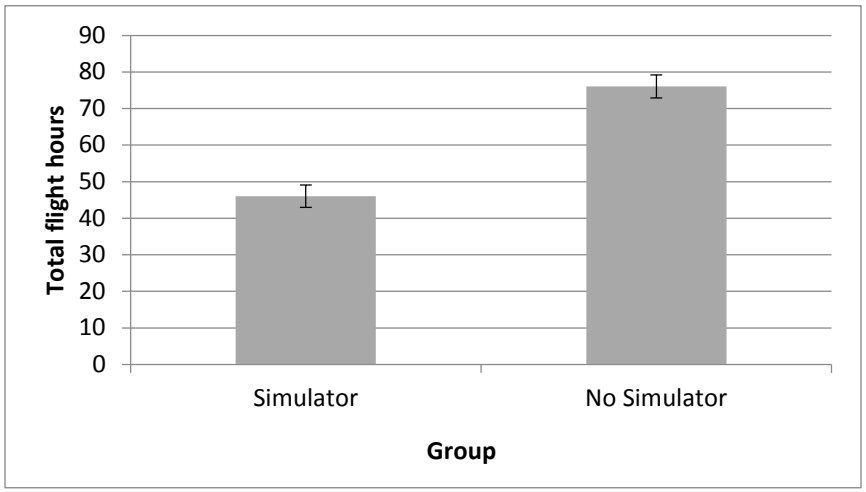


Figure 1. Average flight time for private pilots at completion by group.

At this point, only four students enrolled in Vaughn College's flight professional program have completed instrument training without the benefit of a flight simulator program. However, it should be noted that six members of the cohort groups have completed some or all of their advanced ratings with varying degrees of simulator training (see Table 1).

Table 1.
Average Flight Time at Completion of Advanced Training of Students that Started in the Cohort Groups.

Rating/Certificate	
Instrument	M=78.43 (SD=9.44)
	n=6
Commercial	M=130.6 (SD=18.69)
	n=6

Conclusion

Although an experiment was not conducted, it was found that students who followed the intensive, simulator-based flight training programs earned their FAA Private Pilot certificates in a significantly shorter time than those who attended the conventional flight training program. Given the small numbers of students tested, this finding of statistical significance indicates that there was a large effect related to the timing of the training effectiveness (Cohen, 1988).

From anecdotal evidence, it appears that the simulator training, close spacing of appointments for flying lessons and psychosocial aspects of camaraderie and intensive learning all contributed to the students' successful, rapid completion of their FAA Private Pilot certificates. Students talked about their social bonding, collaboration on flight training and ability to help each other to reduce anxiety and share reward systems as most helpful while they had traveled together and after they returned to New York. They had numerous opportunities to experience the psychological components of reflective learning (Drago-Severson, El; Helsing, Kegan, Popp, Broderick, & Portnow, 2001), such as being able to rehearse, comprehend and retain knowledge. Similarly, the students had opportunities to acquire decision making skills by emulating habits demonstrated by expert pilots. A large amount of time was dedicated to instructor-guided practice so that the students would acquire flight skills needed for safety and proficiency (Lubner, Adams, Hunter, Sindoni, & Hellman, 2003).

The students who attended the intensive, simulator based program did not appear to complete their subsequent pilot qualifications in a comparatively short time, however. Variables including the depth and duration of students' pilot related skills and knowledge, effective components of the methods of instruction, instructor effects, and whether the program conferred any advantages on the students for acquiring subsequent pilot skills and knowledge, all remain to be tested.

A well-designed training program using conventional and simulator-centric training, incorporating camaraderie, and instructor proficiency in this form of instruction, has early indications of being successful. Certainly, a consistent, larger-scale flight training program that successfully limits costs and time-to-completion of initial pilot qualifications would have excellent implications for reducing the looming global pilot shortage (AOPA, 2010).

References

Aircraft Owners and Pilots Association (AOPA) (2010) *The Flight Training Experience: A Survey of Students, Pilots and Instructors*. http://www.aopa.org/

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review, 3*, 385-416.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences. (2nd ed.)* Hillsdale, NJ: Erlbaum.

Dattel, A. R., Kossuth, L., Sheehan, C. C., & Green, H. J. (2013). *Poster presented at the 84th Annual Meeting of the Eastern Psychological Association*. New York, NY.

Dattel, A. R., Durso, F. T., & Bédard, R. (2009). Procedural or conceptual training: Which is better for teaching novice pilots landings and traffic patterns? *Proceedings of the 53rd*

Annual Meeting of the Human Factors and Ergonomics Society. San Antonio, TX.
Drago-Severson, El; Helsing, D; Kegan, R.; Popp, N; Broderick, M; and Portnow, K. (2001). The Power of a Cohort and Collaborative Groups. *Focus on Basics 5, Issue B (October 2001): 15-22*. http://www.eric.ed.gov/PDFS/ED508685.pdf
FAA(2008). Aviation Instructor's Handbook. Retrieved from http://www.faa.gov/about/office_org/headquarters_/avs/offices/afs/afs600
FAA (2008) Advisory Circular *(AC) 00.2-15*. Retrieved from www.faa.gov.
Fenwick, T. (2001). *Experiential learning: A theoretical critique from five perspectives.* Columbus: Ohio State University.
Gopher, D., Weil, M., Bareket, T. (1994). Transfer of skill from a computer game trainer to flight, *Human Factors, 36,* 387-405.
Gopher, D., Sivan, R., & Iani, C. (2001). Comparing learning curves of experts and novices: A novel approach to the study of simulator effectiveness and fidelity. *Proceedings of the Human Factors and Ergonomics Society. Annual Meeting, 2, 1805*.
Harris, D. (2011) *Human Performance on the Flight Deck*. Brookfield, VT: Ashgate.
Hawkins, F.H. (1997). *Human Factors in Flight. (2nd ed.).* Brookfield, VT: Ashgate.
Hays, R. T., Jacobs, J. W., Prince, C., Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis, *Military Psychology*, *4*, 63-74.
Lubner, M., Adams, R., Hunter, D., Sindoni, R. and Hellman, F. (2003). Risks for aviation accidents or incidents among U.S. pilots by pilot training, experience and exposure. *Twelfth International Symposium on Aviation Psychology*. Dayton, OH.
Ortiz, G. A. (1994). Effectiveness of PC-based flight simulation. *The International Journal of Aviation Psychology*, 4, 285-291.
Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology*, *8*, 197-208.
Taylor, H. L., Lintern, G., Hulin, C. L., Talleur, D. A., Emanuel, T. W., & Phillips, S. I. (1999). Transfer of training effectiveness of a personal computer aviation training device. *The International Journal of Aviation Psychology*, *9*, 319-335.
Wicks, F. (2003). Trial by flyer. *Mechanical Engineering*, 4-10.
Wightman, D. C., & Lavern, G. (1985). Part-task training for tracking and manual control. *Human Factors*, *27*, 267-283.

EVALUATION OF AN EYE TRACKING-BASED ASSESSMENT AND DEBRIEF TOOL FOR TRAINING NEXT GENERATION MULTIROLE TACTICAL AVIATION SKILLS

Meredith Carroll & Glenn Surpris, Design Interactive, Inc., Orlando, FL
Greg Sidor & Winston Bennett, Jr., Air Force Research Lab, Dayton, OH

As tactical aircraft become increasingly complex, pilots' cognitive resources will become increasingly strained, especially as more critical and multifaceted information is presented on Helmet Mounted Displays (HMDs). Therefore, it is critical to ensure training results in pilots learning optimal strategies for operating in this information-rich environment, including appropriate attention allocation between different dynamic, adjustable displays and efficient scan strategies. To achieve this, a performance assessment and debrief system was developed that incorporates eye tracking technology into an HMD-enabled multirole fighter simulation to capture and process gaze data to aid in diagnosing why a pilot error occurred. The system utilizes eye tracking to 1) measure attentional focus and scan patterns, 2) diagnose errors in performance and attention allocation, and 3) display performance and attention allocation summaries and mission replay overlaid with pilot scan patterns. A training effectiveness evaluation of the system was conducted with 14 Air National Guard F-16 pilots at the 180th Fighter Wing in Toledo, OH. The F-16 is a current generation multirole aircraft with complicated displays and a variety of mission data displays available for mission execution. Participants were split into two conditions, including a control group which received debriefings utilizing traditional mission replay tools and an experimental group which received debriefings which utilized pilot scan data presented in conjunction with traditional mission replay tools. Results suggest that debriefs utilizing pilot scan data have the ability to support pilots in more quickly adjusting their scan strategies to those most optimal for performance in future, more data intensive tactical fighter environments. The paper will present the system design, experimental methods and a discussion of the results and implications for training.

According to the Air Force Transformation 2010, "…the ultimate source of air and space combat capability resides in the men and women of the Air Force. …[Our] first priority is ensuring they receive the precise education, training, and professional development necessary to provide a quality edge second to none" (United States Air Force, 2010, p. 6). Within the Department of Defense, the ever increasing amount and complexity of knowledge, skills and abilities (KSAs) demanded of their personnel has created the need to develop and utilize tools to increase the efficiency and effectiveness by which training takes place. This is especially true for next generation tactical aircraft. Next generation multirole fighter aircraft will increasingly use a Helmet Mounted Display (HMD) as a primary instrument and sensor display. This comes in conjunction with a significant increase in duties required by the pilots and has the potential to put incredible strain on the pilot's cognitive resources by exposing him to large amounts of data from disparate sources that can quickly exceed natural cognitive processing limits. It is critical to ensure training results in pilots learning optimal strategies for operating in this information-rich environment, including appropriate attention allocation between the different displays and effective and efficient scan strategies. Key to this may be the integration of pilot scan data into assessment and debrief. Visual attention can provide important insights to the information used in task performance, such as the importance of various features or cues (Raab & Johnson, 2007). Several studies (Raab & Johson, 2007; Jarodzka, Scheiter, Gerjets, & van Gog, 2009; Mello-Thoms et al., 2008; and White, Hutson, & Hutchinson, 1997) have demonstrated that eye tracking can aid in the assessment of perception through measurement of visual attention during observation via gaze, scan path, and fixation data. These measures can provide a means for increasing the granularity of performance feedback and a means by which pilots can understand and adjust their scan strategies. Findings in the literature indicate experts have very well-defined scan patterns (Burgert et al., 2007; Jarodzka et al., 2009; and Kasarskis, Stehwien, Hickox, Artez, & Wickens, 2007). This is the result of experts having developed very strong systematic scan strategies in comparison to novices who have less structured scan patterns. Novice scan patterns are typically influenced most by bottom up processes which draw attention to salient features in the environment (Jarodzka et al., 2009). It was hypothesized that, given their experience, it may prove a challenge for legacy pilots such as F-16 pilots to quickly transition their scan strategies to those most optimal for next generation multirole fighter operations. While next generation fighters have similar tactical missions, there is a great deal more and different information presented on the HMD to support those same tactical operations. These pilots will need to redevelop scan patterns to those more optimal for operations and systems of the new multirole fighter aircraft. Thus it was hypothesized that pilots who

receive training debriefs that incorporate feedback related to scan patterns and attention allocation will more quickly alter their scan strategies to those most optimal for the new aircraft operations and related information displays, compared to pilots who receive more traditional debrief methods based on performance measures and serial mission replays. It was also hypothesized that these changes would lead to greater performance improvements in pilots who received eye tracking-based debrief. An experiment was conducted to test these hypotheses.

Method

Participants

Reserve and active duty Air Force pilots (14 men, M_{age} = 36.5 years, *SD* = 5.8) stationed at the Ohio Air National Guard 180th Fighter Wing in Toledo were recruited by email for voluntary participation in the research. Participants were not compensated in addition to their regular salaries; rather, their participation was scheduled during their normal duty day. All participants were required to have either normal or corrected to normal vision by use of contact lenses only, not glasses. Pilot rank ranged from First Lieutenant to Colonel, with an average of 13.5 years (*SD* = 7.3) of military service and 11 years (*SD* = 6.2) of F-16 experience. Participants completed an average of 1800 flight hours (*SD* = 965.2) in the F-16. Two participants had next generation fighter simulator experience, one with 4 hours of experience and another with 175 hours.

Experimental Design and Measures

The experiment used a between subjects repeated measures design. All participants received a pretest in which they performed a series of tactical engagements followed by a debrief per their condition (eye tracking-based vs. traditional). Then each participant performed two brief tactical training scenarios followed by a debrief per their condition. All participants then received a posttest in which they performed a series of tactical engagements. Dependent Variables measured include Instructor Evaluation of Trainee Performance (Correct ID, Shots Valid, Hits, Survivability, Flow Errors, Switch Errors, Communication Errors, Display Utilization, Target Detection, ID, Employment, Cold Ops, Merge Prep, Battle Damage Assessment, Overall Student Performance), System collected measures (Missile Result and Shooter Loft Angle) and attention allocation/scan strategy measures (# fixation and average fixation duration on high/low priority areas, time spent heads up vs. heads down).

Testbed

The testbed utilized Helmet-Mounted Display ASsessment System for the Evaluation of eSsential Skills (HMD ASSESS), a performance assessment and debrief system that incorporates eye tracking technology into an HMD-enabled next generation fighter simulation to capture and process scan data to aid in diagnosing why a pilot error occurred. HMD ASSESS was integrated with a desktop simulation that allows students to interface with the displays and controls via a flight representative hands on throttle and stick (HOTAS; See Figure 1).

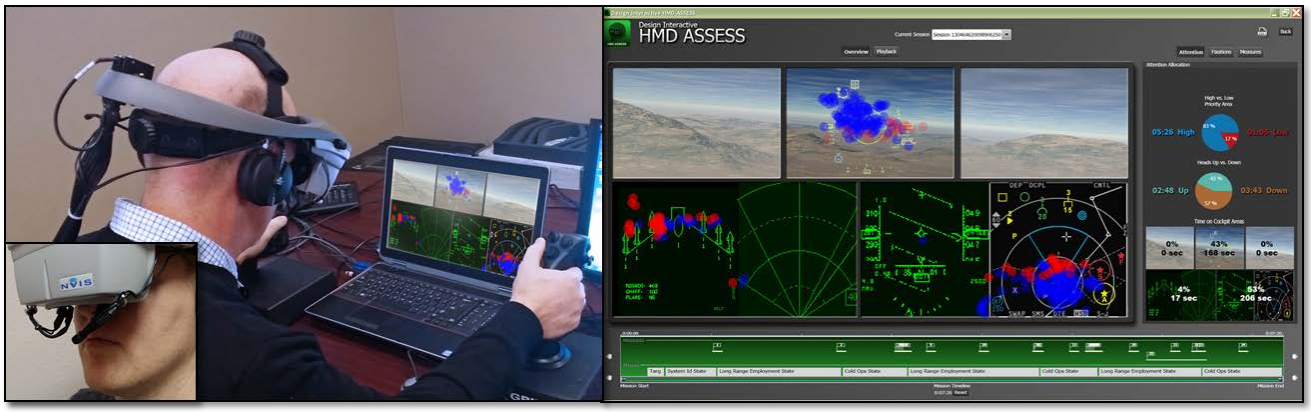


Figure 1. HMD ASSESS Prototype (left) and Instructor Displays (right) with notional iconography and data for the purposes of the paper.

As a pilot flies within the HMD-enabled simulator, the HMD ASSESS measurement component captures gaze tracking measures utilizing the Viewpoint EyeTracker® from Arrington Research Inc. The ViewPoint EyeTracker® integrates miniature high resolution cameras with small infrared lights which allow the determination of pupil location based on corneal reflections. As a pilot monitors different displays within the simulator, the eye

tracker determines the X, Y coordinate associated with each gaze point which is mapped to an area of interest (AOI) (i.e., instrument, display) using a pursuit algorithm which tracks the AOIs as they move across the screen. Behavioral measures from the simulation are also captured, including: 1) simulation events (e.g., flight control inputs, missiles fired, Integrated Caution and Warning System (ICAWS) messages), 2) flight/weapons parameters (e.g., altitude, airspeed, heading, digital maneuvering cue (DMC)), and 3) entity states (e.g., bandits alive, ownship alive). The effectiveness of a pilot's scan and visual attention allocation strategies is then diagnosed by determining where a pilot's attention is fixated and if these fixations intersect with high priority instruments and displays for the task he is currently performing. The high priority displays and instruments for the different segments or phases of the tactical scenarios were determined and defined in a state-engine a priori by fighter pilot subject matter expertise. Additionally, the system determines whether each fixation is associated with the pilot being heads down (i.e., fixating on the panoramic cockpit displays (PCD) panels) or heads up (i.e., fixating out the window). The results are presented in both an interactive multi-level mission timeline overlaid with pilot performance and scan data summaries and an audio/video mission replay with scan data overlays to illustrate the context surrounding errors.

Procedure

All participants were run over the course of six consecutive days at the Ohio Air National Guard facilities. Each experimental session lasted between two and three hours. After reading and signing the informed consent, all participants received familiarization training on the desktop simulator. The familiarization training consisted of a Power Point presentation given by a SME that detailed functionality of the displays and controls necessary to complete the target scenarios. Each participant then donned the HMD and a fitting procedure was performed lasting approximately 5 to 10 minutes. Each participant then completed one familiarization scenario (also approximately 5-10 min) in which they practiced flying the aircraft and targeting enemies. After the familiarization scenario, the eye tracker was adjusted and calibrated for the participant. This procedure involved adjusting the eye-tracking cameras and lighting to get a fix on the participant's pupils. After calibration, participants performed a series of four scenarios containing either two or three adversaries. Each scenario was an air-to-air combat scenario in which the participant had to identify, fix, track and engage enemy aircraft. During each scenario, the instructor played the role of Airborne Warning and Control System (AWACS), providing bullseye picture calls to the trainees. Prior to each scenario, the eye tracker was re-calibrated for those in the experimental condition.

For the control group, the eye tracker was only calibrated for the pretest and posttest scenarios. Eye tracking data was collected for all participants during these scenarios but was not presented to participants in the control condition. Following each scenario, the participant removed the HMD and was debriefed by an instructor on his performance facilitated by the HMD ASSESS debrief system. Participants in the control group received debriefs from the SME who utilized the HMD ASSESS display with eye tracking data disabled. The instructor utilized the HMD ASSESS playback mode to playback the mission, utilizing both the video replay and the values present in the parameters tab to provide feedback. Specifically, participants were provided feedback on the validity (ID and DMC value) of each of their missile deployments as well as the result of their missile deployments. Additionally, participants received feedback on how many red shots were directed towards them; how many flow, switch, and communication errors they made; and tips for avoiding these errors. If the participant did well in any of these categories, they also received positive feedback. For any errors identified, the instructor provided recommendations for improving performance related to these errors in future scenarios. For example, in the control debrief, if a participant lost too much altitude during an out maneuver, the instructor told the participant to note his rate of descent and altitude reading during that phase. Participants in the HMD ASSESS group received a debrief in the exact same format; however, the eye tracking data was enabled. This allowed the instructor to also provide definitive feedback on the root cause of many of the errors. For example, if a participant lost too much altitude during an out maneuver, the instructor was able to see where the participant's visual attention was focused (e.g. tactical situation display), provide feedback on why it was no longer an high priority area to monitor, and instruct them to be mindful of their rate of descent and altitude read outs on future missions. The instructor also used the eye tracking data to validate items he was unable to with the control group. For example, the instructor was able to verify with eye tracking if a participant viewed the Expanded Data Window while making target identifications. The participant also received positive feedback for vigilant scan patterns that incorporated all of the necessary AOIs for a given task. Debriefs were scripted and standardized across both conditions as much as possible. The participant was then fully debriefed with information about the study, including any possible effects the experiment will have and information for access to the results.

Results

Preliminary analyses were conducted to identify potentially confounding variables. One participant was omitted for having 175 hours in a next generation multirole fighter simulation. A multivariate analysis of variance (MANOVA) was performed with a between groups factor of treatment condition (eye tracking-based vs. traditional debrief) for dependent variables (DV) of age, rank, service years, F-16 experience in years and flight hours, F16 training hours – live and simulation, and next generation multirole fighter training hours – live and simulation. There were no significant differences between the two groups in the above demographic variables. A multivariate ANOVA was performed with a between groups IV of treatment condition (eye tracking-based vs. traditional debrief) for DVs of durations associated with each of the four scenarios and each of the four debriefs. There were no significant differences indicating participants in the two training groups received training over approximately the same amount of time.

Scan Data: Next, eye tracking data was analyzed to determine if debrief with scan data led to improved pretest to posttest changes in a trainee's scan strategies compared to the control group. Utilizing 12 of the 14 participants (7 experimental, 5 control; the participant with 175 hours of next generation multirole fighter simulation training was omitted and pretest eye tracking data was lost for one participant), a repeated measures MANOVA was performed with a within groups factor of trial (pretest vs. posttest) and a between groups factor of treatment condition (eye tracking-based vs. traditional debrief). To account for differences resulting from eye tracking data quality and not the treatment condition, calibration quality and eye tracking quality scores were utilized as covariates. As predicted, when comparing pretest to posttest scan data, participants receiving eye tracking-based debriefs altered their scan patterns to focus more on high priority areas as opposed to the control group. These differences were not statistically significant, but trended towards significance with moderately high effect sizes. Specifically, there was an interaction between trial and condition for time spent fixating on high priority areas (F (1, 6) = 3.03, p = .13, η^2= .34), time spent fixating on low priority areas (F (1, 6) = 3.93, p = .09, η^2= .39), number of fixations on high priority areas (F (1, 6) = 2.07, p = .20, η^2= .26),), and number of fixations on low priority areas (F (1, 6) = 1.59, p = .25, η^2= .21). From pretest to posttest, participants receiving eye tracking-based debriefs increased the time spent fixating and their number of fixations on high priority areas while decreasing the time spent fixating and their number of fixations on low priority areas. The control group displayed inverse patterns decreasing time and fixations on high priority areas while increasing time and fixations on low priority areas (see Figure 2). Additionally, there was an interaction between trial and condition that trended towards significance for time spent fixating heads down (F (1, 6) = 3.74, p = .10, η^2= .38), with participants receiving eye tracking-based debriefs decreasing the time they were heads down from pretest to posttest and control participants increasing the time they spent heads down from pretest to posttest.

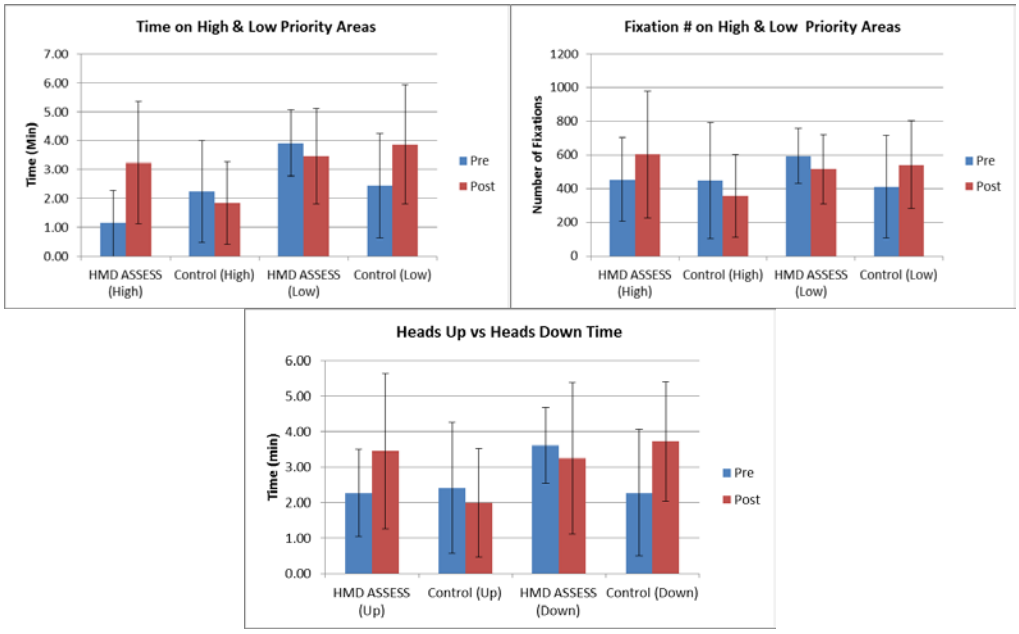


Figure 1. Scan data results.

Performance: These differences in scan strategies did not translate to performance differences. Instructor evaluation measures were analyzed to determine if debriefs with scan data led to improved pretest to posttest changes in a pilot's performance compared to the control group. Utilizing 13 of the 14 participants (8 experimental, 5 control; the participant with 175 hours of next generation fighter simulation training was omitted), a repeated measures MANOVA was performed with a within groups factor of trial (pretest vs. posttest) and a between groups factor of treatment condition (eye tracking-based vs. traditional debrief). There was a significant trial effect for multiple measures with both groups showing improvement from pretest to posttest in overall mission performance (F (1, 11) = 6.04, p = .02, η^2= .39), number of valid shots (F (1, 11) = 12.49, p = .01, η^2= .53), location and utilization of controls and displays (F (1, 11) = 24.5, p = .00, η^2= .69), and hits (F (1, 11) = 9.59, p = .01, η^2= .47 (see Figure 3). There were no between group differences or interactions, suggesting that there were not differential training performance effects between the groups. No other instructor-based measures showed significant differences.

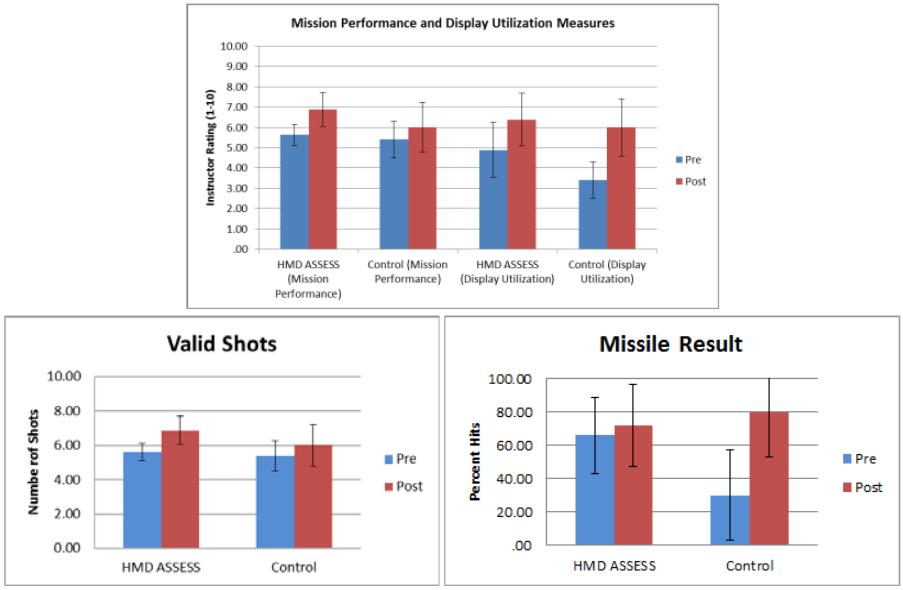


Figure 3. Instructor evaluation measures and system collected measures.

Similar results were seen in system collected performance measures. Due to limitations in the amount and accuracy of data being published by the desktop simulator on the Distributed Interactive Simulation network, only two measures could be calculated across all missiles fired in the pretest and posttest scenarios: Missile Result (0 = miss, 1 = hit) and Shooter Loft Angle (in degrees). These measures were analyzed to determine if debrief with scan data led to improved pretest to posttest changes in a trainee's performance compared to the control group. Utilizing 13 of the 14 participants (8 experimental, 5 control; the participant with 175 hours of next generation fighter simulation training was omitted), a repeated measures multivariate ANOVA was performed with a within groups IV of trial (pretest vs. posttest) and a between groups IV of treatment condition (eye tracking-based vs. traditional debrief). There was a significant trial effect for Result (F (1, 11) = 12.69, p = .01, η^2= .54), with both groups improving from pretest to posttest. There was also a significant interaction for Result (F (1, 11) = 7.68, p = .02, η^2= .41), with control group participants showing greater increases from pretest to posttest, although the control group had a significantly lower pretest score, with posttests being only slightly higher than the experimental group (see Figure 3). There were no between group differences. There were also no significant within or between group differences with respect to Shooter Loft Angle.

Discussion

These results provide promising insight into the potential for eye tracking data to improve the training effectiveness of current tactical aviation debriefs. The results suggest that debrief utilizing scan data has the ability to support pilots in more quickly adjusting their scan strategies to those most optimal for performance. Pilots receiving eye tracking-based debriefs increased the time and number of fixations on high priority areas and the time spent heads up, patterns that were not seen in the control group. Although these did not translate into greater performance improvements over traditional debrief methods, this may be due to the very limited amount of training received or the sensitivity of the performance measures collected. Pilots were trained on a total of four brief

scenarios including the pretest and posttest. This resulted in approximately an hour of training time including debrief. This may not have been enough time for these changes in scan strategies to translate to actual performance improvements. Results of this study must be considered cautiously due to several limitations of the study. There were a very limited number of participants in the study due to the availability of the pilots to support and the time it took to complete the study. The study was also not a blind study, as the instructor had to know which condition in order to give the appropriate debrief. Further, not only was there an unequal number of participants, but treatment group assignments were not random. Those participants in the beginning who had poor eye tracking calibration data were placed in the control group as debrief with scan data was not possible. As the week progressed and methods for fitting the HMD and adjusting the eye tracking cameras resulted in greater proportions of participants with eye tracking data, participants were eventually split up between the groups randomly. Finally, the adversary behaviors in the pretest and posttest scenarios were not consistent. Despite designing the scenarios such that the enemy behavior was scripted to react a particular way, this was not always the case and was not something the system operators could control.

Conclusion

Eye tracking technology provides the ability to go beyond measuring observable performance outcomes (e.g., did the trainees effectively engage the target?), making feasible the measurement of unobservable perceptual and cognitive processes such as pilot scan and attention allocation. These measures aid in the identification of where in the piloting process breakdowns occur (e.g., pilot failed to monitor certain shot parameters in order to identify optimal time to shoot). This facilitates more tailored feedback, potentially leading to significant improvements in training effectiveness and efficiency. Further research is needed in this area, however, this study provides promising data in support of utilizing eye-tracking assessment and debriefs to train next generation multirole tactical aviation skills.

Acknowledgements

This work was funded by the Air Force Research Lab (AFRL) under contract # FA8650-12-C-6303. The views and conclusions contained in this presentation are that of the authors and should not be interpreted as representing the official viewpoints of AFRL or the US government.

References

Burgert, O., Örn, V., Velichkovsky, B. M., Gessat, M., Joos, M., Strauß, G., ... & Hertel, I. (2007, March). Evaluation of perception performance in neck dissection planning using eye tracking and attention landscapes. In *Medical imaging* (pp. 65150B-65150B). International Society for Optics and Photonics.

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146-154.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001, March). Comparison of expert and novice scan behaviors during VFR flight. In *Proceedings of the 11th International Symposium on Aviation Psychology* (pp. 1-6).

Mello-Thoms, C., Ganott, M., Sumkin, J., Hakim, C., Britton, C., Wallace, L., & Hardesty, L. (2008). Different Search Patterns and Similar Decision Outcomes: How Can Experts Agree in the Decisions They Make When Reading Digital Mammograms? In *Digital mammography* (pp. 212-219). Springer Berlin Heidelberg.

Raab, M., & Johnson, J. G. (2007). Expertise-based differences in search and option-generation strategies. *Journal of Experimental Psychology: Applied*, *13*(3), 158.

United States Air Force. (2010). *The Edge Air Force Transformation 2010*. Retrieved from http://permanent.access.gpo.gov/lps40477/edgeweb.pdf

White Jr, K. P., Hutson, T. L., & Hutchinson, T. E. (1997). Modeling human eye behavior during mammographic scanning: Preliminary results. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *27*(4), 494-505.

A MULTI-YEAR STUDY OF THE SAFETY AND TRAINING IMPACTS OF INTRODUCING THE LIVE, VIRTUAL, CONSTRUCTIVE TRAINING STRATEGY INTO NAVY AIR COMBAT

Sarah Sherwood[a]; Kelly Neville[a]; Angus L. M. Thom McLean, III[b]; Jessica Cruit[a]; Katherine Kaste[c]; Melissa Walwanis[c]; Amy Bolton[d]

Embry-Riddle Aeronautical University[a] Daytona Beach, FL

Rockwell Collins[b] Cedar Rapids, IA

Naval Air Warfare Center Training Systems Division[c] Orlando, FL

Office of Naval Research[d] Arlington, VA

In the Navy's proposed *Live, Virtual, Constructive (LVC)* training system, virtual entities that represent pilots in flight simulators and computer-generated constructive entities will be injected into the cockpits of F/A-18 aircraft during live-flight training. The Navy expects LVC to ameliorate many of the economic and environmental impacts of live-flight training and to support future training requirements. However, the potential impact of LVC on training effectiveness and safety is not completely understood. While the naval air combat training system is notably robust, its inherent complexity precludes a straightforward analysis of potential hazards and mitigations. Two years ago, researchers began to identify and assess the numerous human-technology interactions that will characterize the future LVC training system, most notably those that could lead to hazardous LVC training situations. They employed cognitive task analysis methods to interview fighter pilots, F/A-18 weapons systems officers, range training officers, and modeling and simulation subject-matter experts (SMEs). The present paper is a follow-up to their 2013 ISAP presentation on potential LVC training hazards and mitigations identified during Cycle I of their iterative research. An additional two-year cycle of data collection, analysis, and SME review has led to an enriched understanding of the potential training hazards and benefits that could arise from interactions among training system elements. The long-term goal of this research is to develop requirements for the Navy's LVC training system that will enable its safe implementation and eventual optimization. With that goal in mind, the findings from Cycle II are presented here.

A number of economic and environmental challenges threaten the attainment and maintenance of air combat readiness. To mitigate these challenges, the Navy seeks to create a *Live, Virtual, Constructive (LVC)* training system to train F/A-18 pilots. The proposed system leverages advances in high-speed datalink technology to inject virtual tracks (tracks representing aircraft flown by pilots in simulators) and constructive tracks (tracks representing computer-generated aircraft) onto the radar and sensor system displays of live F/A-18 aircraft. In the future, the Navy plans to extend this capability to other platforms.

The naval air combat training system is a highly complex and nuanced product of decades of evolution. The introduction of LVC, a complex simulation system, into the already complex Navy air combat training system carries a heightened risk of induction of unanticipated interactions between elements of both systems, which could result in unforeseen and potentially hazardous emergent system behavior. Two years ago, Office of Naval Research-funded research began to identify and assess the numerous human-technology interactions that will characterize the Navy's future LVC training system, most notably those that could lead to hazardous LVC training situations. The result of the first yearlong cycle of data collection, a preliminary list of potential LVC training hazards, was presented at the 2013 International Symposium on Aviation Psychology (Sherwood et al., 2013). An additional two-year cycle of data collection, analysis, and subject-matter expert (SME) review has led to an enriched understanding of the potential training hazards and benefits that could arise from interactions among training system elements. This paper presents the results of the Cycle II analysis.

Method

Participants

Researchers interviewed 31 participants over the course of the study. The participants' diverse backgrounds provided insight into the potential impacts of LVC from a variety of perspectives.

Cycle I (*n* = 22). The Cycle I participants included twelve Navy pilots, one Marine Corps pilot, one active-duty Air National Guard pilot, and six retired Air Force and Navy pilots [including one who also served as an Naval Flight Officer (NFO)]. Other participants included one active-duty NFO and one active-duty F/A-18 weapons system officer (WSO). The researchers also interviewed two military modeling and simulation (M&S) experts. However, the M&S experts are not included in the participant count.

Cycle II (*n* = 9). The Cycle II pilot participants included six Naval Strike and Air Warfare Center (NSAWC) instructor pilots, one of whom was also an experienced Range Training Officer (RTO), and one adversary pilot. The participants reported an average of 2,330 flight hours in high-performance jet aircraft and flew one or more of the following fighter platforms: the F/A-18 (n = 6), F-16 (n = 1), F-14 (n = 1), F-5 (n = 2), and Hawker Hunter (n = 1). Other participants included a NSAWC WSO instructor and an Air Intercept Controller (AIC) who also has experience as an RTO.

Data Collection

Researchers briefed participants on the LVC training concept, the Navy's goal of using the system to improve training efficiency and to virtually extend training ranges, and the purpose of the interview. The purpose of the interview was to identify potential *training concerns* (potential negative interactions between LVC technology and current training practices), *training benefits* (opportunities for LVC to enhance or supplement current training), and *training hazards* (potential impacts of LVC on training safety if it were implemented in the existing air combat training system without mitigation). Researchers asked participants to discuss potential strategies for the mitigation of identified hazards and training concerns and to suggest ways in which LVC could optimize training. The researchers obtained permission to audio record and transcribe 21 of the Cycle I interviews. Audio recording was not permitted during any of the Cycle II interviews. However, researchers took detailed notes and gave interviewees the opportunity to review the notes and to offer corrections.

Cycle I. During Cycle I, researchers interviewed the majority of participants using an adaptation of the Critical Decision Method (Klein, Calderwood, & MacGregor, 1989). The thematic analysis of the Cycle I interview data yielded a list of potential LVC training benefits, training concerns, and—the focus of this paper—training hazards. Three researchers then jointly organized each hazard into overarching hazard categories according to the fundamental underlying interviewee concerns that they represented. All identified potential hazards were subjected to a 2.5-hour group critique and to a review by two Navy air combat experts, who commented and elaborated on the presented hazards and their supporting data. For a more thorough account of Cycle I data collection and analysis, see Sherwood et al. (2014a, 2014b).

Cycle II. During Cycle II, researchers used a semi-structured interview approach to elicit knowledge from experienced Navy air combat pilots and exercise management personnel. The goal of this data collection was to expand and refine the list of potential training benefits, concerns, and hazards identified during Cycle I.

Researchers interviewed Cycle II participants in sessions that ranged from 30 minutes to 1.5 hours. During each session, a researcher and a naval air combat SME posed a series of questions to a participant based on previously identified training hazards, concerns, and benefits, specifically in relation to training fidelity. However, the findings of the fidelity survey are beyond the scope of this paper except for instances where low fidelity could give rise to hazardous training situations.

Data Analysis

Coding of Cycle II interview notes. While researchers coded the Cycle I interview data using a bottom-up approach, they coded the Cycle II interview data using a top-down approach with respect to the training concerns, benefits, and hazards that emerged from the Cycle I interview data. Two researchers reviewed the interview notes to identify statements made by participants about the previously identified hazards and their mitigation. The researchers then assigned these interviewee statements, referred to as *data extracts*, to their appropriate hazard codes. The wording of the hazards and the organization of the hazard categories evolved over the course of Cycle II as a result of this elaboration. Table 1 lists these changes.

Safety risk level evaluation. After the coding of all the data extracts addressing potential LVC training hazards, the two researchers used the severity and probability scales of the Navy's Operational Risk Management (ORM) system to independently assess the safety risk level associated with each hazard (Department of the Navy, 2010). The researchers based their assessments on the content of the data extracts.

Researchers gave severity and probability ratings for each hazard on four-point scales. They then merged the ratings in accordance with the ORM process to obtain a risk assessment code (RAC), which represents the level of risk associated with a hazard and is expressed as a single Arabic number from 1 (critical) to 5 (negligible) (Department of the Navy, 2010). Researchers calculated three RACs for each hazard: the *baseline risk level* (the risk level of the hazard in the current live air combat training environment), the *unmitigated risk level* (the risk level of the unmitigated hazard in the LVC training environment), and the *residual risk level* (the risk level of the mitigated hazard in the LVC training environment). Researchers assessed inter-rater agreement using a linear weighted kappa, which indicated a very good level of agreement ($\kappa = 0.81$, 95% CI: [0.73, 0.89]). However, the researchers still met to reconcile any disagreements on the baseline, unmitigated, and residual risk levels. They assigned risk levels to each hazard category based on the risk levels of its component hazards. When a hazard category contained hazards of more than one risk level, researchers assigned the category the highest applicable risk level.

Results and Discussion

Researchers grouped 49 identified hazards into five overarching hazard categories: (1) Within Visual Range (WVR) Operations, (2) Human-Machine Interface (HMI) in the Cockpit, (3) Fidelity, (4) Exercise Management, and (5) Cockpit Technology. Table 1 compares the results of the Cycle I and Cycle II hazard analyses, and a summary of each Cycle II hazard category follows. Each summary includes a description of the specific hazards within a category and, as applicable, relevant mitigations suggested by interview data.

Table 1.
Results of the Cycle I and Cycle II Hazard Risk Level Analyses.

Cycle II Hazard Category	Projected Change in Hazard Exposure (Baseline vs. Unmitigated)	Unmitigated RAC	Residual RAC (Mitigated to Baseline?)	Corresponding Cycle I Hazard Categories	Projected Change in Hazard Exposure (Baseline vs. Unmitigated)	Unmitigated RAC
WVR Operations	Increase	1	1 (No)	Unseen Aircraft WVR	Increase	1
HMI in the Cockpit	Increase	1	2 (No)	Inadequate Support for HMI	Increase	2
Fidelity (Live, VC, Display, Environment, and General)	Increase	1	3 (Yes)	Reduced Big Picture Awareness	Increase	3
				Unexpected VC Behavior	No Change	2
				Complacency / Risk-Taking	No Change	2
				Negative Transfer of Training to the Operational Environment	No Change	4
Exercise Management	No Change	1	1 (N/A)	Inadequate Support for Exercise Management	No Change	3
Cockpit Technology	No Change	3	3 (N/A)	Inadequate Support for HMI	Increase	2

Within Visual Range (WVR) Operations

The opinions of interviewed pilots diverged on whether VC aircraft should go to the merge (i.e., dogfight) with live aircraft. Whether future air combat training rules permit VC aircraft to merge with live aircraft remains to be seen, but either case presents its own safety challenges to be mitigated.

Some pilots opposed to the use of VC aircraft WVR asserted that it would provide poor training, while others said that it would increase the risk of midair collision if a live pilot were to merge with a mixed group (a group containing both live and VC tracks). If future LVC training rules do not permit VC aircraft to operate WVR, VC tracks must be safely cleared from radar and cockpit sensor system displays before the merge. Since VC tracks will not likely be differentiated from live tracks on pilot displays, a sudden, unexpected disappearance of VC tracks could induce pilot confusion and attentional tunneling because there are reasons that live aircraft disappear from

radar (e.g., maneuvering, accidents, and electronic attacks). Thus, pilots could end up tunneling on their radar or their out-the-window view while trying to locate the dropped "live" track that is not really there. For this reason, some interviewees suggested that, if exercise management personnel clear VC tracks (from all cockpit displays) before the merge, their action should be accompanied by a comm call indicating that VC tracks have been cleared. Interviewees also suggested that, instead of disappearing, VC tracks should turn away before they reach 10–30 nm.

Interviewees who supported the use of VC aircraft WVR felt that going to the merge with aircraft that one cannot see is good training, as pilots cannot always visually identify (VID) the F-5s flown by live adversary pilots. Therefore, while pilot failure to VID a VC track might result in attentional tunneling, the associated risks are no different from what pilots currently face in live air combat training exercises. Thus, many interviewees believed that a "merging with VC" radio call from the RTO would be all the mitigation that is needed and that, otherwise, the Navy's current air combat training rules and procedures are resilient enough to allow for the safe operation of VC aircraft WVR of live aircraft. However, some pro-merge pilots tempered their support with the suggestion that mixed groups should be kept from the merge because of a perceived increase in the risk of pilots losing track of live aircraft and the resulting increased potential for collision. Merging with VC-only groups mitigates this risk. According to one participant, skills that could be practiced in WVR with a VC entity include merge geometry, look up, look down, time out, and weapons employment. He noted that, while a pilot would lack visual look out and some merge look out, "the goods outweigh the 'others.'"

Human-Machine Interface (HMI) in the Cockpit

To safely conduct LVC air combat training, a number of cockpit HMI design challenges must be resolved. One challenge is to design a solution to address situations in which a pilot must know which aircraft are live. For example, they may become cognitively overloaded and lose situation awareness (SA), or a training incident or accident may occur. Based on participant input, the solution could take the form of either an ever-present, unique VC symbol set or a pilot-controlled switch to clear VC aircraft from cockpit displays. However, most interviewees opposed the artificiality of unique VC symbols (often citing the aphorism that one must "train as you fight"). The use of a pilot-controlled switch to remove VC tracks temporarily met wider support, perhaps in part because the removal of tracks from cockpit displays is not a significant departure from the current training procedure. "If a young guy is overwhelmed, he can currently do this," noted one interviewee. "Leaving the VC entities off for a period of time is okay. We do that currently with the MIDS/Link-16 display." Moreover, since the LVC-off switch would be intended for use by "overwhelmed" pilots or during emergency situations, it must be designed so that a single-seat F/A-18 pilot can operate it without increasing his or her cognitive load or further degrading his or her SA. More specifically, the switch must be readily available and consistently function in the same way, regardless of aircraft mode.

Interviewees expressed an additional concern: ensuring clear indications of the operational status of the live and VC track data feed(s), particularly if VC tracks are sent through a separate data feed. In this case, it is possible that a pilot's Link-16 (i.e., the live track feed) could fail and the VC tracks populating the displays could camouflage the problem. However, this scenario is very unlikely, as pilots already have a clear indication of a lost Link-16. A similarly clear indication of status of the VC track feed should be provided so that pilots' SA is not disrupted by VC tracks unexpectedly disappearing and reappearing. Finally, the LVC-enabled cockpit should support LVC mode awareness. Pilots need to be aware of whether their systems are in LVC training mode or operational mode. Although unlikely, it is possible that pilots could mistakenly enter operational mode and thus could misinterpret information, treat a live aircraft as VC, or take the wrong action for a given context.

Fidelity

Potential fidelity concerns in the LVC environment include the disparate speeds and capabilities of simulated, state-of-the-art adversaries and live F-5 adversaries; the potential for VC tracks to be fed into cockpit displays in ways that do not realistically simulate sensor system noise, ambiguity, and inaccuracy; imperfect correlation of VC behavior across sensor system displays or displays that are out of sync with a pilot's offensive and defensive actions against a VC adversary (e.g., failure to accurately portray broken radar lock); and unexpected VC track behavior.

During training, exercise participants could become cognitively fatigued because of the extra work required to perceive, synthesize, and make sense of unrealistic and imperfectly correlated sensor system data. Fatigued pilots are more likely to make mistakes, lose SA, or exhibit reduced flight discipline. Furthermore, unexpected behavior by VC entities could distract, confuse, or cause a pilot to maneuver reactively. All these factors increase the possibility of a midair collision with a wingman or other aircraft. In the case of "cognitive overload caused by LVC training artificialities, a pilot may use the proposed "clear VC" cockpit switch to regain his or her SA. However, interviewees proposed a number of solutions to prevent pilot SA from being degraded in the first place. For example, the disparity between advanced VC adversaries and live F-5s could be masked or avoided in multiple ways. First, VC tracks could be overlaid on live tracks so that pilots receive both VC indications and live electronic attack (EA) capabilities. Second, VC and live tracks could be blended in the tactical combat training system (TCTS) and then fed to the live aircraft. In this case, radar would be turned off when LVC is on and adversaries are beyond visual range (BVR). Finally, if other interventions do not prove sufficient to compensate for VC artificialities, VC tracks could turn away before the merge, and/or merging with mixed sections could be prohibited.

In addition, LVC artificialities must not impair skills or produce mindset changes that can be carried into combat. Poor VC adversary and sensor system fidelity could ultimately result in pilots and E-2 NFOs going into combat without sufficient training on enemy tactics and perception and coordinated interpretation of multiple sensor system displays, leading to preventable casualties. To increase LVC sensor system fidelity, the LVC system should allow instructors to vary radar strength, clutter, and EA. Furthermore, it should allow instructors to tie VC tracks to live tracks. The live tracks would be detected and tracked by radar, and later, after the adversary section was much closer, the VC entities would break out as separate tracks.

Exercise Management

The addition of VC aircraft to the air combat training environment will increase the workload and complexity of exercise management. The high workload associated with the control of less-intelligent constructive entities is of particular concern, and the task of managing these tracks will likely need to be divided amongst a number of exercise managers depending on the size of the exercise and the availability of work support tools to make the task easier. Otherwise, if overextended, managers could fail to notice and mitigate hazardous conditions or make changes to VC adversary behaviors or plans that are not sufficiently thought out, potentially causing confusion for live pilots. Moreover, additional exercise management roles might need to be created to control the VC entities effectively and to ensure that training rule violations and hazardous situations are detected by managers just as easily as in current, all-live training exercises. The key problem, however, with the creation of additional exercise management roles may be finding the balance between having enough managers to distribute the workload effectively and having too many managers so that it becomes impossible to maintain intra-team communication during time-pressured situations, thus affecting team SA and reducing the managers' ability to identify and address potentially hazardous training situations. As one interviewee put it, "There is a point of diminishing returns. Someone—one person—has to be the bubble and decision-making authority. Too many people is not good."

Cockpit Technology

The LVC architecture must not create a security weakness in network or onboard systems. Furthermore, there are various potential hazards associated with the LVC architecture that depend on its level of integration with the F/A-18 Operational Flight Program (OFP). Architectural options include (a) feeding VC entity data to separate LVC system displays, (b) feeding VC entity data through the same datalink as live track data to appear together on existing cockpit displays, and (c) feeding live track data to LVC training system computers, which, in turn, would feed VC and live tracks to existing cockpit displays.

If the LVC architecture is fully integrated with the OFP, the fusing of VC entities into track files carries the risk that the multi-sensor integration (MSI) function might mis-correlate live track radar data with VC track data. Although unlikely, this includes the possibility that exercise interlopers (e.g., white air and other "non-squawkers") could become correlated with a VC track occupying the same airspace. "F-5s fly around in the FRTC (Fallon Range Training Complex) without radars all the time. I won't see you anyway," said one interviewee, implying that interlopers are not a significant concern. This pilot went on to suggest that, "if feeding a TCTS feed, [exercise managers] could generate… a 'safety' track file [that would represent non-participants in the training arena]. A good guy on the console (the RTO) will handle it." Conversely, if LVC is not fully integrated into OFP displays, possible

hazards could include the increased risk of training accidents because of (a) the increase in cognitive workload required to reconcile OFP displays with LVC displays; (b) the potential for pilots to focus on the LVC displays, thus missing safety-critical data on the OFP displays; and (c) the potential for pilots to use the separate displays to game the system (e.g., to shoot live aircraft first), which might encourage pilot complacency and risk-taking.

In addition, the LVC architecture must prohibit the inadvertent launch of live weapons when the system is in LVC mode. The OFP should prevent pilots from accessing LVC functionality if active weapons are loaded. However, the datalink(s) used during LVC exercises could provide an additional safety net against accidental live fire. "Certain datalinks don't let you send data unless it's valid data. So data associated with intent to fire when an aircraft has live ordnance on board could potentially be blocked, preventing a live launch," according to an air intercept controller (AIC) interviewee.

Discussion

In the coming year, additional data collection, analysis, and SME-review will take place as needed to further elaborate upon the identified hazards. Using risk analysis to prioritize potential LVC safety hazards, researchers created a roadmap created for these efforts, which will focus on identifying the most promising mitigation strategies. The mitigation of potential LVC training hazards associated with WVR operations, cockpit HMI, and training fidelity should be given priority, as they represent the largest potential increases in pilot risk exposure relative to current all-live training. In addition, further research is needed on the potential effects of LVC on exercise management team SA, communication, and coordination to determine what work support tools and/or additional roles should be created to mitigate the increase in workload caused by the addition of VC entities. The long-term goal of this research is the development of requirements for the Navy's LVC training system that will enable its safe implementation and eventual optimization.

Acknowledgements

This research was sponsored by the Office of Naval Research (ONR). The authors wish to thank the interviewees, SMEs, and their extended research team members at Rockwell Collins and the Operator Performance Lab (OPL) at the University of Iowa. The views expressed in this paper are those of the authors and do not represent the official views of the organizations with which they are affiliated.

References

Department of the Navy, Office of the Chief of Naval Operations. (2010). *OPNAV instruction 3500.39C* (OPNAVINST 3500.39C). Washington, DC: Department of the Navy.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical Decision Method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3), 462–472.

Sherwood, S., Blickensderfer, B., Neville, K., Jimenez, C., Fowlkes, J., & Walwanis, M. (2013, May). *Live-virtual-constructive training: Safety concerns and mitigations.* Paper presented at the International Symposium on Aviation Psychology, Dayton, OH.

Sherwood, S., Neville, K., Ashlock, B., Mooney, J., Walwanis, M., Bolton, A., & Martin, T. (2014a). Envisioned world research: Guiding the design of live-virtual-constructive training technology and its integration into Navy air combat training. *Proceedings of the 2014 Annual Meeting of the Human Factors and Ergonomics Society.* Thousand Oaks, CA: Sage Publishing.

Sherwood, S., Neville, K., Blickensderfer, B., McLean, A., Walwanis, M., & Bolton, A. M. (2014b). *The identification and assessment of hazards associated with the introduction of new technology into a complex air combat training system.* Manuscript in preparation.

TRAINING MANNED-UNMANNED TEAMING SKILLS IN ARMY AVIATION

Susan R. Flaherty
Aptima, Inc.
Sierra Vista, AZ
Martin L. Bink
U.S. Army Research Institute
Fort Benning, GA

Current Army Aviation combat operations utilize an employment strategy that teams a rotary wing aircraft with an unmanned aircraft system (UAS), thereby leveraging strategic advantages of each aircraft's unique capabilities, endurance, and payloads. Clear and effective communications between the airborne helicopter pilot and the ground-based UAS operator are critical for successful Manned-Unmanned Teaming (MUM-T) missions. Previous studies have recommended additional training in tactical communications for the UAS payload operator in order to support precision and timeliness in teaming engagements. In accordance with Army Learning Model 2015, an engaging, skills-adaptive, computer game was developed to train critical MUM-T skills for the UAS payload operator. The training game emphasizes the UAS operator's concise tactical communications exercised in doctrinally correct MUM-T mission scenarios with immediate Soldier feedback. Game players are initially exposed to scripted MUM-T training mission scenarios that culminate in a freeplay mission campaign. Performance measures focus on both accomplishment of mission objectives and accurate tactical communications protocol. Game players are given individual scorecards that display skill advancement and potential remediation for knowledge and skill deficiencies. An initial user assessment is scheduled for game and feature refinement. Future implementation of aggregated soldier data will be explored with UAS course instructors. Results from the user assessments and recommendations on tactical communications training will be disseminated when available.

Over the past decade, the primary role for U.S. Army unmanned aircraft systems (UAS) has evolved from simply surveillance and intelligence gathering to participating in tactical scout-reconnaissance missions. The complexity of the evolving UAS role requires the UAS operators not only to develop tactical skills (Stewart, Bink, Barker, Tremlett, & Price, 2011), but also to develop the ability to directly interact with ground units and other aviation assets (e.g., AH-64 attack helicopter). The development of these tactical and communication skills represents a training challenge for UAS operators (Stewart, et al.), especially as manned-unmanned aviation teaming (MUM-T) is formalized into operational requirements. This paper describes recent research and development conducted by the U.S. Army Research Institute (ARI) to train MUM-T skills for UAS operators.

MUM-T is a special case of aviation scout-reconnaissance operations involving a UAS and armed scout or attack helicopter as a tactical team. Each type of air platform has its own complementary asset and sensor advantages, and MUM-T exploits both. The UAS typically

operates above 8,000 feet (above ground level) while the helicopter, in order to evade detection and hide from the enemy, seldom exceeds 1,000 feet on the typical scout-reconnaissance mission. Whereas the helicopter has limited endurance and must return to base after 45 minutes to one hour, the UAS can remain aloft for 6 hours or more depending on airframe. For these reasons, the UAS becomes a persistent combat aviation capability with a very different vantage point. For UAS missions, prolonged time over target yields greater likelihood to detect, identify, and report a threat than shorter rotary wing missions. As part of the Army's Aviation Restructuring Initiative (US Army Aviation Center of Excellence, 2008), MUM-T operations are being formalized in Attack Reconnaissance Battalions. Organic manned-unmanned units are being constructed, whereby the UAS platform assumes the role of scout helicopter.

In previous ARI research, communication skills were identified as the most critical among the set of skills required for UAS operators to effectively conduct MUM-T operations (Sticha, Howse, Stewart, Conzelman, & Thibodeaux, 2012). At present, UAS operators lack the formal training and common terminology that allows them to communicate in a tactical mission with manned aviators. Communication skills such as target handover and battle damage assessment require tactical knowledge as well as knowledge of proper reporting formats. There are additional communication dynamics that contribute to effective tactical communications such as timing and brevity (Stewart, Bink, Dean, & Zeidman, 2015). To address this important training gap, a training game, was developed that incorporated empirically-based training approaches for critical MUM-T skills. The game was developed in collaboration with Night Vision and Electronic Sensors Directorate as part of the Night Vision Tactical Training (NVTT) suite. The game is called NVTT-Shadow.

NVTT-Shadow Overview

NVTT-Shadow is intended to train Soldiers enrolled in Advanced Individual Training for U.S. Army 15W UAS Operators at Ft. Huachuca, AZ. Introduction to the game may occur after initial Phase 1 Common Core Aviation ground school and prior to Phase 2 aircraft-specific instruction. NVTT-Shadow is designed to augment current curriculum instruction of MUM-T-related tasks with specific focus on accurate and timely tactical communications. Gameplay is intended reinforce instruction given. For example, NVTT-Shadow can be used when there is downtime or delays in flight line training.

NVTT-Shadow is aimed at developing the tactical communication skills of the payload operator through accomplishing outlined mission objectives in routine UAS missions (e.g., Route reconnaissance, Convoy Security). Coordination with external agencies and manned aircraft scout/attack teams are emphasized. The tasks presented in the game were derived from U.S Army doctrine (e.g., Department of the Army 2009, 2014) as well as critical MUM-T skills identified in research (Sticha, et al., 2011) and other Army publications (United States Army Aviation Center of Excellence, 2014). Examples of critical MUM-T tasks include the following:

- Utilize standardized radio communication and signal operating procedures
- Provide accurate description of target to support target selection
- Perform battle damage assessment
- Conduct call for direct fires

- Utilize standard execution commands to initiate Deliberate Attack
- Call for and Adjust Indirect Fire
- Perform Target Hand Over to an Attack Helicopter
- Trasmit a Tactical Report
- Request and Adjust Indirect Fire
- Request Close Combat Attack (AH-64 cannon and rockets)

NVTT-Shadow System Description

The NVTT-Shadow software runs on two rack-mounted software servers and is displayed on a 22-in HD 1920x1080dpi monitor. Necessary controls for gameplay are a joystick, keyboard, and monitor. The laptop can be used by a game procter/instructor as a duplicate of the soldier's game display. The ability to monitor correct transcription of speech-to-text is also accessible via voice menu software residing on the laptop. The integration of client and server-side components are synchronized by a web server, which resides in a data center providing links between the User Interface and simulation services. Game software is comprised of the following three integrated simulation services that function across a Distributed Interactive Simulation (DIS) to generate the mission training environment: 1) One Semi-Automated Forces (OneSAF), 2) AVSim Flight Model, and 3) Night Vision Image Generator (NVIG). Where applicable, government-owned or commercial-off-the-shelf simulation tools were leveraged to advance scenario bulding capabilities.

The game uses intelligent, speech-enabled entities as a means to automate training and create an interactive system without requiring human role-players. Underlying technologies include the following: DIS radio simulation, intelligent behavior modeling, and a customized natural language processing (NLP) pipeline. This combination of technologies enables simulation entities to rapidly transcribe, interpret, and respond to trainee radio transmissions. A prescribed set of tactical communications protocol are supported. The NLP pipeline uses a combination of rule-based and machine learning techniques to perform information extraction, named entity recognition, and text normalization. As a result, the system is able to correctly interpret the semi-structured speech from a trainee. System robustness accommodates difficulties such as disfluencies and extraneous words that are often present in radio transmissions from trainees. The output of the NLP module is JSON-formatted data that is readily used by other game functions. All speech transcriptions and natural language interpretations are shared on the DIS network, enabling modular development and integration with performance assessment and after-action-review systems.

NVTT-Shadow utilizes a geographically diverse terrain database depicting the fictional countries of Atropia and Ariana. Terrain features include steep mountains, rivers, valleys, and coastal regions. Correlated satellite imagery overlays the wireframe terrain world, such that cultural features (e.g., bridges, roads) are represented. The presence of simulated ground vehicles and dismounted soldiers is defined within OneSAF and subsequently overlayed onto the terrain database. A single terrain database is utilized for all game missions.

NVTT Shadow Gameplay

NVTT-Shadow serves to reinforce classroom instruction with the Soldier's ability to practice tactical communication skills in an engaging and interactive gaming environment. The game is divided into Training missions and Campaign missions. Each mission begins with a briefing that outlines the high level mission objective (e.g., Area Reconnaissance) and critical actions required for success (e.g., "locate hostile threats and conduct target handover to Apache 11 using laser target marker"). The ten Training missions exercise progressively more difficult skills in increasingly complex situations. Training missions serve as a tutorial for understanding command and control of the aircraft and sensor, as well as providing training on all tactical communications executed as part of the Campaign missions. Training missions require 5-15 minutes of gameplay depending on player's skill level. The 14 Campaign missions incorporate multiple skills and follow typical Combat Aviation Brigade operations. Mission success is predicated on accomplishing mission objectives in a timely and accurate manner with the appropriate tactical communications. Campaign missions are approximately 20 minutes in duration depending on player skill level.

As players progress through the Campaign missions, they receive scores and formative performance feedback. The feedback not only provides corrective information but also guides players to remediation resources. Players are measured and scored on the transmission and content of tactical communications across five primary dimensions:

1. **Accuracy**: Did the trainee accurately describe and report the event in the scenario? Trainee utterances must match one of a set of predefined possible lexical formulations for the event. Moreover, specificity counts: for example, "red truck" is always preferred to simply "truck". Distinctions such as these are reflected in the accuracy score.
2. **Completeness**: Did the trainee report all required information for the event? Utterances are parsed into slots of required information with respect to communication type. For example, for a SPOT report, slots include number, description, activity, location, time, and "what I'm doing". Completeness is computed as the percentage of slots filled by the trainee.
3. **Order**: Did the order in which a trainee reported the event match protocol? Most communication types must follow a structured format where the order of slots of information is prescribed. Order is computed as the distance in terms of "edits" (re-arrangement of a pair of slots) from the prescribed order.
4. **Brevity**: Did the trainee report the event as concisely as possible? Brevity can be operationalized in at least three ways in the context of communications measurement for military training. First, brevity can refer to the trainee's use of "brevity codes" at the appropriate times. Second, the speed or rate of transmission of the trainee's communication can be measured (e.g., in terms of time of utterance from start to completion). Third, the "density" of information conveyed could capture an intuitive notion of conciseness. Currently, we simply measure the use of brevity codes and length of transmission.
5. **Timeliness**: Did the trainee report the event in a timely manner according to protocol? Timeliness is defined as the speed that a communication is formulated and transmitted relative to event observation in the scenario.

Measurement of these dimensions takes into account the content and form of trainee utterances, as well as contextual information from the simulation environment. For example, accuracy of a description in a SPOT report is relative to a known entity or event in the scenario captured from the DIS data stream, and timeliness is measured with respect to the event onset in simulation runtime.

Upon completion of a Campaign mission, a game scorecard is displayed containing performance scores on overall mission objectives as well as the critical MUM-T skills (see Figure 1). Mission scores are totaled utilizing a percent goal accomplished from the stated objectives of the mission. The player has transparency on how effectively he engaged the target in relation to the mission objective. Communications scoring consists of how effectively the soldier completed the appropriate communications report for a prescribed scenario event. A sample audio report of desired communications protocol is accessible via icon within the scorecard for immediate remediation training. Soldier trainees may access their play history through a tabbed window. All missions played-to-date are shown with "Date Played" and "Mission Score." Individual advancement through the game is denoted by rank attainment displayed at the top of the scorecard. Successful mission completions are summed toward rank credit and advancement. Planned game features include the instructor's ability to summarize class advancement against critical tasks as well as drill down capability to display individual soldier ranking data. Capabilities for querying specific critical tasks against individual data are in development plans.

Figure 1. Example mission scorecard and formative feedback.

Utilization

Game Assessment Plans

Initial user feedback will be collected from a sample of U.S. Army 15W AIT soldiers and identified UAS SMEs. Feedback will serve to refine NLP data, scoring measures, and scorecard design. Future implementation of aggregated soldier data will be explored with UAS course instructors. Once design changes are implemented, a game assessment is planned to measure training effectiveness and tactical communications skill advancement for 15W soldier trainees. Data collection is scheduled to commence later this year. Results and critical design considerations will be made available to the training research community in later publication.

Acknowledgements

The authors extend their great appreciation to those who helped conduct the research and development described here: Dr. John Stewart (ARI); Mr. Troy Zeidman (Imprimis, Inc.); Mr. Courtney Dean, Dr. Zack Horn, and Dr. Brian Riordan (Aptima, Inc.); Mr. Steve Berglie and Mr. Steve Webster (Kinex, Inc.); and Mr. JW Dirkse and Mr. Andy Gross (Trideum, Inc.). The authors also thank our colleagues at Night Vision and Electronic Sensors Directorate, Modeling and Simulation Division, Ms. Susan Harkrider and Mr. Chris May. Finally, the authors thank all of the U.S. Army aviation subject-matter experts who contributed their time and knowledge to the development of this work.

References

Department of the Army. (2009). Army Unmanned Aircraft System Operations. (FM 3-04.155). Fort Monroe, VA.

Department of the Army. (2014). Combat Aviation Gunnery. (TC 3-04.45). Fort Rucker, AL.

Stewart, J. E., Bink, M. L., Barker, W. C., Tremlett, M. L., & Price, D. (2011). *Training needs for RQ-7B unmanned aircraft system operators in the scout-reconnaissance role* (Research Report 1940). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADB 367652).

Stewart, J. E., Bink, M. L., Dean, C.R., & Zeidman, T. (2015). *Developing Performance Measures for Manned-Unmanned Teaming Skills* (Draft ARI Research Report). Author.

Sticha, P. J., Howse, W. R., Stewart. J. E., Conzelman, C E., & Thibodeaux, C. (2012). *Identifying critical manned-unmanned teaming skills for unmanned aircraft system operators*. (ARI Research Report 1962). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA565510).

United States Army Aviation Center of Excellence (USAACE). (2008) Aviation Restructuring Initiative Combat Aviation Brigade Unmanned Aircraft System Integration Doctrine and Training Supplement. Fort Rucker, AL.

United States Army Aviation Center of Excellence (USAACE). (August 2014). MUM-T Handbook: Leveraging Aviation Manned Unmanned Teaming (Draft). Fort Rucker, AL.

OPTIMIZING PERFORMANCE OF TRAINEES FOR UAS MANPOWER, INTERFACE AND SELECTION (OPTUMIS): A HUMAN SYSTEMS INTEGRATION (HSI) APPROACH

Ms. Jennifer Pagan
Dr. Randy Astwood
CDR Henry Phillips
Naval Air Warfare Center Training Systems Division
Orlando, Florida

Unmanned Aerial System (UAS) operations research by Williams (2004) found that platforms which employ winged aviators (e.g., Predator) have shown higher mishaps than those that select operators that are nonpilots (e.g., Shadow). One explanation may be negative training transfer from manned to unmanned platforms as operators are separated from the aircraft, thus depriving them of a range of sensory cues (McCarley & Wickens, 2007). Another explanation for higher Predator mishaps may be associated with poor Ground Control Station (GCS) design. These varying explanations for differences in mishap rates across platforms indicate the need to address a number of Human System Integration (HSI) issues including manpower/personnel, training, and design issues. Thus, this presentation discusses an effort investigating which UAS Knowledge, Skills, and Abilities, (KSAs) support the identification and training of candidates best suited to operate UASs. In addition, GCS design considerations directly linked to task workload and KSAs are discussed.

Authors' Note. The views expressed herein are those of the authors and do not necessarily reflect the official position of the organizations with which they are affiliated.

Human Systems Integration (HSI) is the multi-disciplinary marriage of systems engineering and behavioral science (Bost & Miller, 2003). HSI seeks to address issues associated with how the human interacts with other system elements (e.g., hardware/software) to ensure effective performance and safety. Within the DoD HSI consists of a number of disciplines including manpower/personnel, training, safety and health, habitability, survivability, and Human Factors (Bost & Miller, 2003). Manpower/Personnel addresses all aspects of defining requirements for personnel including selecting and retaining those individuals. Training seeks to equip personnel with the necessary KSAs for successful mission completion. Safety and health, habitability, and survivability seek to ensure that systems are designed to minimize personnel risk of injury and error, ensure that all aspects of the working spaces are designed with personnel in mind, and provide personnel with all requisite personal protection needed. Lastly, the Human Factors (HF) component of HSI seeks to ensure that all aspects of the system are designed with the full consideration of the inherent capabilities and limitations of personnel.

Research on UAS mishaps has begun to uncover fundamental HSI problems associated with selection, training, and design for UAS operations. Specifically, Williams (2004) found that UAS platforms utilizing winged aviators as operators (i.e., Predator) have significantly more HF related accidents than those operated by enlisted personnel (i.e., Shadow). Investigation of Predator accidents indicates issues concerning instrumentation, sensory feedback systems, and channelized attention. Conversely, Shadow HF accidents were found to be related to procedural guidance and publications, training issues, overconfidence and crew resource management (Thompson, Tvaryanas, & Constable, 2005). Although this comparison is of UASs from different groups, with Predators (Group IV) flying higher and faster than Shadows (Group III), this difference is unlikely to change the underlying HF issues associated with flying beyond visual range. One explanation for these findings may be negative training transfer from manned to unmanned platforms as operators are separated from the aircraft, thus depriving them of a range of sensory cues (McCarley & Wickens, 2007). This separation of aircraft and pilot puts winged aviators in a situation in which they are unable to employ the psychomotor skills that have been trained into automaticity (Grier et al., 2003), suggesting that winged aviators may not necessarily have the right competencies (i.e., KSAs) to operate a UAS. Additionally, Pagan et al. (2014) and Triplett (2008) found that there are core manned aviator KSAs that go unused when operating UASs (e.g., cognitive/spatial, physical/perceptual, and personality based competencies).

While researchers have begun to address *who* should operate UASs (e.g., McKinley, McIntire, & Funke, 2009; Pagan et al., 2014; Triplett, 2008), selection is only one component to addressing UAS mishaps from the HSI perspective. Another critical aspect is *system design* as multiple studies have cited confusion with interface and

automation modes and difficulty with system management as primary causes for Predator HF related mishaps (Nullmeyer, Montijo, Herz, & Leonik, 2007; Thompson et al., 2005; Tvaryanas & Thompson, 2008).

These varying explanations for varying mishap rates across platforms, coupled with the fact that accident rates for Global Hawk, Predator, and Reaper are still three times higher than any other category of aircraft within the U.S. Air Force (Bloomberg, 2012) suggest further research is warranted. Specifically, research is necessary to identify the right individuals with the capabilities to acquire UAS specific skills and ensure they are trained to the appropriate KSAs, as well as to derive GCS design guidance that is optimized in a manner that improves overall safety and performance. The Optimizing Performance of Trainees for UAS Manpower, Interface and Selection (OPTUMIS) effort was developed to address these HSI concerns. OPTUMIS consists of three phases: 1) KSA Comparison (manned vs. unmanned), 2) Air Vehicle Operator (AVO) KSA Classification (select, train, design), and 3) Performance Differences. This paper describes preliminary results from the second phase of this research effort. Specifically, this paper will discusses our attempt to identify those KSAs that should be used for selection and those that should be used for training U.S. Navy UAS AVOs, as well as discusses UAS GCS design considerations that are directly linked to UAS operator task workload.

Method

Measures

Job Task Analyses (JTAs). The Analysis of Cross-Platform Naval Unmanned Aircraft System Task and Competency Requirements (Mangos, Vincenzi, Shrader, Williams, & Arnold, 2012) was used to identify UAS AVOs tasks and requisite KSAs. This JTA focused on all major UAS systems actively used by the U.S. Navy and Marine Corps. This JTA identified 256 general and system-specific operator (i.e., crew member, by position) tasks and 67 requisite KSAs across platforms. The Mangos et al., 2012 JTA also provided task difficulty, importance, and frequency SME ratings as well as KSA SME importance ratings. Additionally, a qualitative analysis of existing UAS JTAs was conducted to ensure a comprehensive list of KSAs was included for further analyses.[1] This analysis identified another 42 requisite KSAs bringing the total to 109 UAS cross-platform relevant competencies.

Existing Measures for Selection, Training, and Design Classification. An analysis of existing methods for providing selection, training, and system design guidance that is linked directly to requisite tasks and KSAOs was conducted. This analysis involved three steps: 1) identifying overlap among existing methods, 2) identifying unique methods, and 3) expanding/developing a model for design guidance. Results from this analysis deemed it necessary to expand the Brannick and Levine (2002) model for training and selection guidance to include design guidance. This updated model was used to develop techniques and collect required KSA and task information (e.g., ranking, categorizing, and elaborating) from UAS SMEs in order to obtain selection, training, and GCS design recommendations.

AVO KSA Classification Survey. The AVO KSA Classification Survey was developed utilizing the Brannick and Levine (2002) model for KSA Selection and Training classification. Each KSA was presented with a definition and SMEs were asked to provide consensus ratings for each KSA on four scales:

- *Necessary*: Is the attribute necessary for newly hired employees to possess upon entering the job? This is a dichotomous, yes/no response.
- *Practical*: Is the attribute practical to expect of incoming employees in the current labor market? Also a dichotomous, yes/no response.
- *Trouble Likely*: To what extent is trouble on the job likely if this attribute is ignored in selection (compared with the other attributes)? This is a 5-point scale ranging from "very little or none" (1) to "an extremely great extent" (5).
- *Superior from Average*. To what extent do different levels of the KSA distinguish the superior from the average operator (compared to the other KSAs)? The Superior from Average scale was rated on the same 5-point scale as the Trouble Likely scale.

[1]Due to space limitations a complete listing of UAS JTAs utilized for this effort was not provided. For a complete listing of utilized UAS JTAs please contact the study authors.

Participants

Seven U.S. Navy UAS operators were asked to provide consensus ratings for the AVO KSA Classification Survey. Operators' backgrounds consisted of AVOs (2), Mission Commanders (3), and a Sensor Operator (1). Their experience ranged from 10 months to 3.5 years in Groups III -V (e.g. Shadow, Fire Scout, Broad Area Maritime Surveillance Demonstrator [BAMS-D]).

Results

A multi-pronged approached was used to classify KSAs into selection, training, or design categories. First, data from the KSA Classification Survey was used to classify selection and training categories.

Selection. KSAs were classified as required for selection based on the Brannick and Levine (2002) criteria: KSAs rated as "Necessary", "Practical", and 1.5 or higher on the "Trouble Likely" scale. Subsequently, selection KSAs were ranked based on a weighted score derived from multiplying scores on the "Trouble Likely" scale by scores on the "Superior from Average" scale (Brannick & Levine, 2002). Next, the "select to" KSAs were cross referenced with the KSA importance ratings from the Mangos et al., 2012 JTA to ensure all selection KSAs were rated as greater than moderately important on the five point importance scale used (i.e., 3.5 or greater). Finally, KSAs that were considered to be minimum qualifications for job performance we removed (e.g., general health, dynamic flexibility). The resulting KSAs are presented in Table 1. The KSAs presented in Table 1 are broken into four tiers. KSAs within a tier are grouped by importance ranking for selection (i.e., Tier 1 KSAs are most valuable for selection of UAS operators and Tier 4 are least valuable).

Table 1.
UAS Operator "Select To" KSAs

Tier 1		Tier 2	Tier 3	Tier 4
Dependable	Auditory Attention /Localization	Rule Abiding	Interpersonal Skills	Oral Expression
Self-Discipline	Finger Dexterity	Learning Ability	Cooperation	Oral Comprehension
Accountability	Wrist-Finger Speed	Numerical Reasoning	Listening Skills	
Mathematical Ability	Multi-limb Coordination	Work Motivation		
Control Precision	Vigilance	Perseverance		
Manual Dexterity	Resilience	Straightforward-ness		
Hand-Eye Coordination	Moral Interest	Cohesiveness		
Reaction Time	Attention to Detail	Extraversion		
Information Management Skills				

Training. Next, KSAs required for training were classified and ranked using the Brannick and Levine, 2002 methodology: KSAs were classified for training if they were rated as *not* "Necessary" and given a greater than 1.5 rating on the "Superior from Average" scale; training KSAs were then ranked based on their "Superior from Average" score. Then these KSAs were cross-referenced with the Mangos et al., 2012 KSA importance ratings. The resulting "train to" KSAs are presented in Table 2.

Table 2.
UAS Operator "Train To" KSAs

Tier 1			Tier 2	Tier 3
Deliberation	Planning	Threat Categories and Indicators	Mechanical Comprehension	Confidence
Adaptability	Safety Consciousness	Reconnaissance Procedures	Perceptual Speed and Accuracy	Long-Term Memory
Stress Tolerance	Systems Comprehension	Engagement Procedures	Response Selection	Depth Perception
Handling Crisis	Technical Troubleshooting	Meteorology	Organization Skills	Stamina
Disengagement	Decision Making	Aeronautical Terminology	Time Management Skills	
Working Memory	Energy	Flight Rules and Regulations	Critical Thinking Skills	
Initiative	Leadership	Information Orderings	Reasoning Skills	
Concentration/ Selection Attention	Assertiveness	Rate Control	Problem Solving Skills	
Attention Allocation	Map Reading	Situational Awareness	Teamwork Skills	
Task Prioritization	Unit/Command Objectives	Originality	Category Flexibility	
Navigation Skills	Aviation Principles	Resolving Conflicts	Helpfulness	
Spatial Orientation	Basic Operation Procedure			
Spatial Visualization	UAS Operations[2]			
Mental Rotation	Arm-handedness			
Communication Procedures				

[2]UAS Operations includes navigation, sensors, and weapons knowledge.

Design Guidance. The Brannick and Levine (2002) method was also used to identify KSAs relevant to performance that SMEs determined should be addressed through system design rather than through training or selection. This list included any KSA that was rated highly on the "Superior from Average" scale (rating of 3 or greater) but not considered "Necessary" or "Practical", and that was rated low on the "Trouble Likely" scale (i.e., not a candidate for training). The only KSAs that met these criteria and were placed in the "design to" category among the 109 KSAs were Flexibility of Closure (i.e., identifying/detecting known patterns [e.g., figure, word, object] that are hidden in other material) and Pattern Recognition (i.e., detecting a known pattern [e.g., a numerical code]; combining and organizing different pieces of information into a meaningful pattern quickly).
The project team is currently adapting the Brannick and Levine (2002) method to include workload ratings tied directly to individual tasks. Workload ratings provided for those UAS Cross Platform JTA tasks during which the AVO has direct interaction with the system (amounting to 188 of the original 256) are currently being evaluated to identify the optimal candidate tasks for incorporation into automation evaluations. These data are currently being collected and will further inform system design guidance.

Implications

Our analysis found a number of general competencies that should be considered when developing selection and training protocols for UAS AVOs in order to avoid costly mishaps. These competencies are ranked by importance to provide cost-benefit guidance to selection and training decision makers. For example, if funding constraints prevent decision makers from implementing a selection test battery that measures all of the KSAs identified in Table 1 then they can at minimum ensure that a sampling of the Tier 1 competencies are utilized. Additionally, our guidance can be used as a gap analysis tool for current UAS selection and training protocols. Decision makers can use this guidance to ensure that their current selection and training protocols include those KSAs identified in our analysis. These protocols can then be updated accordingly depending on the individual programs requirements and funding.

The methodology used for the initial "design to" competency analysis identified two KSAs (i.e., Flexibility of Closure and Pattern Recognition) that SMEs reported cannot reliably be addressed through selection or training. While all other KSAs should be considered during the design process, *Flexibility of Closure* and *Pattern Recognition* should be considered critical from a GCS design perspective as they adhere to the principles of HSI. For example, the literature has shown that systems can be designed to allow operators to more easily recognize patterns to improve the quality of their decision making and performance (Cummings, Bruni, Mercier, & Mitchell, 2007). Additionally, we are in the process of expanding our analysis to the task level to provide a more robust set of design guidelines linked to both tasks and KSAs.

Limitations

As previously mentioned our effort sought to provide platform agnostic guidance. However, one must not blindly follow the guidance provided. Selection, training, and system developers must be sure that when developing these protocols and technologies the individual competencies indeed meet their platform requirements. For example, *Wrist-Finger Speed* and *Arm Handedness* were identified as Tier 1 competencies; however, these may only be relevant to UAS platforms that use joystick interfaces.

Moreover, the sample used to develop this guidance was service specific, consisting of seven Navy operators with Group 3-5 experience and may not be generalizable to other services or smaller platforms. Further, it is important to note that these findings are preliminary, as the small sample size warrants the need for additional data points.

Additionally, these rating Finally, further research is necessary to better understand the empirical implications from this guidance. Empirical investigation will provide insight as to whether selecting, training, and designing to these specific competencies will in fact improve operator performance and in turn reduce UAS mishaps.

References

Bloomberg, 17 June 2012. Drones most accident-prone U.S. Air Force craft. Retrieved from http://www.bloomberg.com/news/2012-06-18/drones-most-accident-prone-u-s-air-force-craft-bgov-barometer.html

Bost, J. R., & Miller, G. E. (2003, October). *Human Systems Integration Overview: Session 1*. Presentation conducted by the Society of Naval Architects & Marine Engineers, Orlando, FL.

Brannick, M. T. & Levine, E. L. (2002). Hybrid methods. *Job analysis: Methods, research, and applications for human resource management in the new millennium* (pp.99-132). Thousand Oaks, CA: Sage Publications, Inc.

Cummings, M. L., Bruni, S., Mercier, S., & Mitchell, P. J. (2007). Automation architecture for single operator, multiple UAV command and control. *The International C2 Journal Special Issue: Decision Support for Network-Centric Command and Control* 1 (2), pp.1-24.

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., et al. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors, 45*, 349-359.

Mangos, P., Vincenzi, D., Shrader, D., Williams, H., & Arnold, R. (2012). *Analysis of cross-platform naval unmanned aircraft system task and competency requirements*. Unpublished technical report, Naval Air Systems Command (NAVAIR), Patuxent River, MD.

McCarley, J. S., & Wickens, C. D. (2007). *Human factors implications of UAVs in the national airspace* (Report No. AHFD-05-05/FAA-05-01).

McKinley, A. R., McIntire, L. K., & Funke, M. A. (2009). *Operator selection for unmanned aerial vehicle operators: A comparison of video game players and manned aircraft pilots* (Report No. AFRL-RH-WP-TR-2010-0057). Wright-Patterson Air Force Base, OH. Air Force Research Laboratory.

Nullmeyer, R. T., Montijo, G. A., Herz, R., & Leonik, R. (2007). Birds of prey: Training solutions to human factors issues. *Proceeding of the 2007 Interservice/Industry Training, Simulation, & Education Conference* [CD-ROM].

Pagan, J., Astwood, R., & Phillips, H. (2014). Operator Qualification Differences between Manned and Unmanned Aerial System (UAS). *Proceeding of the 2014 Interservice/Industry Training, Simulation, & Education Conference* [CD-ROM].

Paullin, C., Ingerick, M., Trippe, M. D., & Wasko, L. (2011). *Identifying best bet entry-level selection measures for US Air Force remotely piloted aircraft (RPA) pilot and sensor operator (SO) occupation* (Report No. AFCAPS-FR-2011-0013). Human Resources Research Organization (HumRRO), Alexandria, VA.

Thompson, W. T., Tvaryanas, A. P., & Constable, S. H. (2005). *U.S. military unmanned aerial vehicle mishaps: Assessment of the role of human factors using human factors analysis and classification system (HFACS)* (Report No. HSW-PE-BR-TR-2005-0001). Brooks City-Base, TX: United States Air Force 311th Human Systems Wing.

Triplett, J. E. (2008). *The effect of commercial video game playing: a comparison of skills and abilities for the Predator UAV* (Report No. AFIT/GIR/ENV/08-M22). Wright Patterson Air Force Base, OH: Air Force Institute of Technology.

Tvaryanas, A. P., & Thompson, W. T. (2008). Recurrent error pathways in HFACS data: Analysis of 95 mishaps with remotely piloted aircraft. *Aviation, Space, and Environmental Medicine, 7* (5), pp. 525-535.

Tvaryanas, A. P., Platte, W., Swigart, C., Colebank, J., & Miller, N. L. (2008). A resurvey of shift work-related

Fatigue in MQ-1 predator unmanned aircraft system crewmembers (Report No. NPS OR-08-001). *Monterrey, CA: Navy Postgraduate School*.

Williams, K.W. (2004). *A summary of unmanned aircraft accident/incident data: Human factors implications* (Report No. DOT/FAA/AM-04/25). Washington, DC: Office of Aerospace Medicine.

ARMY AVIATION MANNED-UNMANNED TEAMING (MUM-T): PAST, PRESENT, AND FUTURE

Grant Taylor, Ph.D.
U.S. Army Aeroflightdynamics Directorate (AFDD)
Moffett Field, CA
Terry Turpin
Turpin Technologies
Foster City, CA

As the use of unmanned aircraft systems (UAS) in military operations has increased, so too have their capabilities. One recently developed capability is the ability to operate in conjunction with traditional manned aircraft through a process called manned-unmanned teaming (MUM-T), allowing manned aviators to benefit from the unique capabilities of UAS. This paper provides an introduction to the concept of MUM-T, describing the early stages of research and development, current MUM-T capabilities in fielded Army systems, and planned future development efforts to continue to advance the capability.

As unmanned aircraft systems (UAS) continue to increase in number and capability, the user community and technology developers have quickly recognized the tremendous combat multiplier they can provide across the full spectrum of armed conflict. While initially used as intelligence, surveillance, and reconnaissance (ISR) gathering assets, UAS now serve a variety of roles to include scout and attack. UAS also no longer operate in isolation, limited to sending information and receiving commands from a traditional ground control station (GCS). Instead, advanced data links now allow UAS to transmit sensor imagery directly to the aviation and ground warfighters who need it most through a process called manned-unmanned teaming (MUM-T). MUM-T is the cooperative employment of unmanned assets with traditional manned platforms, providing the unique capabilities of each system to be leveraged for the same mission. The primary benefit of this employment concept is to transmit live intelligence captured from the unmanned system to the manned asset, providing the manned operator with improved situational awareness without placing them at risk.

Although MUM-T can describe the coordination between any manned platform (land, air, or sea) and any unmanned platform (land, air, or sea), technologies specifically intended for manned and unmanned aviation platforms have received the greatest attention from the development and currently has the most advanced fielded capabilities. Therefore, this paper will focus exclusively on MUM-T research and technologies intended to support aviation assets. Specifically, the authors present a review of the initial series of MUM-T research programs and technology demonstrations, a description of the current state-of-the-art capabilities, and continuing research being conducted by the Army to further advance the concept.

Past: Previous MUM-T Research Programs

MUM I - IV

Preliminary investigations into the MUM-T concept began in 1997 with a series of four Concept Evaluation Programs titled MUM I, II, III, and IV, led by the Army's Air Mobility Battle Lab at Ft. Rucker, AL (Jones, 2001). These studies sought to evaluate the impact of MUM-T on the efficiency, effectiveness, survivability, and timeliness of the air weapons team, specifically while conducting tactical reconnaissance missions. The information collected through this series of studies established the foundation for all future MUM-T research and development.

The objective of these studies was to determine how many UAS could be controlled at once; the workload associated with controlling between one and four UAS at LOI 4 (see Table 1); appropriate tactics, techniques and procedures (TTPs); and the effectiveness of cognitive decisions aiding systems (CDAS) in reducing workload. The studies were conducted using two networked Comanche Portable Cockpits acting as a scout/attack weapons team, and a notional vertical lift UAS with hover and speed parity with Comanche.

The culminating study (MUM IV) showed that the maximum number of UAS that could be controlled while remaining an active shooter was marginally two. Managing three UAS took the manned aircraft out of the fight due to extremely high workload. Many different tactics were attempted including using the UAS as a wingman

that clearly showed MUM-T to be a force multiplier. The CDAS was never fully implemented due to schedule and cost limitations. This resulted in very high workload managing even one UAS. The Comanche cockpit pilot vehicle interface was also not sufficiently optimized to support the necessary MUM-T tasks which again increased workload and negatively impacted crewmember situational awareness. Even with all of these negatives the knowledge gained from these experiments continues to shape the direction of MUM-T R&D efforts to this day.

Table 1.
Defined Levels of Interoperability (LOI).

LOI Level	Definition: Operator in the manned platform has the ability to…
1	Verbally communicate with UAS operator via radio
2	View UAS sensor imagery in real-time
3	Control UAS sensor payload orientation
4	Control UAS aircraft position via waypoint navigation
5	Assume complete control of UAS, including take-off and landing

Note. Higher levels include all lower level capabilities (e.g. LOI 4 provides control of aircraft position and sensor payload orientation, as well as real-time sensor imagery).

Airborne Manned/Unmanned System Technology Demonstration

The first major follow-on to the preliminary MUM studies was the Airborne Manned/Unmanned System Technology Demonstration (AMUST-D) in 2002 (Colucci, 2004). This program sought to develop and demonstrate new technologies built specifically for interoperability with UAS from manned helicopters. The program consisted of two related efforts: the Warfighter's Associate, led by Boeing, which provided control of UAS from the co-pilot gunner (CPG) station of the Apache; and Mobile Commander Associate, led by Lockheed Martin, which provided UAS control from the Army Airborne Command and Control System (A2C2S) in the back of the Blackhawk. Both systems sought to transition CDAS functionality originally developed for the Rotorcraft Pilot's Associate (RPA, Miller & Hannen, 1999), which included advanced autonomous behaviors, data fusion techniques, and intelligent flight routing – all capabilities intended to free up operator cognitive resources, allowing them to focus their limited attention on the battle rather than aircraft management. The AMUST-D program also sought to overcome the interface shortcomings identified by the previous MUM studies. Although the Mobile Commander Associate system was never formally fielded, due to the A2C2S system never being integrated into the Blackhawk fielding plan, the Warfighter's Associate system continued development until eventually being an integral component in the Apache AH-64D Block III upgrade, as well as the AH-64E model.

Hunter Standoff Killer Team

The Warfighter's Associate and Mobile Commander Associate technologies from AMUST-D were further developed and tested through the Hunter Standoff Killer Team (HSKT) program in 2005, led by the Army's Aviation Applied Technology Directorate (AATD, Colucci, 2004). The HSKT program primarily focused on hardware integration (datalink, sensors, etc.) rather than the operator's control station. The improved hardware demonstrated for the first time that MUM-T could be beneficial beyond just tactical reconnaissance, but for weapons engagements as well. An improved sensor payload (including autotracking capabilities and a laser designator) on the Hunter UAS allowed it to designate a target to be engaged by an attack helicopter (cooperative engagement), increasing the standoff distance, and thus safety, of the manned platform.

Manned-Unmanned Systems Integration Capability

The Army's Program Executive Office for Aviation coordinated the Manned-Unmanned Systems Integration Capability (MUSIC) Exercise in 2011. This capstone event was the largest demonstration of MUM-T interoperability ever attempted (Shelton, 2011). It showcased new technologies that demonstrated the capability of providing interoperability between manned and unmanned assets at a higher technology readiness level (TRL) than ever before. These technologies, ranging from small soldier-portable systems such as the One System Remote Video

Terminal (OSRVT) to major upgrades to the Apache and Kiowa Warrior helicopter platforms, were the final proofs of the concept prior to the capabilities being fielded to live aircraft.

Present: Current MUM-T Capabilities

Although MUM-T capabilities currently exist in portable units like the OSRVT, which will soon be onboard Army utility and cargo aircraft, the most advanced MUM-T functionality currently fielded resides in the CPG station of the AH-64E model Apache helicopter. As such, this system will be the focus of discussion here.

Level of Interoperability

The AH-64E is the first fielded aircraft to provide manned platform crew members with LOI 3 and 4 capability, allowing them to not only view live imagery collected from the UAS sensor, but also take direct control of the sensor and even the UAS aircraft itself if desired. This capability greatly enhances the speed of MUM-T operations by avoiding the need for the traditional "talk on" process, wherein the manned aviator must verbally describe the desired target to the UAS operators in the ground control station. This can be a lengthy and complicated process, requiring a high degree of understanding of the local terrain from both parties. With LOI 3, the CPG can instead take control of the sensor himself and quickly orient it exactly where he wants. At LOI 4, this concept is extended to include the ability to control the position of the UAS aircraft itself, which is particularly useful if the CPG requires a view of a target from a specific vantage point, or needs to ensure that the UAS is in a safe position during a weapons engagement.

System Controls and Displays

One of the primary goals of the Warfighter's Associate system, originally developed under the AMUST-D program, was to utilize existing controls and displays already onboard the aircraft for MUM-T operations. As the Warfighter's Associate system gradually evolved into the AH-64D Block III upgrade, and subsequently the AH-64E model, this design philosophy maintained. In fact, the system allows the CPG to control not only his own aircraft's sensor and weapons systems, but also take up to LOI 4 control of a single UAS, with only one additional switch: a mode selector which alternates the function of the existing Target Acquisition and Designation Sights (TADS) Electronic Display and Control (TEDAC) system between ownship equipment and UAS equipment (Figure 1). This interface design not only provides the most efficient use of size, weight, and power limitations (which are very restricted on attack helicopters), but also minimizes the training requirements and workload imposed on the operator through the use of a new interface. The CPG's TEDAC system provide standard controls for UAS teleoperation, such as manipulating the pan/tilt/zoom of the sensor payload, alternating between a variety of UAS sensors, and activating the laser designator. In addition to these standard control methods, the system also provides two unique control methods that are significant workload reducers for MUM-T operations.

Figure 1. TEDAC system from the Apache helicopter, used to control both helicopter and UAS sensors.

The first unique mode is the *sensor guide mode*, which is sometimes informally referred to as LOI 3.5, because it provides the operator with complete control over the sensor (LOI 3) with partial authority over the vehicle's flight path. However, rather than explicitly commanding a specific loiter point or route for the UAS to follow, the aircraft will autonomously generate its own flight path to provide the optimal viewing angle of the ground region currently in view of the sensor (typically a 45° downward angle). This method of control allows the operator to only focus his attention on the task that is important to him, viewing a particular region of the ground, without the additional workload associated with managing the aircraft itself.

Another method used to reduce workload is the *sensor slave* functionality. This function allows the operator to instantly orient the UAS sensor to image the same geographic position on the ground that is currently viewed by the Apache's own sensor (or vice versa, slaving the Apache sensor to the UAS sensor position). This technique simplifies the common task of coordinating target locations between the manned and unmanned systems, allowing all team members to more quickly and accurately establish a common operator picture.

Future: Ongoing MUM-T Research Efforts

Development of Tactics, Techniques, and Procedures (TTPs)

The most pressing current need for MUM-T development is the formalization of doctrine. Despite the capability being fielded for several years, technological development has outpaced the tactical development to the extent that formalized doctrine prescribing proper tactics, techniques, and procedures (TTPs) to be used for MUM-T missions has yet to be established. This is an uncommon circumstance for the Army, as new technological capabilities are typically developed to overcome an established capability gap, allowing for the TTPs associated with the technology to be well understood prior to fielding. MUM-T has evolved in a unique fashion, wherein the capability was recognized to provide a benefit to the warfighter, but wasn't developed as a deliberate solution to a specific problem. As a result, the capability has been fielded without explicit instruction regarding its associated TTPs, leaving the decision of how to tactically implement the capability to the warfighter.

Although this approach to implementation is uncommon, it seems to have yielded positive results. The users, unconstrained by official doctrine, have been free to test the capability across a variety of situations to establish how it can be used most effectively. The results of these fielded trials are being fed up the chain to user representatives at the Training and Doctrine Command (TRADOC) Capability Management (TCM) offices to be incorporated into the formalized MUM-T doctrine currently under development, an effort led by the TCM for Reconnaissance and Attack as well as the TCM for UAS (POC: CPT Tom Kavanaugh, Thomas.P.Kavanaugh2.mil@mail.mil, 334-255-2108).

MUM-T 2030

As users are working to establish the ideal implementation of current MUM-T capabilities, they are also identifying limitations of the current systems that can be overcome through continued technological development. As with the TTP development, these requests for system modifications are also consolidated by the TCM offices. Of course, the users' desired capabilities will always exceed what can feasibly be delivered due to constrained budgets, time, and technological capabilities. Therefore, the challenge lies in the need to identify from the list of user requests those which are anticipated to provide the greatest benefit. Leading this effort is the TRADOC Analysis Center (TRAC), which is currently conducting complex cost-benefit analyses on a wide variety of desirable MUM-T capabilities that could conceivably be fielded by 2030 (POC: Iris Chavez, Iris.L.Chavez2.civ@mail.mil). Insight into the feasibility, impact, and anticipated cost of development is provided by the research and development community (primarily from the Aviation and Missile Research, Development, and Engineering Center, or AMRDEC) as well as the Program Management offices for UAS, Apache, and Sensors-Aerial Intelligence. The final report from this analysis, expected to release in the middle of 2016, will establish the framework for the research, development, and integration of new MUM-T capabilities targeted for 2030.

Supervisory Controller for Optimal Role Allocation for Cueing of Human Operators (SCORCH)

Although the specific technological improvements expected to have the greatest impact have yet to be established by the TRAC MUM-T 2030 study, the science and technology (S&T) community has already initiated

efforts toward developing improved MUM-T capabilities. One such effort currently in progress is the SCORCH program, a collaborative effort between researchers from the AMRDEC Aeroflightdynamics Directorate (AFDD), United Technologies Research Center (UTRC), and the University of California, Santa Barbara (POC: Amit Surana, SuranaA@utrc.utc.com). The SCORCH program will develop and evaluate cognitive decision aiding tools and sensor tasking automation that will enable aviators to effectively command teams of up to three advanced UAS simultaneously up to LOI 4 in support of a variety of missions and roles. The research is focused on three areas identified through prior research efforts as critical for a single operator to manage multiple UAS in support of a common mission: the *pilot-vehicle interface*, a *sensor management aide,* and *attention allocation aide*.

The pilot-vehicle interface developed for the SCORCH program is representative of cockpit designs expected to be fielded in near-future Army helicopters. The interface follows guidance set forth in Army UAS Roadmap documents (Department of Defense, 2013) and also takes inspiration from modern commercial aviation cockpit design as well as previous experimental interfaces. Significant departures from current cockpit design include the use of multiple large (15”diagonal) high resolution full color displays with touchscreen capability, and a variation on the current Apache TEDAC hand controller which features its own full color touchscreen display and modified button configuration similar to that found on modern video game controllers (Figure 2). The use of touchscreens throughout is expected to reduce workload associated with initiating system functions, which, through traditional cockpit design, can require the operator to navigate through multiple levels of bezel button pages. Using touchscreens to initiate system functions, the interface can be designed to dynamically adapt to the current mission phase and provide the operator with convenient access to the functions most relevant to their current goals (Sarter, 2007). Of course, touchscreens interfaces have limitations as well, most notably a lack of tactile feedback which requires the operator to focus their visual attention to the interface when executing a function. For this reason it is important that the most frequently used functions continue to utilize traditional physical buttons and switches.

Figure 2. Pilot-vehicle interface developed for the SCORCH program.

The sensor management aide is the first of two independent autonomous support systems developed for evaluation in the SCORCH program. The sensor management aide aims to offload operator workload for lower level sensor control tasks, freeing mental resources to focus on higher level information processing and decision making. The system will consist of various intelligent search algorithms that can manage multiple UAS sensors to collectively search ground regions with optimal efficiency. These autonomous behaviors will free the operator from traditional sensor operations, allowing them to focus instead on processing the imagery collected by the sensors. Further, a robust automatic target recognition system allows the aide to further off-load the operator through assistance with the visual search task, leaving the operator free to focus on top-level mission management and decision making tasks.

The final SCORCH system component is the attention allocation aide, an adaptive CDAS with the goal of improving the operator’s visual search behavior. Development of this system will begin by establishing an

algorithmic model of optimal human effectiveness when conducting a visual search with UAS sensors. This model will run in the background as operators conduct their MUM-T missions. Meanwhile, their visual attention will be monitored in real-time through an eyetracker system (a series of cameras that provide continuous measurement of the location of the user's visual focus). Continual comparison of the user's visual search behavior to the known optimal model will allow the system to make real-time recommendations to improve the efficiency of the operator's visual search. Through this method, the human operator and autonomous system can collaboratively conduct the visual search, with both human and system performing functions for which they are respectively best suited.

Synergistic Unmanned-Manned Intelligent Teaming (SUMIT)

The lessons learned from the SCORCH program will feed into the similar, but larger scale, SUMIT program (POC: Ray Higgins, Raymond.T.Higgins.civ@mail.mil). This effort, led by AATD in collaboration with AFDD, NASA Langley, and various industry and academic partners, will investigate a wide variety of pilot-vehicle interface and CDAS concepts to determine the systems best suited to support MUM-T operations on the Army's next-generation helicopters (Future Vertical Lift, or FVL). In addition to the best-performing system components from the SCORCH program, other industry- and government-developed technologies will be included in the evaluations. These technologies are expected to include voice-control systems, head/eye-tracking, head-up and head-mounted displays, touchscreen displays, and a suite of advanced autonomous support behaviors and CDAS capabilities. These technologies will be systematically evaluated to identify the most beneficial systems, which will be demonstrated in a planned live flight test at the conclusion of the program (expected roughly 2020).

Conclusion

As a result of the efforts of the research and development community, current MUM-T capabilities are already providing manned aviators with a diverse benefits, leading to improved situational awareness, survivability, and lethality. The lessons learned from the current fielded systems, as well as continued research and development, will provide future warfighters with even greater capabilities to better confront future threats. However, current systems have reached a level of complexity and sophistication such that continued advancement is only possible through the coordinated efforts of a diverse collection of research and development professionals working together toward a unified goal.

Acknowledgement

This project is sponsored by the U.S. Army Research Office and the Regents of the University of California, through Contract Number W911NF-09-D-0001 for the Institute for Collaborative BioTechnologies. The information presented does not necessarily reflect the position or the policy of the Government or the Regents of the University of California, and no official endorsement should be inferred.

References

Colucci, F. (2004, November). MUM's the Word. *Rotor & Wing Magazine*. Retrieved from http://www.aviationtoday.com/rw/military/attack/1817.html

Department of Defense (2013). Unmanned Systems Integrated Roadmap: FY2013-2038. Reference 14-S-0553. Retrieved from http://www.defense.gov/pubs/DOD-USRM-2013.pdf

Jones, A. (2001, February). UAV-Air Maneuver Integrated Operations. *Army Aviation*, *50*(2). Retrieved from http://www.quad-a.org/images/magazines/MG_Jones/jones_feb01.pdf

Miller, C. A., & Hannen, M. D. (1999). The Rotorcraft Pilot's Associate: Design and evaluation of an intelligent user interface for cockpit information management. *Knowledge-Based Systems*, *12*(8), 443-456.

Sarter, N. (2007). Coping with complexity through adaptive interface design. In J. A. Jacko (Ed.), *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments* (pp. 493-498). Heidelberg, Germany: Springer. doi:10.1007/978-3-540-73110-8_53

Shelton, M. (2011, October). Manned Unmanned Systems Integration: Mission accomplished. *Army.mil News*. Retrieved from http://www.army.mil/article/67838/

REMOTE-SPLIT OPERATIONS AND VIRTUAL PRESENCE: WHY THE AIR FORCE USES OFFICER PILOTS TO FLY RPAs

Lt Col Matt Martin
United States Air Force
Rapid City, SD

Since the advent of Remote-Split Operations (RSO) for MQ-1/9 remotely-pilot aircraft (RPA), where pilots fly aircraft that are thousands of miles away, a popular view is that this distance instills a psychological gap, making it easy to carry out lethal actions. A common further assumption is that RPAs are automated and don't require traditional aviation or leadership skills to operate. But 20 years of combat RPA experience has led practitioners to a different view—that the effective employment of RPAs has been improved by using pilots with previous experience in manned aircraft and undergraduate training where pilot candidates received a foundation of manned flying skills. Furthermore the USAF experience has been that leadership, character, and decision-making qualities needed for effective RPA employment are nearly identical to those required for the effective and efficient employment of manned weapons system.

10,000 feet above Sadr City, an Air Force pilot maneuvers her bomb- and sensor-laden aircraft into position. She's just spent an hour scanning up and down a route as a US Army construction crew erects a wall on the west side of a main thoroughfare. The aim is to seal off the city in order to isolate insurgents as part of an on-going effort to pacify Baghdad.

From the beginning of this effort, enemy snipers had set up positions on the roofs of houses in order to shoot at the construction teams. It was up to the Air Force to support these teams by using armed ISR aircraft to find and engage the snipers.

After an hour of scanning the main route, the pilot and her crew identify a sniper's nest. Thanks to a video downlink, the Joint Terminal Attack Controller (JTAC) in the battalion operations center views the target in real time. The JTAC confers with the Army Battle Captain, does a quick assessment of possible collateral damage, and accounts for the locations of nearby friendly troops. The Battle Captain decides to take the sniper out.

The JTAC coordinates the strike with the pilot via secure radio. He passes target information and instructions, directs the pilot to de-conflict with other aircraft in the area, and tells her to start her attack run.

The pilot coordinates with the airspace controller, notices an aircraft between her and the target, and determines that she has to shoot from a lower altitude. After receiving clearance, she pickles off the auto-pilot to quickly turn the aircraft to the north and heads for a block of airspace where she can safely descend. At 8,000 feet, she turns her aircraft back to the south, and sets herself up for a west-to-east target run. She briefs her crew and arms her missiles.

Once set up, the pilot hand-flying the aircraft turns inbound and announces "in from the west" over the radio. The JTAC responds with: "Type II control, cleared hot." When the pilot reaches the optimal slant range in the heart of the engagement zone, she rifles the laser-guided missile, which strikes the sniper's nest 19 seconds later. There's an incredible explosion, and the sniper's nest (with the sniper inside) is destroyed. The house next door is untouched.

The sniper fire halts. The construction crews are able to continue their work.

This combat action, and many others like it, took place in Iraq in 2007. Was this an application of airpower that Billy Mitchell or Guillot Douhet would recognize? Was the pilot on board the aircraft or in a ground control station 7,000 miles away? And if the pilot *wasn't* on board, is she even a pilot? And was she psychologically connected to the operation in the same manner as pilots of manned aircraft?

The Air Force is changing. The above scenario will happen many more times in future conflicts. RPAs are growing in number, complexity, capability, and prominence as an unparalleled example of the kind of world-class airpower that is the pride of the USAF. And while the rest of the world struggles to understand the role and psychology of remote piloting, the Air Force is drawing on 100 years of airpower lessons to build the RPA pilot of the future—and she looks a lot like the manned aircraft pilot of the past.

Previous Experience and Rapid Growth

In 2003 when the Air Force first employed the RSO model of operations in Iraq and Afghanistan, they had to the ability to fly three aircraft 24/7. This amounted to about 60 hours of flight time per day. Today the Air Force can fly 65 of these Combat Air Patrols (CAPs) at a time, for a total of about 1,300 flight hours *per day.* And while in 2004 the Air Force required an RPA pilot force of only 50 or so, at the end of 2013 the Air Force had 1,366 MQ-1/9 pilots (GAO, 2014), and is now training over 400 more per year. (Drew, 2014). This rate of growth was possible only because until 2010, the Air Force used only previously experienced manned aircraft pilots to fly the MQ-1 and MQ-9. This allowed them to minimize the training needed to produce new crews and focus only on the transition from manned flying to RPAs.

Prior to this it was clear that aviation skills were directly transferable to the employment of the MQ-1 as a weapons system. Several Air Force Research Lab studies (Hall, 1998; Schreiber, 2002; and Chappelle et. Al 2011) concluded that not only did previous and recent experience in aircraft similar to the MQ-1 improve the performance of test subjects in both basic maneuvering and mission tasks, but that pilots with MQ-1 experience overwhelmingly agreed that manned flying experience was necessary for success in flying the MQ-1.

Based on this conclusion, and under pressure to increase both the number of CAPs and the range of MQ-1 support (reconnaissance, special operations, close air support, and air interdiction) to ongoing operations, the Air Force tailored the MQ-1 initial qualification course to provide a bare minimum of training with the expectation that previous manned operational experience will carry the day. And it did. Between 2003 and the US withdrawal from Iraq in 2011

not only did USAF MQ-1/9 CAPs grow by a whopping 1,403%, but they became the critical enabling capability for US counter-insurgency operations (McCaffrey, 2007).

This approach is evident in the Initial Qualification Training (IQT) syllabus for MQ-1 at the height of the surge. In 2007, the syllabus contained 101.5 hours of academics, 39.5 hours of part-task trainer time, and 31 hours of flying (USAF, 2002). In 2008, the flying hours were *reduced* in a new syllabus of 84 academic, 40.5 simulator, and 20 flying hours. To do this the Air Interdiction phase and the Combat Search and rescue phase were eliminated (USAF, 2008). Both of these syllabi specified that previous operational experience in manned aircraft or graduation from USAF Undergraduate Pilot Training were required to enter the course. It was not until the creation of a true Basic Course for MQ-9 pilot candidates with no previous manned operational experience that the training was increased to 105.5 academic, 61 simulator, and 40 flying hours. (USAF, 2010). This period also saw the introduction of the first hi-fidelity simulator for IQT.

At the same time, the Air Force created the RPA Pilot Training course to prepare officers with no previous flying experience to enter MQ-1/9 IQT. After some testing and adjustment, this undergraduate pipeline now includes 35 hours of manned light aircraft flying (the equivalent of a FAA Private Pilot License), 40 hours of instrument time in a T-6 simulator (for a basic level of instrument flying skill), and 135 hours of academics (Jean, 2010). This move established the model of transference of manned flying skills as the foundation for USAF RPA training.

Why Officer Pilots?

There is a popular notion that flying RPAs is fundamentally different than flying manned aircraft. To quote a prominent researcher: "The Shadow (and Hunter) can effectively do the same mission as the Predator [but using enlisted operators] because Army operators leverage higher levels of autonomy onboard the Shadow than do their Air Force counterparts." (Cummings). But this doesn't capture the difference between the control aircraft and the application of airpower. The ease or difficulty with which an operator can control an aircraft is not the issue.

Obviously technology has made it possible to automate almost every task needed to control an RPA. Tasks can be automated. Judgment cannot. It is the fact of the complexity of the aircraft, the airspace, the mission, and the desired effects that demand the judgment of a trained and mature aviator to employ these aircraft as weapons systems. In exactly the same manner as manned fighter and bomber pilots, MQ-1 and MQ-9 pilots must be prepared to employ weapons and provide target guidance across the entire spectrum of conflict—from major combat ops to counterinsurgency—in every possible type of terrain from open, rural areas, to dense, urban environments—both in the vicinity of and independent of ground forces. In Libya during Operation Unified Protector, MQ-1 conducted suppression of air defenses as well as air interdiction and strike coordination (Etchells, 2011). Right now in Iraq and Syria MQ-1s and MQ-9s are identifying ISIS elements and engaging them both independently and in support of Kurdish forces on the ground (Cole, 2014). It's in those types of scenarios—strategic scenarios in phases 0 through 4 of a conflict—where a mistake can have strategic consequences. In a heavy air-only battle, or when employing weapons within close proximity to friendly forces and non-combatants, making the proper split-second decisions can be a matter of life or death.

The reason the Army is able to use enlisted troops with much less training in a weapons employment role is the fact that they shift the decision-making responsibility away from the crew. A senior NCO or Warrant Officer inside the Tactical Operations Center (TOC) always supervises Army crews. The decision to employ ordinance always rests with the Battle Captain—a field-grade officer who's in charge of the entire operation (Martin, 2014). So while the Army may save on manpower expenses for its aircrew, it still has to pay a bill in the form of supervisory personnel located elsewhere.

This concept highlights the key difference between Army and Air Force employment of similar capabilities. The Army might put a junior NCO in charge of a 60-ton tank, but it wouldn't then send that NCO 1,000 miles behind enemy alone lines to employ it. The Army does everything big, organic, and with lots of supervision. The Air Force on the other hand must retain the flexibility to conduct global interdiction as well as close air support.

And the evidence can be found in the rate of employment. The Army has had the ability to launch the Viper Strike missile off of the Hunter for over several years. But so far there are only been two engagements—both in September '07. That's two engagements in over 200,000 hours of combat time (Harper, 2007). Since the Army always teams unmanned aircraft with manned assets, they typically bring in dedicated shooters to finish engage targets.

By contrast, the MQ-1 and MQ-9 have flown over two million combat hours and conducted thousands of successful weapons engagements since 2003 as independent assets. For instance, during the Surge in Iraq, the MQ-1 fired 112 Hellfire missiles averaging 18 per month **(**Johnson, 2013**).** To date, there has only been a single documented instance of friendly fire incident against US troops from an Air Force MQ-1 or MQ-9 **(**Zucchino and Cloud, 2011**).**

The success of the MQ-1 and MQ-9 across the spectrum of combat continues to drive demand for this capability, and the Air Force now plans to maintain the current 65 CAPs and transition to an all-MQ-9 fleet (Schogol, 2015). The sheer size of this enterprise means that the majority of hardware and manpower resources must be devoted to operations, leaving minimum residual forces for training, supervision, and management. Of the 1,366 pilots on hand in 2013, only 179 were assigned to training duties. 111 more serve in leadership and staff positions. While a small number of pilots within a flying squadron are available for in-unit instruction and supervision, the vast major of squadron pilots fly combat missions as their primary duty. This means that they must operate with minimal direct supervision and must be able to exercise a wide latitude of responsibility and judgment—the very definition of officership.

The Psychology of Remote Combat

Another popular notion of remote combat is that the physical distance between a target and the pilot of an RPA means that the pilot has no emotional or psychological connection with the target. Journalist Mark Bowden wrote that RPA piloting is "like a video game; it's like Call of Duty." (Bowden, 2012). Professor Brennan-Marquez asserts that the "numbness that results from using machines rather than soldiers to carry out our dirty work" produces "the nightmarish image of an 18-year-old drone operator basically playing video games from the detached safety of a Nevada bunker." These attitudes are evidence of a widespread suspicion that RPA pilots

might casually cause collateral damage or otherwise employ these aircraft in a reckless manner. But these attitudes are wrong. And the evidence at hand along with the and experience of RPA pilots leads us to the opposite conclusion.

In fact the psychological connection between the pilot and her target is not a function of distance, it's a function of cognition. Anyone who has ever felt empathy for a fictional character in whatever form (literature, film, etc.) understands that it's very easy for people develop emotional bonds with those they observe—even if the subject of observation doesn't actually exist. As one researcher put it, "The experiences with fictional characters resonate with us because of the fact that we've had deep experiences with people throughout our lives." (Nuwer, 2013).

This phenomenon emerging in long-distance operations is further evidenced by the chronic stress suffered by air traffic controllers (Martindale, 1977), the guilt that some B-17 bombardiers felt during World War II (AP, 1987), and the fact that RPA pilots are just as prone to stress disorders as their manned fighter and bomber counterparts (Dao, 2013). It's clear that no amount of electronic removal or distance between RPA pilots and the targets of their efforts is enough to overcome a lifetime of human empathy and emotional experience.

The strongest evidence of all is the experience of the MQ-1 and MQ-9 pilots themselves. In a multitude of studies and interviews, RPA pilots again and again stress the psychological connection and urgency they experience during operational flying. As one study found: "SMEs also noted the ability to control emotions during urgent situations (e.g., aerial strikes or reconnaissance of enemy combatants, interaction with ground forces, targeting of high-value assets) as especially critical. The attribute of emotional composure is also considered critical to the selection of successful military pilots and high-demand, high-operational military personnel." (Chappelle, 2011). Emotional control and the ability to stay focused on the task at hand in the face of emotional distress simply wouldn't be important if remote piloting by its nature removed the RPA pilot for the human realities of combat.

Conclusion

Since the early days of the RPA experiment, the United States Air Force has made the conscience decision to leverage a century of experience employing airpower to guide the organization, training, equipment, and employment of RPA forces. Indeed, the Air Force approach has been to treat MQ-1s and MQ-9 as much like manned multi-role combat aircraft as possible. By leveraging the manned flying experience of seasoned aviators, the Air Force has been able to expand combat capability at a rate that simply wouldn't have been possible if they had taken a blank-sheet approach. And once the manned flying model was established, it followed as a matter of course to use manned flying as the foundation for the creation of a dedicated RPA career field that trains officers first to fly manned aircraft, and then RPAs, before setting them on a dedicated RPA career track. And despite popular notions that RPAs don't need pilots or that they place a psychological distance between the operator and the target, the Air Force and its RPA practitioners have found that these assumptions don't hold up in combat. And that the use of officers trained as manned pilots is the best approach to building and sustaining a massive MQ-1/9 enterprise in an efficient, effective, and ultimately humane manner.

References

Associated Press (20 Jul 1987). World War II Bombardier Apologizes to Germans. http://www.apnewsarchive.com/1987/World-War-II-Bombardier-Apologizes-to-Germans/id-fa95e130aa23f26514a8c95251ae919a Accessed on 04 Mar 15.

Chappelle et. al. (May 2011). Important and Critical Psychological Attributes of USAF MQ-1 Predator and MQ-9 Reaper Pilots According to SMEs. Air Force Research Lab

Cole, C. (Jul 2014). Drone in Iraq and Syria: What We Know and What We Don't. In *Drone-wars.net.* http://dronewars.net/2014/11/07/drones-in-iraq-and-syria-what-we-know-and-what-we-dont/ Accessed 04 Mar 15.

Cummings, Missy (August, 2008). Of Shadows and White Scarves. *C4ISR Journal*.

Dao, J. (22 Feb 2013). Drone Pilots are Found to Get Stress Disorders Much as Those in Combat Do. In *The New York Times.*

Deptula, D. (2009). *Air Force Unmanned Aerial System (UAS) Flight Plan 2009-2047*. www.af.mil/shared/media/document/AFD-090723-034.pdf. Accessed 6 Jan 2010.

Drew, James (Oct, 2014). MQ-1, MQ-9 Formal Training Units To Produce Fewer Aircrews In FY-15. *InsideDefense.com.* Retrieved from http://insidedefense.com/inside-air-force/mq-1-mq-9-formal-training-units-produce-fewer-aircrews-fy-15

Etchells, A. (Dec 2011). NATO Airpower: Lessons Learned from Libya. In *Defense Viewpoints. http://www.defenceviewpoints.co.uk/military-operations/reflections-on-op-unified-protector* Accessed on 04 Mar 15.

Harper, Douglas (November 12th, 2007). Hunter Delivers a Decade After Cancellation. *Defense Systems Daily*.

Jean G. V. (Nov 2010). Teaching Non-Pilots to Fly Predators Requires More Cockpit Hours in Manned Aircraft in *National Defense Magazine.* http://www.nationaldefensemagazine.org/archive/2010/February/Pages/TeachingNon-PilotstoFlyPredatorsRequiresMoreCockpitHoursinMannedAircraft.aspx Acc. 03 Mar 15

Johnson et. al. (2013). *The 2008 Battle of Sadr City: Reimagining Urban Combat.* Santa Monica: RAND Corporation.

Kiel Brennan-Marquez, "A Progressive Defense of Drones," Salon, 24 May 2013

Mark Bowden, "The Killing Machines: How to Think about Drones," Atlantic, 14 August 2013

Martin et. al. (2014). "Finnishing" the Force: Achieving True Flexibility for the Joint Force Commander. In *Air & Space Power Journal*. May-June 2014.

Martindale, D (1977). Sweaty Palms in the Control Tower. In *Psychology Today,* 10: 71-75.

McCaffrey, B. R. (15 Oct 2007). Memorandum for Colonel Mike Meese, Subject After Action Report. *United States Military Academy.* 5.

Nuwer, R. (Jul 2013). The Psychology of Character Bonding: Why We Feel a Real Connection to Actors. In *The Credits.*

Schogol, J. (06 Feb 2015). Air Force Increases Combat Air Patrols for Reaper Pilots. In *Air Force Times.*

United States Government Accountability Office (Apr 2014). *Actions Needed to Strengthen Management of Unmanned Aerial System Pilots.*

USAF (Dec 2008). *MQ-1B Initial Qualification Training Syllabus.* Air Combat Command.

USAF (Jan 2012). *MQ-9A Initial Qualification Training Syllabus.* Air Combat Command.

USAF (Nov 2007). *MQ-1B Initial Qualification Training Syllabus.* Air Combat Command.

Zuccino and Cloud. (14 Oct 2011). US Deaths in Drone Strike Due to Miscommunication, Report Says. In *The Los Angeles Time.*

COMPARED EVALUATION OF B-ALERT'S ENCEPHALOGRAPHIC WORKLOAD METRICS USING AN OPERATIONAL VIDEO GAME SETUP

Sami LINI[1]
Christophe BEY[2]
Lucille LECOUTRE[1]
Quentin LEBOUR[1]
Pierre-Alexandre FAVIER[2]
[1] Akiani, 109 avenue Roul, 33400 Talence, France {first.last@akiani.fr}
[2] ENSC, 109 avenue Roul, 33400 Talence, France {first.last@ensc.fr}

When it comes to operational human factors studies, the use of a number of different means (psychophysiological, questionnaires, performance indexes) to complete expert behavioral observations allows specialists to issue practical recommendations despite of the variability of the few operators involved. When it comes to mental workload, literature has identified several different physiological ways to assess it. We used Heart Rate Variability (HRV) and pupillometry for previous works (ISAP'11, '13) and both have strong limitations: HRV can only be analyzed over 5-minutes time periods and pupil dilation is subject to light variability.
During this study, we tested the electroencephalography B-Alert X10 system (Advance Brain Monitoring, Inc.) mental workload metrics. We set up an experiment on a video game in real life conditions in order to evaluate the reliability of this index. Participants were asked to play a video game with different levels of goal (easy vs. hard) as we measured subjective, behavioral and physiological indexes (B-Alert mental workload index, pupillometry) of mental workload. Our results indicate that, although most of the measure point toward the same direction, the B-Alert metrics fails to give a clear indication of the mental workload state of the participants. The use of the B-Alert workload index alone is not accurate enough to assess an operator mental workload condition with certainty. Further evaluations of this measure need to be done. As we observed in a previous study, pupil dilation is a reliable index of mental workload as it correlates significantly with most measures.

Introduction

The integration of Humans in the design and evaluation of complex systems is an approach that is becoming increasingly important. There now is a real interest in assessing the impact of such systems on operators who need to handle them. Humans have features of their own, with their constraints and limitations that it is necessary to identify in order to correctly adapt the systems.

When evaluating a system, the concept of mental workload is of particular interest to qualify the operator state. An overload situation can have tragic consequences onto the performances of an operator. Disposing of appropriate tools to assess an operator mental state becomes crucial when evaluating a system and the reliability of such tools is an important issue.

Electroencephalography (EEG) is a good candidate for measuring and monitoring mental workload (Antonenko et al. 2010; Tsang & Vidulich 2006). EEG has some advantages for use in operational environment. In particular, wireless solutions like the B-Alert system (Advance Brain Monitoring, Inc.) (Berka et al., 2005, 2007; Johnson et al., 2011) seem very promising, as they allow more ecological experimental situations. The implemented classification algorithm allows one to use it without requiring extensive medical expertise. We were particularly interested in the mental workload gauge and decided to evaluate this tool following a "black-box" approach.

We set up an experiment on a video game. We chose this particular set-up because it allowed us to put the participants in operational conditions (they sat on an office chair, in front of TV flat screen, with PlayStation controllers in their hands and were able to move freely) making it close to an ecological situation.

To get closer from operational constraints, we made the choice of constraining our recruitment: very few subjects, recruited on the basis of their availability.

We used Rayman Origins, developed for PlayStation 3 for it is a 2D platform game. This allows us to reduce the degrees of freedom (compared to a 3D game) and ensure us the scenarios reproducibility despite the ecological environment. Like for most platform games, the player has to collect items along the level, which defines several performance steps as a function of the number of collected items. Moreover, some scrolling levels were particularly suited to our needs.

(Cegarra & Chevalier, 2008) advise to cross several measures when estimating mental workload. Following this rational, we chose a set of subjective, behavioral and physiological measures to address the issue of the B-Alert workload index reliability.

We had the participants take the Nasa-TLX questionnaire (National Aeronautics and Space Administration Task Load indeX, Hart & Staveland 1988). Physiological measures are also used

Based on the hypothesis that an overload has a negative effect on performances (Wickens, 1992), we also collected two performance indicators: the number of collected items and the number of times the participant dies within a level. A higher number of collected items and lower number of times a participant dies mean better performances.

We had two tests for the EEG workload index: its sensitivity to our experimental manipulation of the mental workload (easy *vs* hard condition), and its confrontation to our control measure.

We hypothesized that:

(H1) Our control measures were sensitive to our task manipulations of the mental workload.
(H2) They were correlated with each other.
(H3) The EEG index of mental workload was sensitive to our task manipulation of the mental workload.
(H4) It was correlated with our control measure.

Methods

Participants

Eight healthy participants (mean age: 22.1 ± 2 years old, 7 males) took part in the study after signing a consent form. They were informed of the purpose of the study.

Measures

Subjective measure

Subjects were asked to evaluate their mental workload after each of the four runs of the experiment with the Nasa-TLX questionnaire. We used an approved French version of this questionnaire (Cegarra & Morgado, 2009). In order to avoid any ambiguity, the participants took a first dry questionnaire during the setup. They then took the questionnaire after each run.

Behavioral measures

Each participant was asked to fill a short questionnaire about information such as his nicotine consumption, sleep deprivation, video games familiarity. We also measured during each run the number of collected items and the number of time the participant died as a measure of their performance.

Physiological measures

- EEG: a portable sensor headset, the B-Alert X10 System was used to record electrophysiological data from 9 electrodes sites (Fz, F3, F4, Cz, C3, C4, POZ, P3, P4), with left and right mastoid as reference. Each EEG channel was sampled at 256 samples per seconds, with a 50-Hz notch filter applied to remove environmental artifact. We also recorded electrocardiogram (ECG) from two electrodes linked to the B-Alert system. The signal decontamination procedure and the classification algorithms are already implemented in the B-Alert system. We were interested into two specific metrics: the probability of being in a high mental workload state, and the probability of being in a high engagement state. The mental workload model and algorithm is described in an article from Berka and colleagues (Berka et al., 2007). For our analysis, we averaged the index for each run of the experiment to derive a mean workload index. 1-second samples with error values were taken out of the analysis as advised by the constructor prior to averaging.
- HRV : we measured Heart Rate Variability (HRV) as another physiological indirect indicator of cognitive load (Backs, Lenneman, & Sicard, 1999; Wilson, 2002). We used the ECG channels of the B-Alert system to derive another measure of the Heart rate variability (HRV). Calculations were performed in both the 0.05-0.15 Hz band (low frequency, LF), and 0.15-0.3 Hz band (high frequency, HF). Total HRV retained for analysis was the ratio of these two values (LF/HF).
- Eye Tracking: subjects were fitted with a mobile eye tracker, Tobii's glasses, a head-mounted eye tracking system resembling a pair of glasses. Tobii studio, the data processing software, allows dealing with mean pupil dilation, i.e. the percentage of dilation compared to the mean dilation measured during the calibration phase.

Procedure

We had the participants take the B-Alert baseline tasks after the EEG was set-up. The choice vigilance task, and standard eyes open and eyes closed vigilance tasks each lasted for 5 minutes. The eye tracker was calibrated after the baseline acquisition. The experimental room brightness was maintained constant, and the participant was in the room long enough before the beginning of the experiment so that we can assume he was accommodated.

Participants were asked to play on two particular levels of the game, chosen for their duration and similarity in difficulty. The display of the levels moves on automatically, forcing the player to move forward. These levels are scripted, events always occur at the same times independently of the actions of the player. This ensured reproducible scenarios from one condition to another and from one participant to another. The participants could train on some other levels with the same game play while they were being set-up.

Each of the two levels was played twice, either with an easy-to-achieve goal, or a hard-to-achieve goal. The easy-to-achieve goal was to collect at least 150 items. The hard-to-achieve goal was to collect at least 300 items. The order of presentation of the levels and goals were assigned for each participant following a latin square procedure to avoid learning effects on the group level analysis. We kept a record of the number of time the participant died as a measure of performance.

Results

Sensitivity of the measures

We first tested the sensitivity of each measure regarding the change in mental workload. We separated the data into two groups: easy goal and hard goal, regardless of the level of the game. One-tailed Wilcoxon's tests are used to analyse inter-individual variability between both conditions (paired samples). The side of the test was given by hypothesis.

As displayed in Figure 1, the Nasa-TLX score (**p=.025**) and the pupil dilation ratio (**p=.022**) were sensitive to our task manipulation of the mental workload, confirming the validity as control measures. The HRV measure (**p=.066**) and the collected item (**p=.080**) measure of performance showed a trend in the direction of our hypothesis. The other measures did not show a significant result.

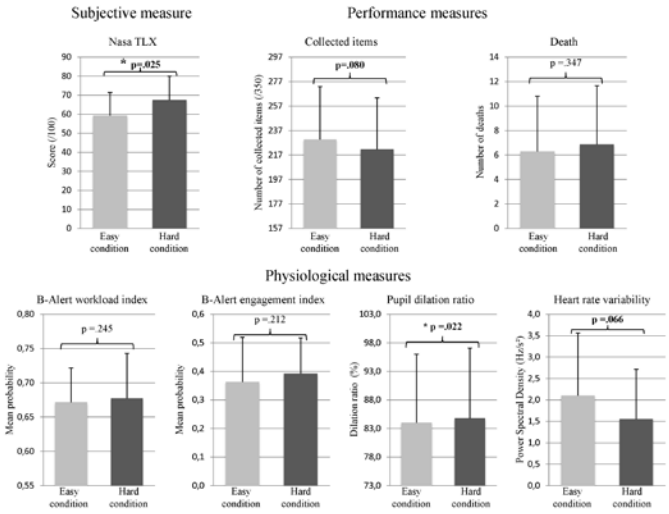


Figure 1. Across subject mean value for each measurement. The easy and hard conditions are compared. Error bars represent one standard deviation from the mean. P-value results from one-tailed Wilcoxon's tests are displayed.

Consistency of the measures

We then analyzed the relations between the measures by computing Spearman correlations (non parametric) between the measures.

The NasaTLX score is correlated with both performance measures. As hypothesized, the higher the NasaTLX score, the lower the number of collected items (R=-0,432, p=0,025) and the higher the number of death (R=0,461, p=0,016).

The pupil dilation measure is also correlated with the NasaTLX score (R=0,486, p=0,042) and both performance measure (items: R=-0,663, p=0,004 ; deaths: R=0,609, p=0,007), confirming its validity as a control measure. It is also strongly correlated with the EEG workload index (R=0,623, p=0,005), but not with the EEG engagement index.

The EEG workload index is also correlated with the NasaTLX score (R=0,430, p=0,026) and the EEG engagement index (R=0,442, p=0,019), although the EEG engagement index is correlated neither with the NasaTLX score nor with the pupil dilation measure. The HRV measure is not correlated with any of the other measures.

Discussion

Validation of the experimental design

The NasaTLX subjective index and the pupil dilation measure were both significantly affected by our task manipulation of the mental workload, whereas the HRV did not. Hence, we can confirm that the experimental manipulation we used was successful at eliciting differential mental workload conditions (**H1**).

The performance measures showed a trend for the number of collected items.

The pupil dilation metrics correlates with the Nasa-TLX score, showing that this index is sensitive to both objective and subjective estimates of mental workload (**H2**).

Pupil dilation and Nasa-TLX scores are both correlated with the performance measures. They are negatively correlated with the number of collected items and positively correlated with the number of deaths, confirming again that the participants were indeed in a situation of overload, which impaired their performances.

Evaluation of the EEG indices

As revealed by the Wilcoxon's tests, the EEG workload and engagement indices were not sensitive enough to our task manipulation of mental workload. The lack of results might be due an important inter-individual variability. We made the choice of recruiting few participants only.

For the workload index, the model is based on a group of individuals independently of our study and is not adjusted to each participant. The consumption of nicotine or caffeine thus has even more impact, potentially leading to ceiling effects masking the B-Alert metric sensitivity. For the engagement index, we interpret the lack of difference between the easy and hard condition as the fact that both conditions require a similar amount of perceptual and attention related processing.

When confronting the measures, however, we can observe that the EEG workload index is correlated with the pupil dilation metrics and the NasaTLX score (**H4**), suggesting that it is sensitive to the same underlying phenomena.

Conclusion

We conducted an experiment using an ecological gaming situation to reproduce an operational environment. Subjective and physiological measures of mental workload suggest that we succeeded in placing the subjects in an overload situation.

Our analysis indicates that the B-Alert system does not capture the variation of mental workload as can be observed with the pupil dilation and the subjective measures. However, the B-Alert workload index is correlated with both of these measures, suggesting that this index varied consistently with our hypothesis, but was not sensitive enough to capture our mental workload variation.

The B-Alert system might be a useful option depending on the environmental conditions: the workload index is not sensitive to a change of brightness (p=1), whereas pupil dilation is (p=.009). Nonetheless, our results show that the B-Alert metrics are not precise enough to offer a reliable option in operational conditions.

References

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*(4), 425–438.

Backs, R. W., Lenneman, J. K., & Sicard, J. L. (1999). The Use of Autonomic Components to Improve Cardiovascular Assessment of Mental Workload in Flight.. *The International Journal of Aviation Psychology*, *9*(1), 33–47.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., … Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*(Supplement 1), B231–B244.

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., … Stibler, K. (2005). Evaluation of an EEG workload model in an Aegis simulation environment. In *Defense and Security* (pp. 90–99). International Society for Optics and Photonics.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2012). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*.

Cegarra, J., & Chevalier, A. (2008). The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. *Behavior Research Methods*, *40*(4), 988–1000.

Cegarra, J., & Morgado, N. (2009). Étude des propriétés de la version francophone du NASATLX. In *Communication présentée à la cinquième édition du colloque de psychologie ergonomique (Epique)*.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., & Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology*, *87*(2), 241–250.

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics, Third Edition*, 243–268.

Wickens, C. D. (1992). *Engineering psychology and human performance* . HarperCollins Publishers.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, *12*(1), 3–18.

‘WE NEED PRIORITY PLEASE’
MITIGATED SPEECH IN THE CRASH OF AVIANCA FLIGHT 052

Simon Cookson
J. F. Oberlin University
Tokyo, Japan

On 25 January 1990, Avianca Flight 052 was flying from Columbia to the United States when it crashed after a missed approach to JFK Airport in New York. The direct cause of the accident was fuel exhaustion but the NTSB investigation identified multiple causal factors. The Avianca captain, who was flying the aircraft, repeatedly instructed the first officer to notify ATC about the fuel emergency. The first officer, however, did not use the word ‘emergency’ but instead requested ‘priority’ and told ATC that the airplane was ‘running out of fuel’. Why did the first officer mitigate the captain’s instructions? This paper hypothesizes that a range of factors relating to national culture, professional culture, organizational culture and stress may have contributed to the first officer’s use of mitigated speech. The implication is that the communication breakdown was not simply caused by inadequate English language proficiency.

This paper examines English language communication problems experienced by the flight crew of Avianca Flight 052, prior to its crash at Cove Neck, near New York, on 25 January 1990. The direct cause of the crash was fuel exhaustion, but the National Transportation and Safety Board (NTSB) investigation found that the accident occurred as a result of multiple causal factors. Two of the factors were directly related to communication: the crew’s failure to declare a fuel emergency, and the lack of standardized terminology for minimum and emergency fuel states. The accident is summarised in Table 1.

Table 1.
Summary of the Avianca 052 accident.

Date of accident	25 January 1990
Location	Cove Neck, Long Island, New York, USA
Aircraft type	Boeing B-707-321B
Operator & flight number	Avianca Flight 052
L1 of flight crew	Spanish
L1 of air traffic controllers	English
Type of accident	Fuel exhaustion
Number of fatalities	73

Note. L1 is an abbreviation of first language, or mother tongue.

Despite taking place more than 25 years ago, the Avianca 052 accident was cited in justification of the major English language program that ICAO brought into full effect in 2011 to improve the English proficiency of commercial pilots and air traffic controllers around the world. The official manual for the program, ICAO Doc 9835, refers to the Avianca crash as one of several major accidents in which ‘insufficient English language proficiency on the part of the flight crew or a controller had played a contributing role’ (ICAO, 2010, p. 1-1).

A few years after the accident, Helmreich (1994) provided a wide-ranging and thought-provoking analysis from a system perspective, using data uncovered in the litigation that followed the accident to try to explain the causal factors identified by the NTSB investigation. His paper examined organizational factors such as training, maintenance and dispatch operations at Avianca, as well as the group dynamics of the crew during the flight. This paper leans heavily on Helmreich’s analysis, but is more narrowly focused on the communication breakdown that occurred between the flight crew and air traffic controllers. It does not address group processes within the Avianca cockpit. The sources of data for this analysis are the cockpit voice recorder (CVR) and air traffic control (ATC) transcripts in the NTSB accident report. It is important to note that the CVR data covered only the final 40 minutes of the flight.

The Accident

On 25 January 1990, Avianca Flight 052 was scheduled to fly from Bogota to Medellin in Columbia, then – after a short stop for refuelling – on to John F. Kennedy International Airport (JFK) in New York. The weather was poor in the north-eastern part of the United States, which meant that the aircraft had to enter three separate holding patterns for a total of 77 minutes. During the third holding period, at 20:46 EST, the flight crew notified ATC that they could only hold for about five more minutes and could no longer reach their alternate in Boston because they were running out of fuel. As the aircraft finally descended towards JFK it encountered wind shear and the crew executed a missed approach at 21:23. While trying to return for a second approach, all four engines suffered a loss of power and the aircraft crashed at approximately 21:34 at Cove Neck, Long Island. Of 158 passengers and crew on board the plane, 73 died as a result of the crash. The fatalities included all the flight and cabin crew, with the exception of one flight attendant.

NTSB Investigation

The NTSB investigated the accident and published its report in April 1991. The conclusion of the report listed 24 findings, before stating one probable cause plus a number of contributory factors. The probable cause 'was the failure of the flightcrew to adequately manage the airplane's fuel load, and their failure to communicate an emergency fuel situation to air traffic control before fuel exhaustion occurred' (NTSB, 1991, p. 76). Table 2 has a summary of the probable cause and contributory factors.

Table 2.
Causal factors for the Avianca 052 accident (NTSB, 1991).

Probable cause	flight crew failed to adequately manage fuel load
	flight crew failed to communicate emergency fuel situation to ATC
Contributory factors	flight crew failed to make use of airline operational control dispatch system
	FAA traffic flow management was inadequate
	lack of standardized, understandable terminology for minimum / emergency fuel states
	first approach to JFK hindered by windshear
	first approach hindered by flight crew fatigue
	first approach hindered by flight crew stress

The accident was clearly caused by the co-incidence of multiple factors, and might have been averted if any of these factors had been absent. The NTSB report notes, for example, that the accident would not have occurred if the crew had not been prevented from successfully completing the first approach by a combination of wind shear, stress and fatigue. Two of the causal factors relate directly to English language communication: the flight crew's failure to communicate an emergency fuel situation to ATC; and the lack of standardized, understandable terminology for minimum and emergency fuel states.

Flight Crew

The flight crew of Avianca 052 consisted of the captain, first officer and flight engineer. The 51-year-old captain was a very experienced pilot, with no record of previous accidents, and was also a pilot in the Columbian Air Force Reserve. The 45-year-old flight engineer was likewise very experienced, and had more than 3,000 flight hours in the Boeing 707. By contrast, the young first officer was, as Helmreich (1994, p. 280) notes, 'inexperienced overall and particularly in the B-707', with just 64 hours in this aircraft type and a total flight time of 1,837 hours. Table 3 gives details of the crew's flight experience.

The Avianca captain had previously flown on international flights with the first officer, and also with the flight engineer, but this was the first time that all three flew together as a crew (NTSB, 1991). Citing NTSB research, Helmreich and Merritt (1998, p. 12) note that 'a disproportionate percentage of accidents happen to crews who are flying together for the first time'.

Table 3.
Flight experience of the Avianca 052 crew (NTSB, 1991).

	Age	Flight hours		Flight hours in B707		Flights to NY 1989-90	
		Total	Night flying	Total	Night flying	Total	B707
Captain	51	16,787	2,435	1,534	478	14	14
First officer	28	1,837	408	64	13	13	5
Flight engineer	45	10,134	2,986	3,077	1,062	7	5

Communication with ATC

During the final stages of the flight, the Avianca first officer was communicating with controllers in the New York Air Route Traffic Control Center (ARTCC) and Terminal Radar Approach Control (TRACON). While Flight 052 was still in the third holding period, at 20:44, an ARTCC controller informed the crew they could expect further clearance information at 21:05. The first officer read back the time and said, 'I think we need priority we're passing [unintelligible].' This exchange is recorded in the ATC transcript (NTSB, 1991, pp. 177-179) but not in the CVR data because the latter covered only the last 40 minutes of the flight. At 20:46 the first officer reported to ATC that they could only hold for about five more minutes and, when asked to repeat the alternate, he said, 'It was Boston but we we can't do it now we, we, don't, we run out of fuel now.'

Thus the first officer informed ARTCC about the fuel problem more than 45 minutes before the crash occurred. Crucially, though, neither then nor later did he declare a fuel emergency. His message about the fuel problem and being unable to reach the alternate was not passed on to New York approach control because the handoff controller was on the phone when the transmission was made and so he did not hear it. The aircraft subsequently received routine vectors including a 360° turn for spacing.

The Avianca crew did not contact ATC again about their fuel situation until after the first landing attempt failed at 21:23. Then, during the final eleven minutes of the flight, there were several exchanges about the fuel problem. The CVR transcript indicates that the captain three times declared in Spanish 'we don't have fuel'. In terms of communicative functions, at 21:23:43 the captain gave information ('we don't have fue-'); at 21:25:08 he gave an order to the first officer ('advise him we don't have fuel'); and at 21:25:28 he requested confirmation ('did you already advise that we don't have fuel'). In response to these instructions and other prompts by the captain, the first officer three times said to the controller in English 'we're running out of fuel'. The captain also twice instructed the first officer to declare an emergency: at 21:24:06 ('tell them we are in emergency') and at 21:24:22 ('advise him we are in emergency'). The first officer did not do so. Finally, after two engines had flamed out, the first officer made a request at 21:32:49: 'we need priority please'.

Mitigated Speech

A mitigated form of speech may be defined as 'one which expresses a given propositional content in such a way as to avoid giving offense' (Linde, 1985). During the final section of the flight the captain repeatedly instructed the first officer to notify ATC about the fuel emergency. However, the first officer used mitigated speech in his messages to ATC: he did not use the word 'emergency' but instead requested 'priority', and he told controllers that the airplane was 'running out of fuel'. Cushing (1994, p. 2) observes that the accident was in part due to the first officer using 'the formal English phrase *running out of fuel* rather than the technical aviation term *emergency*, thereby failing to convey to the controller the intended degree of urgency.' Why did the first officer use mitigated speech, when he could have simply translated the Spanish word 'emergencia' into its English equivalent?

Cultural Factors

Helmreich (1994) hypothesises that the behavior of the Columbian flight crew can be attributed at least partly to national culture, and makes use of the cultural dimensions identified by Hofstede (1980) in his analysis of the accident. Two of these cultural dimensions are relevant to the communication between the flight crew and ATC: collectivism-individualism and power distance. In addition to national culture, both the professional culture and organizational culture of the flight crew may also have contributed to the communication breakdown.

National Culture

The cultural dimension of collectivism-individualism is a measure of the degree to which people act as members of cohesive groups rather than as individuals. Coming from a strongly collectivist culture in Columbia, Helmreich (1994) suggests that the Avianca flight crew may have been reluctant to declare an emergency and push themselves ahead of other crews that they perceived to be in a similar situation. In other words, a strong sense of collectivism may have made the first officer reluctant to use the word 'emergency'. Other aircraft were indeed running low on fuel that night. At 21:02 an American Airlines crew transmitted the following: 'American six ninety two I want to advise you we're at minimum fuel uh we're uh about uh twelve or fourteen minutes from declaring an emergency' (NTSB, 1991, p. 219).

Helmreich (1994) also observes that power distance, or the degree to which people accept unequal power relationships, is typically high in Columbia. In a lengthy discussion of the Avianca accident, Gladwell (2008, pp. 192-209) develops this idea, noting that authority is highly respected in Columbian society, and explaining that the first officer – only 28 years old and lacking flight experience – would have seen himself as subordinate to both the captain and the 'domineering Kennedy Airport air traffic controllers'. Gladwell (2008, p. 194) suggests that the first officer, in deference to the authority of the captain and controllers, used mitigated speech 'to downplay or sugarcoat the meaning' of his communications.

Professional Culture

The use of mitigated speech within the pilot community was studied by researchers in the 1980s and 1990s. Using data from eight airline accidents that occurred between 1972 and 1982, Linde (1985) reports that requests made by subordinates to superiors were more mitigated than those made by superiors, and requests were less mitigated during emergency situations. Linde goes on to discuss politeness theory and the concepts of negative face and positive face. Requests are speech acts that threaten the negative face of hearers by pressuring them to act and restricting their freedom of action. Speakers may lessen the damage to hearers by using indirect requests. Applying this reasoning to the case of Avianca 052, it is possible that the first officer, rather than declaring an emergency, instead used an indirect request ('we need priority please') as a strategy to lessen the imposition on ATC.

In another study involving several hundred pilots from the United States and Europe, Fischer and Orasanu (1999) report that first officers were more likely than captains to use indirect communication strategies such as hints, and communications were more direct in emergency situations than in normal flight. This and the previous research involved intra-cockpit communications between native-speaker pilots. By contrast, the Avianca accident featured native speaker air traffic controllers communicating with a first officer who was a non-native speaker of English. Nevertheless, both studies highlight a tendency for first officers to use mitigated speech, and they also suggest that native speaker controllers may expect pilots to use direct communications in emergency situations.

Organizational Culture

Training at Avianca, the airline for which the pilots worked, may have led the first officer to believe that the terms 'emergency' and 'priority' were interchangeable. The NTSB (1991, p. 63) report includes the testimony of another Avianca captain who stated that training provided by Boeing gave the impression that 'the words priority and emergency conveyed the same meaning to air traffic control'. Indeed, Boeing issued a bulletin to all B-707 operators in 1980 advising that during operations with very low fuel quantities 'priority handling from ATC should be requested' (NTSB 1991, p. 28). This terminology was critical but the correct usage was ambiguous. The Avianca crew may have believed that 'priority' conveyed the same meaning as 'emergency', but air traffic controllers questioned during the investigation stated that only the terms 'Mayday', 'Pan-pan-pan' or 'Emergency' should be used to declare a fuel emergency (NTSB 1991, p. 63). In a human factors analysis, Krause (2003 pp. 90-107), noting that the first officer twice asked for 'priority' and four times advised ATC the plane was low on fuel, states that 'it would seem reasonable and logical' for the controllers to have asked for clarification.

Effects of Stress

Earlier in the day the Avianca crew had flown a 54-minute leg in Columbia from Bogota to Medellin. Following that, the actual flight time from Medellin to JFK was 6 hours 26 minutes, much longer than the planned time of 4 hours 40 minutes. The aircraft was in three separate holding patterns for a total of 77 minutes. By contrast, another Avianca captain interviewed after the crash stated that holding delays for JFK were normally 'a maximum

of 20 to 30 minutes' (Cushman, 1990). The unexpectedly long delays, mechanical problems with the aircraft's autopilot and flight director, the worsening fuel problem, adverse weather including reports of windshear, and the missed first approach would all have increased the stress on the flight crew.

Furthermore, the first officer was probably suffering additional stress due to his lack of flight experience. With only 13 hours of night flying experience in the Boeing 707, it was unlikely that he had ever faced a situation like this before. The cognitive processing demands of having to cope with a novel and difficult situation would have been considerable, compounded by the need to continually code switch between Spanish (to talk with the other crew members) and English (for the ATC communications).

As noted in Table 2, the investigation found that flight crew stress and fatigue contributed to the missed first approach. The NTSB report briefly mentions laboratory experiments which show that demanding flight conditions cause communication performance to be significantly degraded, but does not examine the effects of stress in detail. Since the Avianca accident a lot more research has been carried out into the impact of stress and fatigue on flight operations, and it is now evident that the effects are complex and multi-faceted. Two stress-related effects may have impacted upon the first officer's communications: attentional tunneling and regression to earlier behavior.

Attentional Tunneling

Stokes and Kite (1994, pp. 112-116) describe a number of ways in which communication can be degraded by stress, and they observe that the degradation may manifest itself in a 'decreased ability to receive and interpret messages'. They note that stress can cause pilots to 'miss advice, information, or instructions from ATC or flight deck colleagues' due to working memory limitations and attentional tunnelling, or a narrowing of the field of attention. The CVR data indicates that the first officer was having problems processing information accurately: five times between 21:25 and 21:26 he incorrectly reported a new heading of 'zero eight zero' (=080°) as 'ciento ochenta' (=180°). Under these stressful conditions, it is possible that attentional tunneling prevented the first officer from understanding the significance of the word 'emergency' in the captain's instructions. His subsequent use of the word 'priority' as a synonym for 'emergency' might have been the result of him reducing the criteria for accuracy, which is one way that individuals deal with the demands of multiple concurrent tasks (Loukopoulos et al., 2009).

Regression to Earlier Behavior

The first officer lacked flight experience but according to Helmreich (1994, p. 272) he spoke 'excellent, unaccented English'. Well-learned tasks can be carried out automatically with little cognitive effort, but new or recently-learned tasks may require considerable controlled processing (Loukopoulos et al., 2009). When speaking English, the first officer could probably produce well-learned conversational structures automatically, but he may have needed considerable cognitive effort to produce recently-learned standard phraseology. Under conditions of high stress and workload, the difference between these two types of speech, and the mental workload they require, becomes critical. Stokes and Kite (1994, p. 65) observe that 'individuals under stress are prone to revert to behaviours, strategies, and schemata learned earlier'. Such regression to earlier behavior may have led the young first officer to use conversational structures (such as mitigated speech) instead of phraseology. This tendency could have been reinforced by the first officer hearing plain language being used by other flight crews and controllers, a number of instances of which are recorded in the ATC transcripts. For example, the NTSB (1991, p. 222) report shows that at 21:06 the following exchange took place:

TRACON controller:	'American six ninety two how are we making out'
AAL 692:	'We got enough fuel for the approach and landing and that's it'
TRACON controller:	'Ok understand'

Hypothesis

Why did the first officer use mitigated speech to communicate with ATC? The hypothesis put forward in this paper is that a combination of factors created conditions conducive to his use of mitigated speech. First, a strong sense of collectivism may have made the first officer reluctant to declare an emergency when other flight crews were also in difficulty. High power distance made it more likely for him to use mitigated speech with the controllers. Being a first officer, not a captain, he was more likely to use mitigated speech and may have used it as a strategy to lessen the imposition on ATC. Training in his organization may have led him to think that the word 'priority' carried the same meaning as the word 'emergency'. A high level of stress may have caused attentional tunneling making it difficult for the first officer to interpret the captain's instructions accurately and to comprehend the distinction

between ‘emergency’ and ‘priority’. Finally, stress-related regression may have led him to use conversational speech forms including mitigated speech rather than standardized phraseology in his communication with controllers.

Helmreich (1994, pp. 271-272) stresses that ATC did not realize how serious the Avianca fuel situation was, and he lists several ways in which the first officer’s transmissions misled the controllers: he did not declare an emergency; he communicated about the fuel problem ‘in an offhand manner’; his English was excellent; and he spoke in a ‘monotone voice’. It is now possible to add one more factor that may have misled ATC: the first officer’s use of mitigated speech signalled to the controllers that the fuel problem was not severe since research indicates that communications are *more direct* in emergency situations (Linde, 1985; Fischer & Orasanu, 1999).

Conclusion

Research conducted since the crash of Avianca 052 allows a new hypothesis to be out forward to explain why the first officer used mitigated speech to communicate with ATC. This paper suggests that a combination of factors relating to national culture, professional culture, organizational culture and stress led him to use mitigated speech forms. The implication is that the communication breakdown between this flight crew and ATC was not simply due to inadequate English proficiency. Indeed, as noted above, the first officer’s English proficiency was high. This was a system accident involving numerous causal factors ranging from mechanical problems to adverse weather and fatigue. A combination of factors coincided to leave an inexperienced pilot, who was simply trying to do his job, critically exposed. The Avianca 052 crash was a tragic accident, and this paper highlights the need for continued research into the effects of stress and culture on communication.

References

Cushing, S. (1994). *Fatal Words: Communication Clashes and Aircraft Crashes*. Chicago: The University of Chicago Press.

Cushman, J. H. (1990). *Avianca Flight 52: The Delays That Ended in Disaster*. The New York Times. 5 February 1990, Section B, p. 6.

Fischer, U., & Orasanu, J. (1999). *Cultural Diversity and Crew Communication*. Paper presented at 50th Astronautical Congress in Amsterdam, October 1999, published by American Institute of Astronautics & Astronautics.

Gladwell, M. (2008). *Outliers: The Story of Success.* London: Allen Lane, Penguin Books Ltd.

Helmreich, R. L. (1994). Anatomy of a System Accident: The Crash of Avianca Flight 052. *The International Journal of Aviation Psychology.* 4/3, pp. 265-284.

Helmreich, R. L., & Merritt, A. C. (1998). *Culture at work in aviation and medicine: National, organizational and professional influences*. Farnham, Surrey: Ashgate Publishing.

Hofstede, G. (1980). *Culture’s Consequences: International Differences in Work-Related Values*. Beverley Hills: Sage

International Civil Aviation Organization (ICAO) (2010). *Manual on the Implementation of ICAO Language Proficiency Requirements*. ICAO Document 9835, 2nd edition. Montreal, Canada: ICAO.

Krause, S. S. (2003). *Aircraft Safety: Accident Investigations, Analyses, and Applications.* 2nd edition. New York: McGraw-Hill.

Linde, C. (1985). *The Quantitative Study of Communicational Success: Politeness and Accidents in Aviation Discourse*. Paper presented at the Annual Meeting of the Linguistic Society of America, Seattle, WA, December 27th-30th, 1985.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009). *The Multitasking Myth: Handling Complexity in Real-World Operations.* Farnham, UK: Ashgate Publishing Limited.

National Transportation Safety Board (NTSB) (1991). *Aircraft Accident Report: Avianca, The Airline Of Columbia, Boeing 707-321 B, HK 2016, Fuel Exhaustion, Cove Neck, New York, January 25, 1990*. NTSB/AAR-91/04. Washington, D.C.: NTSB.

Stokes, A., & Kite, K. (1994). *Flight Stress: Stress, Fatigue, and Performance in Aviation*. Hampshire, England: Avebury Aviation.

THE EFFECT OF ASYNCHRONOUS DATA ON PILOT-CONTROLLER COMMUNICATION IN A DYNAMIC ENVIRONMENT WITH SUBJECT-MATTER EXPERTS

Samuel Lien
University of Waterloo
Waterloo, ON, Canada
Jonathan Histon
University of Waterloo
Waterloo, ON, Canada

Integrating Unmanned Aerial System (UAS) into controlled airspace may introduce communication challenges if there are time delays associated with the distribution of a common surveillance source of those UAS. Termed "information asynchrony" by Yuan et al (2012), an earlier, static image study showed large time delays had an observable impact on controller-pilot communication, but the effect was not present for time delays of less than 1 minute. A follow-up study is being conducted using an online ATC-flight simulator with professional pilots and controllers as participants. Effects on communication are being analyzed objectively through measurable characteristics of communication breakdown, and subjectively through trial questionnaire and survey. Limited results to-date showed no observable effects on pilot-controller communication with time delays of less than 100 seconds and illustrated the challenge of identifying measures robust to the inherent variability in working methods and communication styles.

Civil applications of Unmanned Aerial Systems (UAS) such as security surveillance, disaster response and aerial photography are steadily increasing. The Federal Aviation Administration's 5-year roadmap on the integration of UAS into National Airspace stated that more research will be needed on procedures and operating rules for UAS (FAA, 2013). Smaller, non-cooperative UAS, that do not have the capacity or desire to participate in traditional surveillance techniques (such as secondary radar, or ADS-B) are generating significant public and research interest. New technologies, procedures and operation rules being developed will need to take into account how surveillance data about these vehicles is distributed to both pilots and controllers. Asymmetric time delays, termed "information asynchrony" by Yuan et al (2012), can occur in the distribution of any surveillance data from a common source to a pilot and a controller. Challenges can arise if different surveillance sources are used, or there are time delays associated with the distribution of a common surveillance source.

Yuan et al (2012), in a preliminary study using static pictures and naïve university students as participants, showed that longer time delay values had a clear effect on pilot-controller communication. The operator with the most up-to-date information had a consistently better communication experience with less frustration, better communication effectiveness and performance. However, the method used had several limitations including the lack of time pressure on participants as a result of static radar displays, and the use of university students as participants lacking professional experience, knowledge and judgement.

In order to investigate the effects of information asynchrony at shorter time delays and address the previous study's limiations, a follow-up, dynamic study, with subject-matter-experts as participants has been developed and data-collection is ongoing. This paper reviews previous works on asynchrony, describes the experiment design and presents findings to-date.

Previous Work

There has been significant work in the past on asynchrony in the control loop with a majority of findings showing that latency in feedback can have detrimental effects performance such as increased errors and movement time in tasks involving target acquisition and telemanipulation (MacKensie 1993, Currie & Rochlies, 2004, Lum et al, 2009). This is also true in collaboration tasks, for example, where latency caused increased time to completion, and over and undershoot errors over targets during multiple robot manipulator control tasks, as well as collaboration breakdowns from jigsaw-puzzle task (Allison et al, 2004, Gergle et al, 2006).

In aviation, numerous studies have also looked into data asynchrony topics and found similar degradation effect on pilot-controller communication causing misunderstandings and degrading operator's performance. For example, Nadler et al. (2009) found increasing transmission delays on air traffic control communications can cause communication blocks and thus lapses in transfer of critical information due to simultaneous transmission between pilots and controllers. This was one of the identified causes to the Tenerife disaster (Nalder et al., 2009). Also, actual incidents such as the crash of a Eurocopter AS350 and Piper PA-32-360 due to NEXRAD weather imagery of more than five minutes old also serve as examples of the degradation effects of time delays (NTSB, 2011).

However, a limitation in these studies on asynchrony was that they were conducted with a singlular delay in the feedback loop, such as a delay in operator responding, or a one-way propagation delay in transmission. Little research has been done to date in the case of two or more parties receiving data from the same source with different latency applied. For example, surveillance data on non-cooperative objects (UAS, birdflocks) may be passed to a System Wide Information management (SWIM) architecture, where data may be processed through different system paths to be delivered to pilots and controllers. Each path will have different amount of inherent latency due to hardware, software, human operations and cognitive complexity in information processing (Yuan et al, 2013, Yuan and Histon, 2014). Consequently, pilots and controllers may have a different situation awareness due to the presence of information asynchrony; this would be expected to affect their communications (Yuan & Histon, 2014). Understanding the conditions (e.g. amount of time delay) that trigger such effects is important for the design of future surveillance distribution systems affecting the integration of UAS into controlled airspace.

A Dynamic Study of Information Asynchony

To address this challenge, a dynamic study utilizing experienced subject matter experts has been developed. The study uses a simulated environment to provide time pressure and is recruiting professional pilots and controllers as subject-matter experts. The research objectives of the follow-up study include: 1) to identify any observable effect of information asynchrony on pilot-controller communication applied to non-cooperative surveillance data, and 2) to identify if the effects differ between the pilot versus controller role, and 3) to identify if the effect depends on whether a participant is in the role that is ahead (no time delay applied) or behind (experiencing the applied time delay).

Methodology

In the study two subject matter experts are paired together to perform in an online simulator session while sitting at home using own choice of computers. One participant assumed the role of a pilot while the other assumed the role of a Terminal (TRACON) controller. By live-streaming a video broadcast of simulator displays, participants were provided a dynamic, real-time navigation display and ATC radar display for pilot and controller participants, respectively. They were instructed to communicate with each other to resolve a potential traffic conflict in each trial. In each trial a time delay value was applied to only the non-cooperative objects on one of the pilot or the controller's display.

Experiment Setup, Task Details, Scenario Design, and Time Delay Choices

A schematic of the experiment setup is shown in Figure 1. The experiment moderator interacts with the pilot and controller participant on the web from a physical lab room through Google Hangout. The web platform delivers the required simulator video feed separately to each participant.

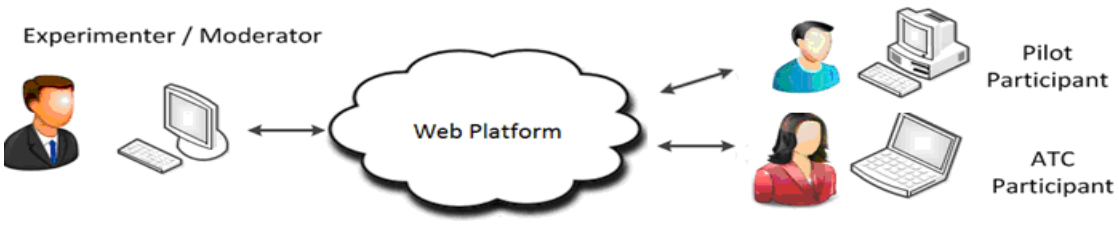


Figure 1. Experiment Setting in an Web Environment.

Participants began each trial reading a brief paragraph on the traffic situation of the scenario, customized for their respective role. Participants monitored their respective display and communicated with each other following standard procedures as they would do their in professional work. The controller participant was responsible for the safety of all airline traffic, while the pilot participant was responsible for the safety of their own aircraft. The goal for both parties was to avoid colliding with non-cooperative objects including UAS and birdflocks. Surveillance data consisted of normal airlines traffic (cooperative), UAS and bird flock (both non-cooperative). For each scenario, there were eight cooperative and four non-cooperative surveilance object. No weather data were presented in the study. The moderator implemented all commands given by the controller to the pilot participant and "imaginary pilots" of other airliners on the simulator.

Five departure and five takeoff scenarios were designed and each uniquely and randomly assigned to nine time delay values of (96c, 48c, 12c, 6c, 0, 6p, 12p, 48p and 96p) applied for each participant pair. The time delays were in seconds and the "c" and "p" designate either controller or pilot participant's screen has the delay applied to the non-cooperative objects. Two additional generic training scenarios, one arrival and one departure, were designed as well. Thus, for a simulator session, there were two training and nine formal trials.

Participants and Training

The required audience for recruitment included active or retired commercial pilots with at least commercial licence (CPL) or above, as well as active or retired air traffic enroute or terminal (TRACON) controller. As of March 1st, five groups (five pilots and five controllers with ages between 27-65) have participated in the study. The effect of gender was not controlled due to very few female participants signing up. Participants were paired based on schedule availability and did not previously know or have worked with each other. Training preparation included briefing documents and two training scenarios for familiarization of simualtor procedures. The training scenarios were repeated depending on participants' experience and familiarization levels.

Data Collection

Objective Data. Audio transcripts were collected for each trial. The transcripts were analyzed for indications of communication breakdowns including the number of phrases: 1) expressing conflicting understanding, 2) asking for clarifications, 3) seeking confirmations, 4) expressing confusion with respect to the traffic situation, 5) spoken in an urgent tone, 6) making direct "interventional" commands. Timing data was also extracted from the audio transcript showing when and the length of particular phrase of interest.

Subjective Data. A demographic questionnaire was administered to gather information on the background of each participant. A self-rated post-trial questionnaire was administered at the end of each trial, with 7-point Likert scale "agree-disagree" self-rating questions on 1) confusion, 2) awareness of own traffic situation, 3) awareness of other's traffic situation, 4) communication effectiveness, 5) controller's satisfaction with the pilot's reaction to the initial resolution from controller, or pilot's satisfaction with the controllers' issued initial resolution, and 6) controller's satisfaction on pilots reaction to final resolution from controller, or pilot's satisfaction to a controllers' final resolution . A post-experiment survey was administered asking about participant's insight and professional experience with information asynchrony.

Other Data. Latency data was collected as the simulation was done online, by using an online clock broadcasted on the participant screen to count out loud while noting down the discrepency.

Preliminary Results

For space reasons only examples of the objective communications data analysis are shown. For each trial, one coder listened to the recorded audio data and determined the total number of communication events, coding them according to the six categories discussed above. Due to the limited number of participants, initial results are presented in the form of box and whisker plots to emphasize the spread in the data collected to date. Boxes show the inter-quartile range (Q1-Q3) while whiskers show the minimum and maximum observed values.

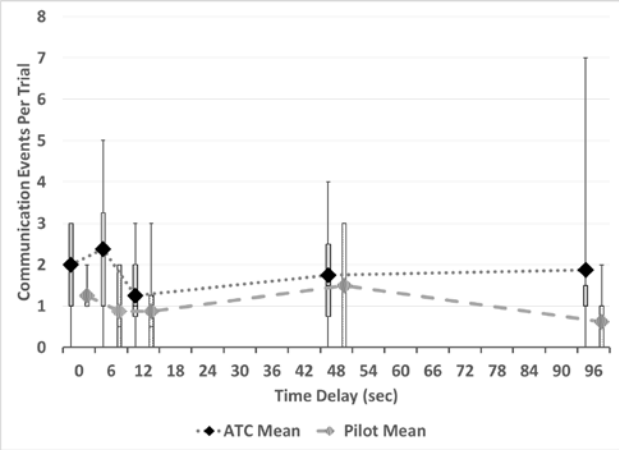


*Figure 2.*Communication rates per trial as time delay increases for pilot and controller.

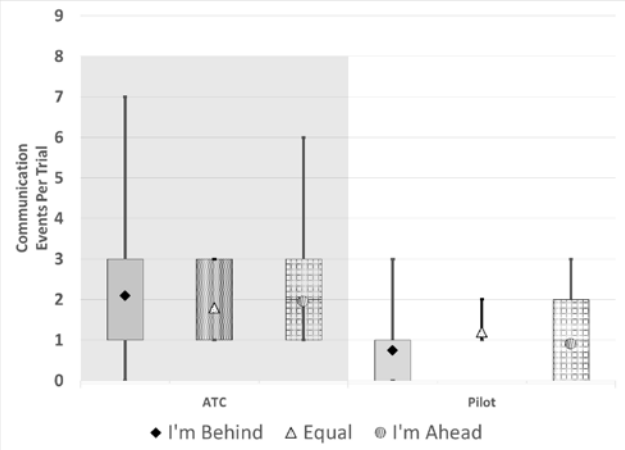


*Figure 3.*Communication rates per trial as time delay increases for pilot and controller grouped by behind , no delay, ahead.

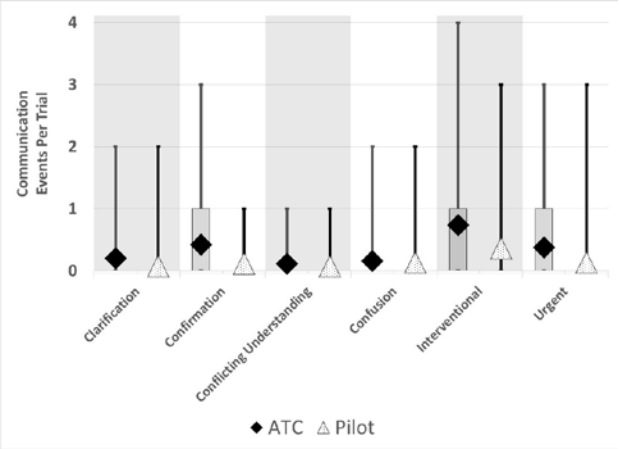


*Figure 4.*Communication rates per trial for pilot and controller grouped by individual types of communication events.

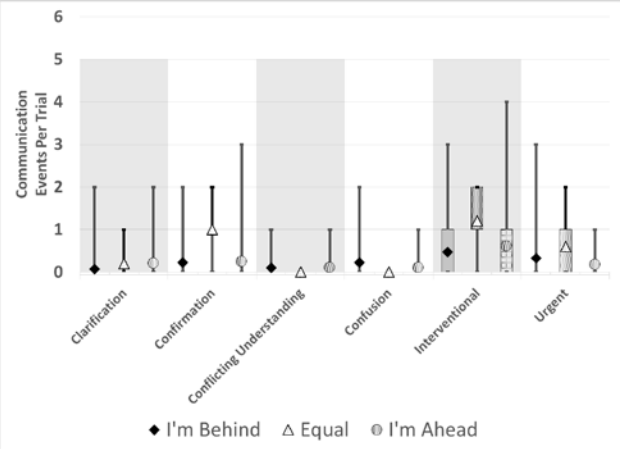


*Figure 5.*Communication rates per trial for each communication events overall, grouped by behind, no delay, ahead.

Figure 2 presents the average number of communication events per trial for all participants; results are shown separately for pilots and air traffic contollers. The time delay is in the x-axis and does not distinguish between which role (controller/ pilot) is experiencing the time delay. Three preliminary trends are found in the figure. First, the average number of communication events appears to be higher for air traffic controllers compared to pilots; this is not surprising given the traditional division of tasks between pilots and controllers.

Second, time delay does not appear to be having any impact on the total number of communication events, irrespective of the role (pilot / controller). This suggests the effects of time delay may be subtle, without causing dramatic changes in the communication behaviour. It may also be that the effects manifiests themselves more dramatically only in particular aircraft configurations (e.g. scenarios); this is an area that needs further investigation.

The relatively low number of events per trial suggests other, more sensitive, measures beyond the communication events analyzed so far also need to be considered. Alternatives such as examining each pairs event counts as a difference from the count in a baseline (no delay) condition will be considered in the future.

Third, the previous two points need to be considered relative to the wide spread in the results collected to date. There are wide absolute ranges (spread of whiskers) and large inter-quartile ranges in Figure 2. However,

Figure 3 also shows the challenges of representing low count discrete data with box plots as there are several places where the inter-quartile range is either non-existent (Q1=Q3) or the maximum observed value corresponds with Q3 and hence no whisker is visible. Alternative visualizations will be considered in future to show the spread of the observed rates.

Analyses are also being conducted to investigate the hypothesis that there may be differences in the effect of time delay depending on whether the participants is the one experiencing the time delay or not. A distinction is drawn for each participant between trials where they are receiving up to date data (e.g. not experiencing a time delay, "I'm ahead" in the following figures) and trials where they are receiving the delayed data (e.g. are experiencing a time delay, "I'm behind " in the following figures). As shown in Figure 3, results to date indicate there may be a slight effect for air traffic controllers with I'm ahead and I'm behind showing slightly higher average communication event rates. The situation is reversed for pilots, with equal time delays having the highest average communication event rate. What is also clear from Figure 3, is that the spread in the data collected limits the confidence in these effects until much more data has been collected.

In order to examine whether effects of time delay are limited to only a subset of the communication events, Figure 4 breaks out the rates of the individual types of communication events. Unsurprisingly, interventional events are the highest communication type for both air traffic controllers and pilots. When the rate of event types are examined with respect to time delay experienced (Figure 5), somewhat surprisingly interventional events are the highest in the no time delay ("Equal") condition. This is also true for the Urgent and Confirmation event types. More data collection, as well as examination of the association of specific scenarios with the no time delay ("Equal") condition, will be performed to investigate this further.

Discussion, Limitations, and Future Work

The limited results to-date showing no observable effect on pilot-controller illustrate the challenge of identifying measures robust to the inherent variability in working methods and communication styles. The limited results to-date showing no observable effect of time delay may point to several limitations of this study. First, only a limited number of pairs have been run. Second, it was very difficut to maintain a consistent procedure tailored to the collision designed into each scenario. This appeared to be due in part to the control strategies used by participants due to their different background, experience and training from region of work (e.g. Africa vs North America). Thirdly, the absence of the visual, out-the-window view for pilots, and the ability to listen to radio chatter from other aircrafts, limited cues for more accurate identification of object location and decision-making. Consequently, results to date have illustrated the challenge of discerning an effect due to the high variability in participant responses.

Finally, in a dynamic environment, it is now recognized that time delay may not have direct effect on communications. Rather, it has been recognized that time delay ultimately manifests itself in a difference in the physical location of an object on the display screen. Pilots and controllers directly observe the history and current positions of surveilled objects (assuming no out-the-window view) to form a mental model of the situation and behavior of the aircrafts (Reynolds, 2006). The change in physical location between displayed and 'actual' will be a function of the speed of the object; thus, rather than designing the experiment around consistent time delay values, it may be more appropriate to design for consistent screen distance impacts (e.g. pixel difference on the display screen as seen by participants). As screen distance depends on the speed of the non-cooperative object in question and the ratio of the size of the radar screen, limitations with online experiment environment, such as participants use of different monitors for the study, and fluctuating network latency, may have affected the accuracy of captured effect on communication mapped to each time delay measurement.

References

Allison, R. S., Zacher, J. E., Wang, D., & Shu, J. (2004). Effects of network delay on a collaborative motor task with telehaptic and televisual feedback. *Proceedings VRCAI 2004 - ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry, New York, NY, USA*, 375–381. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.72.5771.

Currie, N. J., & Rochlis, J. (2004). Command and telemetry latency effects on operator performance during Interational Space Station robotic operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting September 2004,* 48: 66-70, doi:10.1177/154193120404800115.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2006). The impact of delayed visual feedback on collaborative performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06*, 1303. doi:10.1145/1124772.1124968.

Lum, M.J.H., Rosen, J., King, H., Friedman, D.C.W., Lendvay, T.S, Wright, A.S., … (2009). Teleoperation in surgical robotics -- network latency effects on surgical performance. *31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009*, 6860–6863. doi: 10.1109/IEMBS.2009.5333120.

MacKenzie, I., & Ware, C. (1993). Lag as a determinant of human performance in interactive systems. *In Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, Amsterdam, Netherlands, April 24 - 29, 1993, pp. 488-493.* doi:10.1145/169059.169431.

Nadler, E., Mengert, P., Disario, R., Sussman, E. D., & Grossberg, M. (2009). Effects of satellite- and voice-switching- equipment transmission delays on air traffic control communications. *International Journal of Aviation Psychology. 10/1993;* 3:315-325. DOI: 10.1207/s15327108ijap0304_5.

National Transportation Safety Board. (2011). *SAFETY ALERT: Actual age of NEXRAD data can differ significantly from age indicated on display* (Safety Alert SA-017). Retrieved from http://www.ntsb.gov/safety/safety-alerts/Documents/SA_017.pdf.

Reynolds, H. J. D. (2006). *Modeling the air traffic controller's cognitive projection process*. (Technical Report ICAT-2006-1). Retrieved from DSpace@MIT Website: http://dspace.mit.edu/bitstream/handle/1721.1/34894/Hayley-ICAT-2006-1.pdf?sequence=1.

US Department of Transportation, Federal Aviation Administration. (2013). *Integration of civil unmanned aircraft systems (UAS) in the national airspace system (NAS) roadmap* (Roadmap Report, First Edition – 2013). Retrieved from https://nppa.org/sites/default/files/UAS_Roadmap_2013.pdf.

Yuan, X., Histon, J. M. , Waslander, S., Dizaji, R., & Schneider C. (2012). Distributing non-cooperative surveillance data: A preliminary model and evaluation of potential use cases. *2012 Integrated Communications, Navigation and Surveillance Conference, Washington DC, IEEE, pp. F5–1–F5–10, 2012*, 2155-4943. doi: 10.1109/ICNSurv.2012.6218397.

Yuen, X., Histon, J., Burns, C., & Waslander, S. (2013). Controller-pilot communication in the presence of asynchronous unmanned aircraft system (UAS) radar surveillance data. *International Symposium on Aviation Psychology (ISAP), Dayton, OH, pp. 1–6.* Retrieved from https://uwaterloo.ca/humans-complex-systems-lab/publications/controller-pilot-communications-presence-asynchronous.

Yuan, X., Histon, J. (2014). Distributing non-cooperative object information in next generation radar surveillance systems (Master's thesis, no. HCOM-2014-01, Human in Complex System Lab, University of Waterloo, Waterloo, Ontario, Canada). Retrieved from https://uwaterloo.ca/humans-complex-systems-lab/publications/distributing-non-cooperative-object-information-next.

SOCIAL COMPLEXITY: THE MISSING LINK IN A CRITICAL INCIDENT REPORTING SYSTEM

Jaco van der Westhuizen
Dept. of Human Resource Management
University of Pretoria, South Africa
Karel Stanz
Dept. of Human Resource Management
University of Pretoria, South Africa

The safe operation of complex socio-technical systems depends on the reporting of safety critical incidents by operators within a system. Through the action of reporting, systems develop the capability as learning organizations to improve human and organizational performance. The research paper will provide a social construction understanding of reporter behavior that is influenced by the safety management system and the context of reporting, within an Air Navigation Service Provider (ANSP) in Africa. A case study methodology was applied with complementing inductive coding and thematic content analysis to explore underlying explanations for underreporting behavior. Four main themes emerged: Knowledge Management, Decentralized Safety Power Distance, Shared Safety Logic and Social Construction of Safety/Efficiency. This broad thematic landscape from the four data sets was reflected against the characteristics set by Complexity Theory to produce a framework that guides the systemic approach to reporting that facilitates organizational learning and greater insight into safety risks and opportunities.

"To not be allowed to err is not to be allowed to learn" (Rochlin, 1999, p.1552).

The above statement may be true for all humans, though society expects people in certain industries to perform at an elevated standard that constitutes an error free environment (Bosk, 2000). These industries are typically aviation, petro-chemical, nuclear engineering and health care. The industries where critical incidents resulting from human error are avoided at all costs operate as complex socio-technical systems in high risk environments (or high reliability organisations) based on the realisation that a minor human error can have catastrophic consequences (Rochlin, 1999). This applies to the ANSP that participated in this study. Staender (2011) refers to the learning from critical incidents as experience that leads to expertise and states that incident reporting offers this experience in the form of a window to system weaknesses that become visible. The reporting of critical incidents is therefore an essential part of this learning process and the underreporting of incidents is likely to have a negative effect on both the occurrence of incidents, subsequent organisational learning as well as the performance of the system or insight into system risks.

Critical incidents are events or non-standard situations, with "...origin[s] in the processes, the technique, the environment and the human/team or in any combination of all these factors" (Staender, 2011, p. 209). Research indicates that the mitigation of risk inherent to complex systems in high risk industries remains dependent on the information flowing within such a system (Griffin, Young and Stanton, 2010) and this is dependent on the reporting of critical incidents. Unfortunately, literature also shows significant underreporting of critical incidents in several industries. In a study on the underreporting of maritime accidents, an estimated 50% of accidents were underreported across Canada, Denmark, Norway, Sweden, The Netherlands, The United Kingdom (UK) and The United States (Hassel, Asbjornslett and Hole, 2011). In a 1996 study in the UK health care sector it was found that 39% of hospital service workers had not reported one or more occupational injuries. This high percentage of non-reporting occurred despite the fact that 64% of such unreported injuries required medical treatment and 44% resulted in lost work time (Weddle, 1996). A study by Zellman (1991) found that 40% of health care personnel in the USA admitted to underreporting perceived child abuse.

Literature clearly shows a remarkable and persistent reality of underreporting, although critical incident reporting research was mainly focused on system design, modeling and enablers and barriers to reporting. This study applied Social Construction Theory to critical incident reporting to further the understanding of reporting behavior.

Method

As the research seeks to explain the working of the social phenomenon of reporting in detail an interpretive case study methodology was considered most appropriate. The explorative study focussed on an ANSP in Africa where the target population consisted of operational Air Traffic Controllers (ATCs) and their managerial staff. Firstly, a random sample of six Air Traffic Service Units (ATSUs) was approached for three to seven volunteers to participate in focus groups. Secondly, the same set of questions developed for the focus groups were then applied on seven interviews on Air Traffic Controllers from another five ATSUs that were purposively sampled based upon their experience of a critical incident report that was subjected to an incident investigation. Thirdly, there were seven line managers representing five of the seven ATSUs that participated in the focus groups that volunteered for managerial interviews and fourthly, a further four purposively sampled senior management interviews were conducted that explored reporting from the managerial perspective. Questions elicited observations complemented by storytelling and actual examples of experiences on the actions and inactions of managers, peers and the organisation that influence reporting decision-making. Each of the four data groups were analysed separately with specific attention afforded to differences in the social construction of reporting stemming from exposure to investigations through to investigation expectations and local construction of reporting.

Inductive thematic analysis was performed according to Hycner's (1985) thirteen step approach to gain an overall understanding of the organisation's approach to the reporting of critical incidents, its effects within the organisation, and its perceived degree of success against the backdrop of corporate structures, positions and roles, social interactions and reproduced social structure (Edvardson, Tronvoll, and Gruber, 2011). This was achieved through a classification process of coding, thematic identification and clustering from collection to analysis phase.

Results

The thematic content analysis showed no noteworthy differences in the responses from frontline operator focus groups versus the interviews with reporters exposed to incident investigations. More detailed themes were however prevalent amongst the investigated reporters than the focus groups, for example with themes such as value contribution, multiple realities, trust and self-preservation. Interestingly though, line manager responses had more similarities with investigated reporters, especially on the themes of understanding context of a report or unreported incidents. Other similarities included the realisation of a need for multi-faceted methods to facilitate organisational safety learning and the detrimental implications of a risk measurement tool that they shared with reporters. These managers rather measured reporting success against the amount of verbal queries from their subordinates on reportability of incidents more so than the actual number of written reports filed.

Senior management themes had similarity with line managers and investigated reporters on the part of system optimisation although the context of incidents and especially underreporting was not prevalent. Critical incident reporting at an upper echelon was purely a matter of mandatory reporting guided by a disciplinary code. This linear approach to reporting was echoed in incident reporting as an expression of a corporate safety ratio – a mere number to report on, while on the contrary the other three respondent groups experienced this phenomenon as the cause of an organisational disconnect that dilutes the essence of reporting – improving safety. Decentralised safety governance was a prominent theme that surfaced across all four data sets as a necessity for safety.

Across the four data sets and 1455 codes, the main themes emerged with associated subthemes. Twenty nine supporting themes were identified that are not discussed due to volume constraints. The main themes are:

Table 1.
Overarching Themes across Four Data Sets.

Main Theme	Associated Sub-themes
Decentralized safety management power distance	Value contribution
	Self-preservation at multiple levels
Shared safety logic	Corporate positioning of reporting
Social construction of safety versus efficiency	Context focused
Knowledge management	System wide trust
	Social coherence on reporting

To illustrate the strengths of themes some paraphrases are noted and explained. For example, respondents showed signs of despair through the shrugging of shoulders, frowns or a deep sigh, resulting from the competing demands that are perceived as forcing an operator not just to judge the context but also to evaluate the potential consequences of reporting the critical event. In other words self-preservation emerged from competing consequences. When operators realized that an incident had occurred, their first thought was likely to be about self-preservation, as described by a participant: "The first thing I think about is how it is going to affect me?" This consideration is in addition to the strong emotional impact that incidents have on operators: "A very stressful time, very stressful, it takes a toll on literally everything, family, personal. Emotionally you're up and down the place." This causes underreporting in some instances, not because of personal loss or gain but the measure of safety applied can sway reporting behaviour, 'Some people actually, you know, have an incident without anyone noticing and they wouldn't report it because now it would sort of increase the safety ratio.'

What drives corporate reporting? In this case the researcher was referred to the Disciplinary Code – a stimulus to socially constructing reporting. This stemmed from an industry regulation that makes incident reporting compulsory.

Some of the focuses on reporting and the associated organizational response was depicted as '… because now you feel like your boss is already standing against you and they haven't even started the investigation yet.' '… it looks like they're out to get you, you know.' '… but you must be narrowed down to be a culprit.' On the other hand, the decision to report a critical incident requires judgment on the part of the operator and usually involves specific situations that require the operator to interpret possibly applicable rules with the situation to conclude on a decision to report. One respondent described this process as follows: 'It's a procedural reduction but I mean it's not like … a safety event [critical incident], it's not, it doesn't really affect safety.' Critical information goes astray…

Discussion

Power distance was reflected in the need for self-preservation and consistent with other studies that identified the risk of liability, the burden of reporting (Evans, Berry, Smith, Esterman, Selim, O'Shaughnessy & De Wit, 2006) and self-interest (Blanthorne and Kaplan, 2008) as reasons for underreporting. The word choice of respondents, such as 'culprit,' 'afraid' and 'punishment' reflected Mahajan's (2010) findings that reports were inhibited by views that only 'bad' professionals make mistakes. The Blanthorne *et al*. (2008) findings that the value and purpose of a report was not the only driver of reporting, but that individual consequences were a deciding factor, holds true for this study. The study also found that team consequences were a deciding factor and moreover that managerial levels also responded to incident reporting in a self-preservation fashion, despite evidence of ownership towards safety across organisational levels.

The professionals who participated in the research considered the context of an incident to play a substantial role in judging whether an incident was reportable. This is supported by findings from Wagner, Harkness and Gallagher (2012); however the themes disclosed that reporting comes to its own right in a safety management system when the focus is on value contribution beyond judging context for reportability. In the same light Tourtier, Auroy and Grasser (2010) has cautioned against oversimplifying reported incidents masking hidden risks. Therefore, shared safety logic is required with an appropriate amount of organisational energy afforded to reporting. The level of energy afforded will depend on the level of tension between the requirement for incident reporting and the performance measure of reducing incidents that presents a contradictory dilemma. Although accountability is needed and assumed by participating respondents, the tension is aggravated by misaligned social constructions of efficiency and safety when not seen as two polarities on the same continuum (Dekker, 2007).

Knowledge management closes the thematic loop as experiences of blaming and reprimand, as discovered by Firth-Cozens, Redfern and Moss (2004) in the medical industries inhibited reporting, while the theme of relationship management shaped by trust surfaced as key to reporting. Social coherence on reporting changes the view on what to report but also how to report as it links again with the value contribution of a report as oppose to human error. The understanding of themes from an organisational behaviour perspective was found to be better understood from a Complexity Theory stance.

A Complexity Framework

The main themes were mapped against the complexity characteristics set by Cilliers (1998) and a framework was derived that guides the approach and understanding towards reporting behaviour as a social construct within a complex socio-technical environment beyond the formal design of a reporting system in Figure 1.

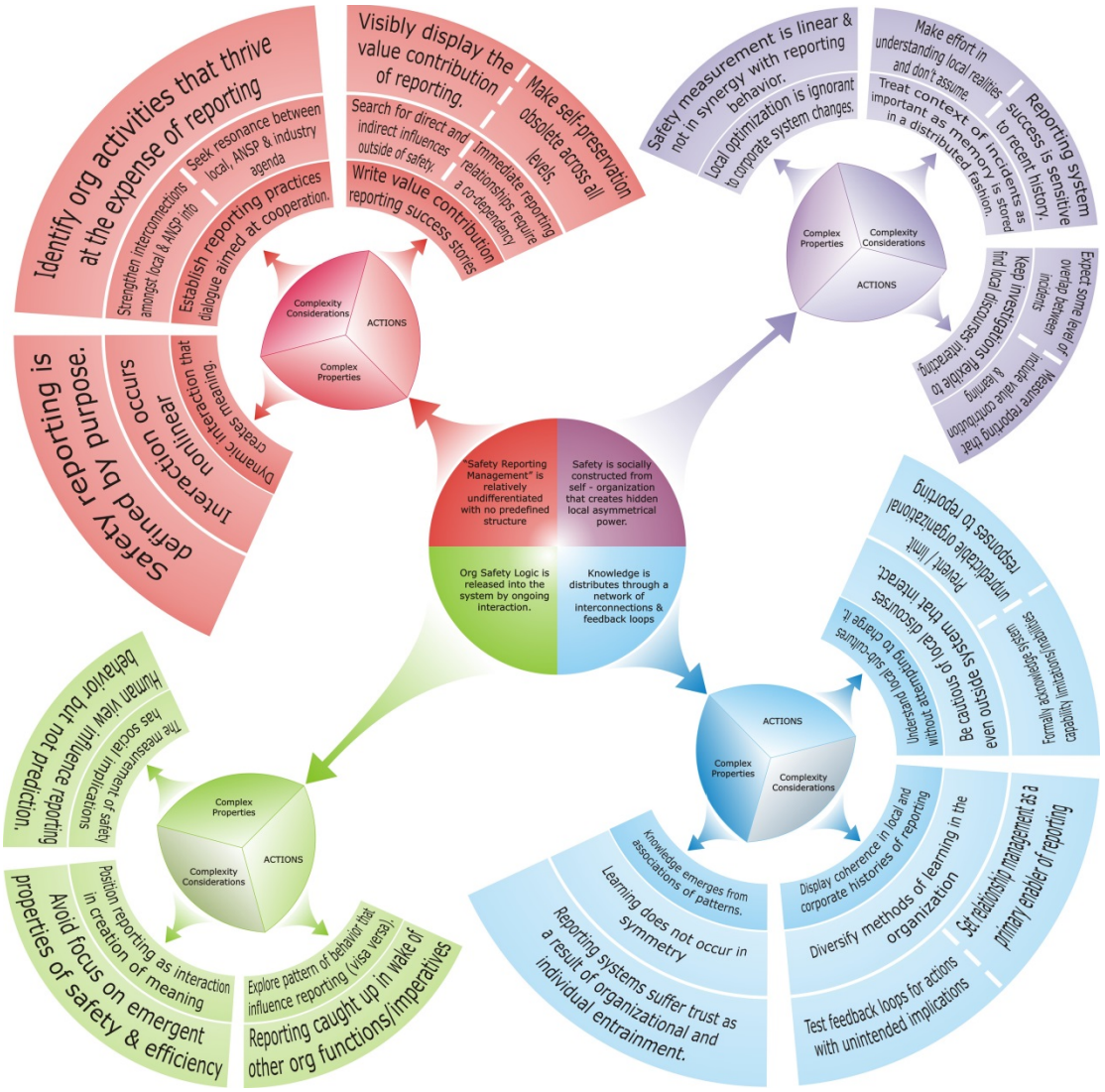


Figure 1. *A Complexity Framework to Map Reporting Markers*

Decentralized Safety Management Power Distance (value contribution & self-preservation at multiple levels)

The framework brings to the fore a fulcrum of organisational behaviour that gently directs the flow of organisational energy pertaining to reporting. For instance, embracing the asymmetrical power within the system and allowing it to shift as the relativity of safety actions shift and framing a co-dependency as the reporter to be as accountable and responsible to safety as the senior level manager embraces the reality of complex reporting behaviour. This embracement evolves from dynamic dialogue across hierarchical lines, a system's approach that seek individual and organisational gain from reporting as well as defining reporting by its value contribution. The reporting landscape also has to be scanned for power distance risk markers that can distress reporting behaviour, for example, a lack of resonance between corporate and local safety aims, organisational activities such as a disciplinary practice/code that exist at the expense of reporting (entrainment). A final prominent risk marker is that self-

preservation across levels within an organisation because a tolerance for self-preservation only exists when the focus distracts away from safety (red sector).

Shared Safety Logic (corporate positioning of reporting)

As a safety system is dependent on the information flowing internally, an Organisational Safety Logic is required that can only be affected from thorough interaction amongst stakeholders that creates meaning. This of course means that reporting has to be correctly positioned in the system as a safety pillar and a basic activity while not becoming a mere capturing tool of human fallibility or diluted by perceived more important system activities. The organisational safety logic (that reporting forms part of) can only be made prominent through patterns made explicit by continuous coherent communication across levels and units that is unique but that overlaps (green sector).

Social Construction of Safety versus Efficiency (context focused)

Safety and efficiency is socially constructed at a local level in an ANSP and this also fuels self-organisation and an automatic emergence of asymmetrical power upwards – the power of choosing not to report. This is possible because local performance is ignorant to system wide behaviour. To explain this better, complexity theory indicates that reporting behaviour is non-linear and fluid and therefore not bound by any corporate rules while being sensitive to recent history – how the last reporter was treated influences my behaviour. In the same light a reporting memory (knowledge) is stored in a distributed fashion which makes the context of each incident important to the extent that flexibility is required when investigations are performed for safer systems. Safer systems expects measurement, however reporting in itself was not measured, which in turn creates the construction of being less valued and where measured, it was limited to volumes of reporting instead of what is important to reporters and the system, e.g. learning and value contribution (purple sector).

Knowledge Management (System wide trust & social coherence on reporting)

Once a report has been tendered, knowledge management lies at the heart of that act. The themes of the data illustrates that the complexity of knowledge has to be applied to improve knowledge management. In other words, knowledge distribution occurs through a network of interconnections with multiple recursive feedback loops and learning therefore does not occur in symmetry but rather by association of patterns. The most important of these patterns are relationship patterns as reporting requires an appropriate level of intimacy amongst the transmitters and receivers. This suggests that an ANSP should be cautious of local discourses' interaction that may distort the predictability that reporters expect from the system while also demanding honesty towards the limitations of the system if quality reporting is expected corporately but the processing capability contains inabilities (blue sector).

The limitations of the study include the small sample size as well as the fact that the sample was restrained to a single organisation within the African aviation industry. Moreover, the study was performed by an ex-air traffic controller that introduces a particular frame of reference. Further research is proposed in complex high risk socio-technical organisations to test the proposed framework for applicability. It may also be valuable to explore how cultural differences may influence the social construction of reporting across different demographic group, although no noticeable differences surfaced during this study within a multi-culture environment. It is foreseen that the framework can even possibly be adapted into a scoring mechanism to assess safety critical reporting practices across industries.

Conclusion

The application of the complexity characteristics on the social construction of reporting themes from the study, point towards fluidity in a reporting system despite the assumption that a safety service is highly regulated. However, high risk industries or so called high reliability organisations can benefit exponentially from an understanding of the social construction of reporting as an integral part of the design and operation of a critical incident reporting system. This though requires a decentralised safety governance (not structure) approach and a knowledge management philosophy that embrace value contribution that makes self-preservation across levels obsolete. Inevitably such core practices improve organisational learning from critical incidents and performance insights of systems that revolve around safety - a journey that can be facilitated by the Complexity Framework.

Acknowledgements

The paper is based on a doctoral dissertation conducted by a Human Factor Specialist in the employment of the ANSP. The views of the research reported do not reflect the views of the ANSP necessarily.

References

Blanthorne, C. & Kaplan, S. (2008). An egocentric model of the relations among the opportunity to underreport, social norms, ethical beliefs, and underreporting behaviour. *Accounting, Organisations and Society*, 33, 684-703.

Bosk, C. (1979). *Forgive and remember: managing medical failure*. Chicago: Chicago University Press.

Cilliers, P. (1998). *Complexity theory and postmodernism*. Oxon: Routledge.

Dekker, S. (2007). *Just Culture: balancing safety and accountability*. Hampshire, UK: Ashgate Publishing Company.

Edvardsson, B., Tronvoll, B. & Gruber, T. (2011). Expanding understanding of service exchange and value co-creation: a social construction approach. *Journal of the Academic Marketing Science*, 39, 327-339.

Evans, S. M., Berry, J. G., Smith, B. J., Esterman, A., Selim, P., O'Shaughnessy, J. & De Wit, M. (2006). Attitudes and barriers to incident reporting: a collaborative hospital study. *Quality, Safety and Health Care*, 15, 39-43.

Firth-Cozens, J., Redfern, N. & Moss, F. (2004). Confronting errors in patient care: the experience of doctors and nurses. *Clinical Risk*, 10:184-190.

Griffin, T. G. C., Young, M. S. & Stanton, N. A. (2010). Investigating accident causation through information network modelling. *Ergonomics*, 53(2), 198-210.

Hassel, M., Asbjornslett, B. E. & Hole, L. P. (2011). Underreporting of maritime accidents to vessel accident databases. *Accident Analysis and Prevention*, 43, 2053-2063.

Hycner, R. H. (1985). Some guidelines for the phenomenological analysis of interview data. *Human Studies*, 8, 279-303.

Mahajan, R. P. (2010). Critical incident reporting and learning. *British Journal of Anaesthesia*, 105(1), 69-75.

Rochlin, G. I. (1999). Safe operation as a social construct. *Ergonomics*, 42(11), 1549-1560.

Staender, S. (2011). Incident reporting in anaesthesiology. *Best Practice and Research Clinical Anaesthesiology*, 25, 207-214.

Tourtier, J., Auroy, Y. & Grasser, L. (2012). On violation management: lessons from aviation. *International Journal for Care of the Injured*, 43, 386-393.

Wagner, L. M., Harkness, K. & Gallagher, T. H. (2012). Nurses' perceptions of error reporting and disclosure in nursing homes. *Journal of Nursing Care Quality*, 27(1), 63-69.

Weddle, G. M. (1996). Reporting occupational injuries: the first step. *Journal of Safety Research*, 27(4), 217-223.

Zellman, G.L. (1991). Report decision-making patterns among mandated child abuse reporters. *Child Abuse and Neglect*, 14, 325-333.

PILOTS' WILLINGNESS TO REPORT AVIATION INCIDENTS

Andreas Haslbeck
Institute of Ergonomics, Technische Universität München
Munich, Germany
Carsten Schmidt-Moll
Commercial Airline Captain
Germany
Ekkehart Schubert
Institute of Aeronautics and Astronautics, Technische Universität Berlin
Berlin, Germany

This paper reports results from a survey-based study among eighty-six airline pilots investigating their willingness to report safety-relevant events and incidents. Pilots have been asked to report how many events they have experienced in thirty-five different contextual areas and how often they have reported these cases. Thus, underreporting rates, respectively dark figures, were calculated and listed. These results and the willingness to report are discussed within an aviation operation's background. Most of these surveyed underreporting rates are very high, which means a substantial source of uncertainty in airlines' safety reporting databases, and thus for airlines' safety management systems.

In safety-critical domains like the nuclear, intensive healthcare or aviation industry, reporting systems play an important role in safety management: as *to err is human*, error reports are crucial to learning from errors and avoiding future errors. A well accepted reporting culture supports this process. It is often unclear, however, whether the willingness to report observed or self-committed errors is distinctive or not (Tani, 2010). Underreporting, or dark figures, means that safety-critical events occur and are observed by operators, however, are not being reported to superiors, safety departments or authorities. These dark figures are frequently rather high and unknown due to the fear of being blamed for committed errors in professional life. Heinrich (1931) carried out studies about occupational safety and health. He proposed that in a workplace, for every accident that causes a major injury, there are 29 accidents with minor injuries and 300 accidents with no injuries - named as incidents or *near misses* (1-29-300). The detailed analysis of such a large number of incidents could prevent accidents because they share common root causes. With a further investigation of more than 1.5 million reported events reported by 297 cooperating companies, Bird and Germain (1985) determined that the number of *near misses* is even greater than once thought. They revised the ratio of one accident (with a serious or major injury) to ten occurrences with minor injuries to 30 property damage incidents up to six hundred *near-miss* incidents with no visible damage (1-10-30-600). Illustrated as the *iceberg model*, the works of Heinrich, Bird, and Germain expose that accidents, serious incidents or property damage only represent the small visible part above the water surface. An occurrence reporting system allows companies to gather larger quantities of incident information by individuals. Detailed analysis of the data allows a look beneath the water surface of the *iceberg model.* Related safety precautions derived from this data could prevent accidents, but only if the incidents are reliably and truthfully reported by the individuals.

There is much literature about incident reporting in general; however, precise domain-specific under-reporting rates, respectively dark figures, are rarely published. Factors influencing incident reporting, like a manager's reactions to reports, were found in literature and discussed by Clarke (1998). She also conducted an experiment among train drivers to find out reasons for not reporting. Her main finding indicates that the managers' reactions to reports play an important role on operators' willingness to report. Pransky, Snyder, Dembe, and Himmelstein (1999) have reviewed literature about underreporting of work-related disorders and collected data about unreported wrist/hand symptoms: 53% of these have not been reported. Reasons for underreporting were the fear of consequences on the job and disciplinary actions, or only minor symptoms. Four factors influencing underreporting were derived from literature by van der Schaaf and Kanse (2004): the fear for disciplinary action, an attitude of risk acceptance, a feeling of uselessness of reporting, and several practical reasons. Tani (2010) listed reporting schemes and systems in aviation from several countries in her thesis. Her research questions addressed different aspects of the use of these tools, and in the empirical part of her work she figured out six factors affecting the willingness to report: seriousness of errors, direct or indirect involvement in errors, working environment, legal protection of the reporter, motive of the wrongdoer, and relationship to the wrongdoer (Tani, 2010). Knowing these factors can help an organization to establish a reporting system or to improve the willingness to report. One

important issue about reporting in general is postulated by Strauss (2002): reports are always processed and evaluated by technical staff with shifting focus. Some reports were written and submitted, but were sorted out and so this number biases underreporting. When looking for precise numbers of underreporting, most sources deliver crime or medical data; only few works have been published in transportation. Jayasuriya and Anandaciva (1995) have conducted an experiment to evaluate compliance in reporting systems in anesthesia and found a dark figure of 70%. Barach and Small (2000) mentioned dark figures in medicine between 50% and 96% for adverse events in the U.S. They also deliver a descriptive list of non-medical incident reporting systems and a list of barriers and incentives for reporting. For underreporting of aviation wirestrikes in Australia, a dark figure of 40% can be found (ATSB, 2012). An estimated dark figure of 20% for birdstrike reporting in the U.S. is given by Cleary, Dolbeer, and Wright (2000). For maritime tanker operations, Psarros, Skjong, and Eide (2010), deliver dark figures of at least 59% and 70%, depending on the data source. To our knowledge there is no publication dealing with situation- or task-specific underreporting in aviation. Literature and experience from practice offer a differentiated insight into an aviation industry, which is highly cost driven, strongly influenced by international competition and focused on a constant increase of productivity, sometimes neglecting safety.

Method

Research Question

The research question of this paper is about the underreporting rate of incidents from airline pilots during their daily flight operation: how many events and incidents do pilots report to their safety department in relation to the total amounts of experienced incidents? In addition a subsequent research question is whether there are specific types of events and incidents showing higher underreporting than others. The research question cannot be answered with the help of any other database. German aviation accidents and severe incidents are defined by law (FlUUG of 1998, 1998); they have to be reported by national regulation (LuftVO of 1963, § 5, 2012) to the German Federal Bureau of Aircraft Accident Investigation (BFU) and by European Commission Regulation (EU) No 965/2012 via the German Federal Aviation Office (LBA) into a database. Even though an underreporting is assumed not to appear, any official database is just the sum of the individual databases of the different airlines.

Questionnaire

The focus of this survey was to ask pilots about the number of events and incidents they have experienced (*m*) in the last twelve months prior to the experiment and, in a second question, the number of events they have reported (*n*). The under-reporting rate (Probst & Estrada, 2010), respectively the dark figure, is then calculated as follows: *dark figure = (1- n/m)*. To specify the severity of experienced events, four different categories were defined (Table 1). There was not only one question for such events, but 35 questions divided into three different thematic areas: ground operations, in-flight operations, and human factors. All questions can be seen in Table 2.

Table 1.

Four succeeding severity categories for experienced events and incidents.

Category	Description: A safety-relevant event …
A	whose consequences were completely covered by the crew (‘*problem was solved’*)
B	whose consequences were only partly covered by the crew (‘*that was close’*)
C	whose consequences were not covered by the crew (‘*by a whisker’*)
D	where the situation was completely out of control *(‘oh my gosh’*)

Participants

The participants of this survey were scheduled for a non-voluntary full-flight simulator research experiment (Haslbeck, Gontar, & Schubert, 2014). They work as pilots for the same airline, either on an Airbus A320 short-haul or an Airbus A340 long-haul fleet. From altogether one-hundred-twenty pilots, eighty-six pilots participated in this survey on an anonymous basis during their waiting time for the experiment. Thirty-four participants can be assigned

to the group of A320s (15 CPT, 19 FOs) and thirty-five can be assigned to the group of A340s (16 CPTs, 19 FOs). With six female participants (five FOs and one CPT), the group very well represents partner airline proportions at approximately five percent.

Results

Data Analysis

To calculate a dark figure, all events of one contextual-category are added and compared to the number of reports afterwards according to the abovementioned formula. When analyzing the obtained data from pilots, the values given in category A show a remarkably higher variability compared to the other three categories. In theory, someone might expect very high numbers in category A and, the more severe the categories become, constantly falling numbers, like postulated by Heinrich (1931). However, there are at least two different types of entries in category A: some pilots have mentioned up to 1,000 events for deviations from standard operating procedures (SOP), while other pilots mentioned no such events in the same category at all. From their personal perspective, both might be right: one pilot calculates 20 flights per month (on short-haul operation), 10 working months per year, and roughly estimates 5 SOP deviations per flight – another pilot notices no SOP deviations at all. These differing perceptions might be the cause of a lower reliability of data in category A. Another question considering this category must be discussed: even when these events need to be reported in theory, it can be disbelieved whether such category A occurrences will be reported and, in addition, whether these occurrences will be added to the database, as Strauss (2002) postulated his doubts. Because of these limitations concerning category A, two different methods to calculate the dark figures were applied. In one calculation, representing a lower boundary, only the numbers of categories B, C, and D were added (method 1); in another approach, representing an upper boundary, all four categories have been taken into account (method 2). The data analysis also allows the identification of the highest risks within the flight operation. These incidents with frequent D-categorized occurrences are marked in bold letters. The calculated dark figures for safety-relevant events as described in the above paragraph are shown in Table 2.

Table 2.

Dark figures for safety-relevant events as reported by 86 commercial airline pilots.

	Method 1 (B+C+D)		Method 2 (A+B+C+D)	
Description	number of entries	calculated dark figure	number of entries	calculated dark figure
deviations from SOPs (procedures, callouts, wording, etc.)	140	.971	2707	.999
complacency (airborne use of cell phone/laptop, reading, programming of FMA too late, distraction below FL100)	142	.993	2664	1.0
time pressure induced by organizational deficits (ground handling, turn-around, dispatch, technical condition of aircraft, etc.)	**376**	**.939**	**1577**	**.985**
errors due to external factors on ground (time pressure, difficulties due to operational reasons, e.g. the technical status of the aircraft.)	108	.852	840	.981
reduced capability (fatigue, illness, stress)	**338**	**.962**	**687**	**.981**
ATC-induced time pressure (start-up, slot, parking position, push-back)	82	.963	610	.995
weather induced problems airborne (fog, visibility, turbulence, thunderstorm, icing, wind, lightning, etc.)	**70**	**.814**	**513**	**.975**
problems with paper-based documentation (flight plan, fuel calculation, NOTAMS, weather, technical, HIL)	48	.854	489	.986
neg. operational factors at the airport (ATC language problems, airport closures, congestion, FODs, birds, etc.)	95	.821	457	.963
clear and organized arrangement of the airport (taxiways, signs, etc.)	102	.961	397	.99

weather induced problems on ground (contamination, wind, gusts, fog, rain, RVR or visibility)	44	.909	303	.987
problems with electronic documentation (LIDO, EFB, eRoute Manual, performance calculation, OM A - C)	42	.738	291	.962
autoflight system (A/P and F/D-modes, ILS-capture, lat. and vert. control, manual handling, mis-selection, etc.)	30	.933	288	.993
incorrect inputs on the FMS (route, flight plan, arrival, departure)	25	.8	184	.973
incorrect take-off data or performance calculation	25	1.0	156	1.0
birdstrike	36	N/A	123	.455
smell in cabin	9	N/A	127	.858
unruly passengers	24	.167	108	.815
unstabilized approach (no go-around performed)	32	.969	103	.99
low fuel due to OPS, economic considerations, weather, etc.	26	.846	76	.947
smell on flight deck	15	.4	70	.871
unintended deviation from flight path on heading or altitude	18	.944	45	.978
lack of flight training, lack of manual practice, manual flying	20	.95	42	.976
loss of control (unusual attitudes, early flap retraction, overspeed, turbulence, IR or IAS disagreements, etc.)	6	.5	37	.919
incorrect weight and balance loadsheet	6	.833	36	.972
loss of separation with other aircraft	12	.25	34	.735
insufficient / inaccurate de-icing	9	.556	24	.833
near-miss on ground (with other aircraft, ground vehicle, people, lamp pole, ground power, etc.)	**9**	**.333**	**18**	**.667**
hard landing	4	.000	17	.765
flight crew incapacitation	3	1.0	10	1.0
visible smoke on flight deck or in cabin	3	N/A	5	.2
taxiway and apron excursion	2	.5	4	.75
landing with fuel for less than 30 mins.	**2**	**1.0**	**2**	**1.0**
runway incursion	1	.000	1	.000
RWY excursion at the end or on the side	1	.000	1	.000

Discussion

Reasons for Underreporting

Pilots were asked if their reporting had changed within the last ten years. Exactly 50% of the pilots said their amount of written air safety reports did not change, 47% stated they write less reports, and only 3% write more reports compared to 10 years age. The reasons for not writing air safety reports differ significantly between short and long haul pilots. For short haul pilots, it is the complexity of writing a report with a complicated and time-consuming database software. Thereafter, it is the negative feedback by their superior and third, the lack to initiate change. On the contrary, for long haul pilots, it is first the negative feedback by their superiors and then the lack to initiate change. The third reason for not writing is the generally felt meaninglessness and insignificance of any air safety report. Moreover, the events above can be clustered into different contextual accident and incident categories: organization (ORG), environmental (ENV) threats, technical failures (TEC), and human (HUM) performance issues.

Organizational Challenges

With this survey, it also became possible to identify the five most severe events (D category), representing the highest risks. The first two are *time pressure induced by organizational deficits* (ground handling, turn-around, dispatch, technical condition, etc.) and *reduced capability* (fatigue, illness, stress). Both event types may be mitigated by measures and changes, taken by one's own organization like ground handling or crew scheduling. Competition in business and permanent cost pressure continuously lead to longer duty times and an increased workload with less time for recreation. Furthermore, every airline aims to shorten the time span for airplanes standing on ground. And every airline is interested in minimizing the amount of personnel involved. For the future, it is a challenge for company leadership to keep the balance between continuously improving the existing processes and at the same time, minimizing organizational deficiencies. Frequent irregularities and operational difficulties like delays or technical problems of the airplane have to be considered. Understandably, the deviations from the standard operation procedures lead to a higher workload and increased time pressure, which again lead to an increase in errors and incidents. The results illustrate that operational difficulties put significant stress on pilots who seem to be unprepared and untrained for handling these organizational difficulties. In support of the international accident statistic, todays training of flight crews is mainly focused on flight safety regarding the handling of abnormal situations in-flight. In the future, crew training should be expanded to practice organizational difficulties on ground. This approach intends to help crews to cope with these changes and lead to more efficient organizational circumstances during ground operation.

Other high-risk D-categorized events imply *weather induced problems in the air*, being *low on fuel* and *near-miss on ground* (with other aircraft, ground vehicle, people, lamp pole, ground power, etc.). It may be assumed that there is a correlation between *bad weather* and *being low on fuel*. Ending up with minimum fuel during bad weather could be a planning deficiency by the crew, but could also be the consequence of economic pressure within the airline. It could well be the organization trying to make the crew take less fuel than practicable. The management of an airline is able to counteract both risks. Offering better weather information in combination with additional fuel on board could be a simple answer to the problem. It might even be reasonable for an airline to dictate minimum fuel values for certain weather phenomena, giving flight crews more time to deal with hazardous weather situations during flight.

It may be assumed that underreporting is more severe among low cost airlines. They frequently employ pilots via employment agencies (Bachman & Matlack, 2015). Staff members become *self-employed-pilots* which imply weaker pilots' labor rights. Their fixed-term contracts do not include permanent positions. A pilot's safety report criticizing organizational deficiencies including the superiors would not only be unpopular, but could also have consequences for the continuation of a pilot's labor contract.

Safety Management

According to ICAO Doc 9859 (ICAO, 2009), there are many approaches to risk assessment. Each time when evaluating a risk, both the probability of occurrence of a hazardous effect and the severity potential of that effect need to be taken into account. It is a common practice to use a risk classification matrix in support of this two-dimensional judgment. To know the probability, safety management does need information about the actual number of incidents within their operation. The underreporting shown in this survey is too extensive to give a reliable base for such an approach like the risk matrix. The survey pointed out substantial underreporting among human factor issues like SOP deviations or fatigue. Risks due to fatigue could, therefore, be underestimated, whereas technical risks with lower underreporting rates, like mid-air-collisions, could be overestimated. Prospective safety work should try to estimate the amount of occurrences. Regular and anonymous surveys among employees of airlines could be an additional possible and suitable approach for the future.

Air transportation will continue its growth. To keep today's accident rates low, it is necessary to further improve flight safety. For staying successful, airline management must continue optimizing its organizational structures, and they must simultaneously utilize all available safety information to avoid and reduce risks. This balance will be trendsetting and a key item for the long-term success of an airline. This survey has shown that existing reporting systems alone are no longer able to gather important safety data. In today's airline industry, additional means to collect data must be implemented.

Acknowledgements

This work was funded by the German Federal Ministry of Economics and Technology via the Project Management Agency for Aeronautics Research within the Federal Aeronautical Research Program (LuFo IV-2). The authors acknowledge their thanks to Sören Merzky and David Schopf for supporting this study and their contribution to the project.

References

ATSB. (2012). Under reporting of aviation wirestrikes: Aviation Research Report (No. AR-2011-004). Canberra.

Bachman, J., & Matlack, C. (2015). Budget Airlines Shop the World for Cheaper Pilots: Ryanair and Norwegian Air Shuttle are turning to the cockpit for ways to trim costs: BloombergBusiness. Retrieved from http://www.bloomberg.com/news/articles/2015-02-12/budget-airlines-shop-the-world-for-cheaper-pilots

Barach, P., & Small, S. D. (2000). Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. The BMJ, 320(7237), 759–763. doi:10.1136/bmj.320.7237.759

Bird, F. E., & Germain, G. L. (1986). Practical loss control leadership. Loganville: Institute Pub.

Luftverkehrs-Ordnung (LuftVO), Bundesministerium für Verkehr 1963.

Gesetz über die Untersuchung von Unfällen und Störungen bei dem Betrieb ziviler Luftfahrzeuge (Flugunfall-Untersuchungs-Gesetz - FlUUG) Bundesgesetzblatt, Bundesrepublik Deutschland 1998.

Clarke, S. (1998). Organizational factors affecting the incident reporting of train drivers. Work & Stress, 12(1), 6–16. doi:10.1080/02678379808256845

Cleary, E. C., Wright, S. E., & Dolbeer, R. A. (2000). Wildlife strikes to civilian aircraft in the United States 1990-1999. Washington, DC.

Commission Regulation (EU) No 965/2012 Official Journal of the European Union, European Commission 2012.

Haslbeck, A., Gontar, P., & Schubert, E. (2014). How Can Procedures and Checklists Help Pilots in Abnormal Flight Situations? In N. A. Stanton, S. J. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), Advances in Human Aspects of Transportation. Part II (pp. 456–461). AHFE International.

Heinrich, H. W. (1931). Industrial accident prevention: a scientific approach. New York: McGraw-Hill.

International Civil Aviation Organization. (2009). Safety Management Manual (No. Doc 9859). Montréal. Retrieved from www.icao.int

Jayasuriya, J. P., & Anandaciva, S. (1995). Compliance with an incident report scheme in anaesthesia. Anaesthesia, 50(10), 846–849.

Pransky, G., Snyder, T., Dembe, A., & Himmelstein, J. (1999). Under-reporting of work-related disorders in the workplace: a case study and review of the literature. Ergonomics, 42(1), 171–182. doi:10.1080/001401399185874

Probst, T. M., & Estrada, A. X. (2010). Accident under-reporting among employees: testing the moderating influence of psychological safety climate and supervisor enforcement of safety practices. Accident Analysis and Prevention, 42(5), 1438–1444. doi:10.1016/j.aap.2009.06.027

Psarros, G., Skjong, R., & Eide, M. S. (2010). Under-reporting of maritime accidents. Accident Analysis & Prevention, 42(2), 619–625. doi:10.1016/j.aap.2009.10.008

Strauss, B. (2002). Avionics interference from portable electronic devices: review of the Aviation Safety Reporting System database (Digital Avionics Systems Conference).

Tani, K. (2010). Under-reporting in aviation: An investigation of factors that affect reporting of safety concerns (Dissertation). Massey University, Manawatu. Retrieved from http://mro.massey.ac.nz/handle/10179/1998

van der Schaaf, T., & Kanse, L. (2004). Biases in incident reporting databases: an empirical study in the chemical process industry. Safety Science, 42(1), 57–67. doi:10.1016/S0925-7535(03)00023-7

EFFECTS OF WORKLOAD ON MEASURES OF SUSTAINED ATTENTION DURING A FLIGHT SIMULATOR NIGHT MISSION

Hans-Juergen Hoermann
Institute of Aerospace Medicine, Department of Aviation and Space Psychology
German Aerospace Center (DLR), Hamburg, Germany

Patrick Gontar and Andreas Haslbeck
Institute of Ergonomics, Technische Universität München
Munich, Germany

N=60 commercial airline pilots holding valid ATPLs flew a manual ILS approach following a weather induced missed approach during a night mission in full flight simulators. Measures of subjective fatigue, sustained attention, and the NASA Taskload Index were collected before and after the mission. In addition, sleep history data were available covering three days prior to the simulator. Both subjective and objective measures of fatigue showed significant ascent over the three hours of the experimental procedure. While sleep history data and roster information were related to both the overall level of fatigue and to reaction times, pilots who experienced a higher degree of workload during the simulator exercise showed a significant increase in subjective fatigue scores after the mission. The findings provide some evidence for lasting effects of a sleep deficit as well as for a multifactorial model of fatigue risk.

Most models of fatigue risk in aviation can be traced back to the classical two-process model of sleep regulation (Borbély, 1982), which explains sleepiness through the interaction of homeostatic (sleep pressure by time awake) and circadian influences (circadian phase). In order to achieve more accurate predictions, some fatigue risk management systems (FRMS) consider additionally sleep inertia (Åkerstedt & Folkard, 1990), task-related factors (i.e. time-zone transitions, workload, work-schedule), individual factors (i.e. life-style, chronotype) or cumulative effects (VanDongen et al., 2003). However, empirical validation data for task-related and individual factors with cognitive effectiveness within the aviation environment are rare and contradictory (Tritschler & Bond, 2010; Williamson et al., 2011).

This study aimed to explore the relationship between subjective and objective measures of fatigue with factors of workload and work scheduling. Our data were gathered before and after a simulator night mission with a sample of long- and short-haul pilots who had been awake for more than 16 hours. It was expected that individual sleep history and scheduling factors are equally related to the overall level of fatigue before and after the simulator mission. In addition to that, we analyzed whether workload as experienced by the individual pilot during the simulator mission can be identified as a moderator variable for an increase of fatigue after the mission.

Method

Experimental Procedure

This study originally aimed for measuring manual flying skills of airline pilots in a full-flight simulator (*JAR STD 1A Level D*) night mission. However, the data of flying performance itself are reported elsewhere (Haslbeck, Kirchner, Schubert, & Bengler, 2014). All participants were asked to get up as usual in the morning and not to sleep during the daytime prior to the simulator session. The procedure started at 9:30 p.m. with dinner at a restaurant located close to the simulator facility. Three pilots per night participated in the overall 20 simulator missions. Between 11 p.m. and midnight, the baseline measurement of subjective fatigue and sustained attention took place (*base*). Thereafter, the whole group of three pilots went to the simulator and a second *pre*-simulator measurement of fatigue scores was conducted for the second and the third participant (*pre*) while the first participant started the 45-minute experiment at about 12:30 a.m. After another 15 minutes, during which the simulator was reset, the next participant started the experiment at about 1:30 a.m; the third participant started at about 2:30 a.m. For all participants, a final fatigue testing was scheduled in a briefing room immediately after they finished their simulator trial between 1:30 a.m. and 3:30 a.m. (*post*). During this *post*-session participants also assessed the level of workload during the simulator mission. In this scenario, all pilots had to perform an approach scenario towards Munich (EDDM) ending with a missed approach decision due to a strong tail wind situation (Haslbeck, Eichinger, & Bengler, 2013). After the crew performed their go-around, the tower changed the runway direction. When the pilots

turned to their final runway heading, we evoked a malfunction leading to a failure of the autopilot and the flight director. Consequently, the participant had to manually fly and land the aircraft (raw data ILS).

Participants

All 60 pilots participating in this study were scheduled for a full-flight simulator research experiment; participation was part of their working schedule in terms of an additional simulator event and not discretionary. That meant they were randomly selected from the crew planning department, and participation in the experiment was part of their normal duty time (Haslbeck et al., 2012). 30 long-haul captains (Airbus A340) and 30 short-haul first officers (Airbus A320), representing a wide range of experience considering the level of practice and training (listed in Table 1.) of the co-operating partner airline participated in the study. However, only data from N=57 pilots can be included in the following statistics because one simulator mission was cancelled due to technical reasons.

Table 1.

Demographical data of all participants, adapted from Haslbeck et al. (2014)

	Age		overall flight hours	
	mean	SD	mean	SD
CPTs A340 (n=27)	50.4	4.0	15,019.7	2,938.4
FOs A320 (n=30)	30.4	3.0	3,373.9	1,703.9

Measures

A number of objective and subjective measurements were administered according to the procedure described above in order to collect data on fatigue, level of attention, workload and the three-day sleep history.

NASA Raw TLX (RTLX): A German version of the NASA Task Load Index (Hart & Staveland, 1988; Hart, 2006; Pfendler, Pitrella & Wiegand, 1994) was used as a measure of subjective workload during the simulator mission. With respect to the overall demanding procedure, the RTLX was chosen which omits the comparison of the subscales. This method delivered a total score (RTLX) and six subscales reflecting three different workload factors:

- Task related:
  - *Mental demands* (TL-MD)
  - *Physical demands* (TL-PD)
  - *Temporal demands* (TL-TD)
- Behavior related:
  - *Performance* (TL-PE)
  - *Effort* (TL-EF)
- Subject related:
  - *Frustration* (TL-FR)

All RTLX scores were scaled between 0 and 100.

Psychomotor Vigilance Task (PVT): As a measure for sustained attention, a ten-minute version of the PVT (Dinges & Powell, 1985; PEBL version by Mueller, 2011) was administered on portable computers. A simple visual stimulus was presented about 7 times per minute with variable inter-stimulus intervals. Subjects were asked to press the space key as soon as they see the stimulus. Different performance scores were calculated:

- Number of lapses with reaction times > 500ms (P-LAPS)
- Mean reaction times for the 10% slowest responses (P-RT10)
- Mean reaction times for the 10% fastest responses (P-RT90)
- Overall mean reaction times (P-MRT)

Subjective Fatigue Checklist (FAT): With the FAT (Samn & Perelli, 1982) the subjective level of fatigue was assessed subsequent to the PVT. The FAT provided subjective fatigue scores between 0 (lowest) and 20 (highest). Scores above 9 are regarded as “mild fatigue”, above 12 as “moderate fatigue”, and above 16 as “severe fatigue” (Samn & Perelli, 1982, p5).

Visual Analogue Scale (VAS): A visual analogue scale (VAS) was used for a subjective alertness/sleepiness assessment. A score of 100 was labeled as “very alert” and 0 as “very sleepy”.

Sleep diaries and roster information: Subjects started writing sleep diaries three nights before their scheduled simulator mission. The recorded parameters reflected

- Sleeping time (SD-ST)
- Wakeup time (SD-WU)
- Sleep quality (SD-QU – analogue scale from 0 – "very bad" to 100 - "very good")

From this information we calculated two further scores:

- Accumulated sleep deficit (SD-DEF): 24 hours minus the time asleep during the previous three nights
- Time awake before the simulator mission in minutes (SD-TAW)

With respect to roster information we considered here two parameters:

- Number of duty days within three days before the simulator event (SD-DUT)
- Last flight over 3 or more time zones within 3 days before the simulator (SD-TZN) (Samel et al., 1995)

Reliable Change Index (RCI): In order to receive a measure of change for the fatigue scores, the Reliable Change Index (Jacobson & Truax, 1991) was calculated between the *base*-scores and the *post*-scores. The RCIs compensate change information for unreliability of measurement.

Results

All pilots had been fatigued when they came for their simulator mission. The simulator by intention was scheduled during the circadian low. At that time, the pilots had already been awake for 16 to 22.5 hours. In addition, almost half of them (28 pilots) had accumulated a sleep deficit of more than 2 hours. 14 pilots crossed more than three time zones within the past three days (SD-TZN) and another 9 pilots had just one or no off-days before the simulator (SD-DUT). The distribution scores for SD-TAW and SD-DEF are shown in figure 1 and 2.

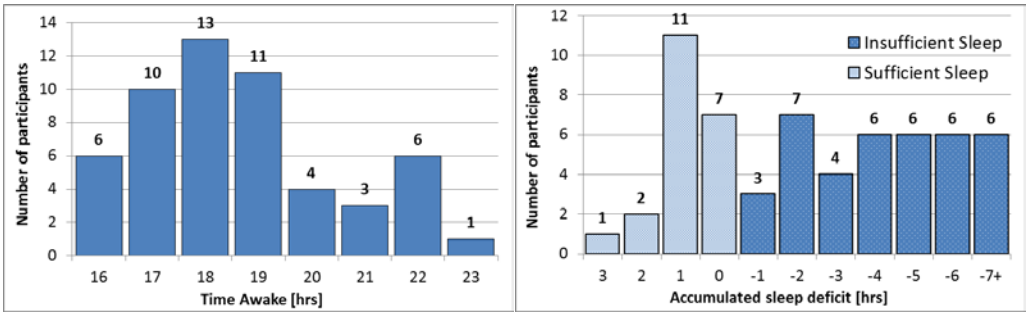


Figure 1 and 2. *Distribution of the number of hours being awake (left) and the accumulated sleep deficit over the past three days (right) before the simulator mission.*

Under these conditions of fatigue and the difficulty of the scenario, it was expected that the pilots would face a very demanding simulator mission. The mean scores of the RTLX-subscales are shown in figure 3. Subjective workload was highest with respect to M*ental Demands* and E*ffort*.

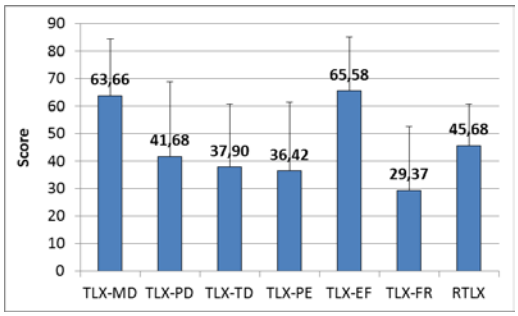


Figure 3. *Mean scores of the RTLX subscales and overall subjective workload. Error bars indicate one standard deviation.*

Paired sample T-Tests of the mean scores from the PVT and the subjective fatigue assessments showed, in several parameters, a significant ($\alpha < .05$) decrease of attention levels and an increase in fatigue. According to Cohen's d (Cohen, 1992), the effect sizes are medium for P-LAPS (d=.45), P-RT10 (d=.32), and FAT (d=.33) and small for VAS (d=.14). As illustrated in figures 4 to7, the inter-individual variances also increased from the *base*- to the *post*-measurement. FAT scores varied around 12, which means a moderate but not yet severe amount of fatigue.

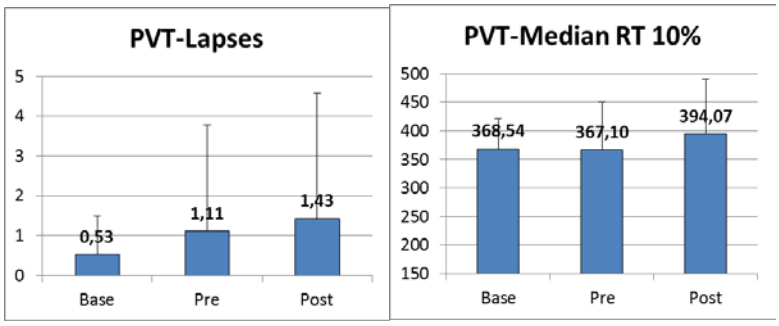


Figure 4 and 5. *Mean scores of PVT lapses (left) and the 10% slowest reaction times (right) during the experimental procedure. Error bars indicate one standard deviation.*

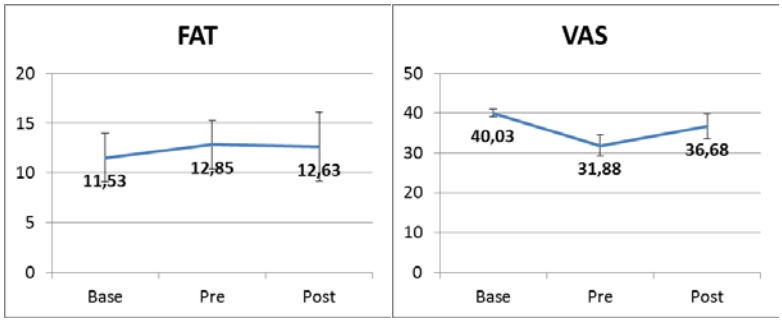


Figure 6 and 7. *Mean scores of subjective fatigue checklist scores (left) and visual analogue alertness assessment (right) during the experimental procedure. Error bars indicate +/- one standard deviation.*

Correlation analyses were conducted to explore the relationship between scores of sustained attention and workload. While the RTLX total score had no significant correlations to any PVT-scores, the task-related and subject-related RTLX-subscales showed some significant correlations ($\alpha < .05$) primarily with PVT-scores during *base*- and *pre*-measurement. Significant coefficients were for TL-PE .26 ($P\text{-}LAPS_{base}$), .23 ($P\text{-}MRT_{base}$), .41 ($P\text{-}MRT_{pre}$) and .28 ($P\text{-}MRT_{post}$). TL-EF had significant correlations of .33 ($P\text{-}RT10_{pre}$) and .32($P\text{-}MRT_{pre}$). TL-FR showed significant correlations of .41 ($P\text{-}MRT_{base}$) and .32 ($P\text{-}MRT_{pre}$). No significant correlations with the RCI of the PVT were observed.

Looking at the correlations with subjective fatigue assessments, the relationship to workload appeared stronger. For the RTLX total score, we found significant correlations to the *post*-measurements FAT_{post} of .29 and VAS_{post} of -.24. Also, pilots who experienced a higher workload during the simulator felt increasingly more fatigued afterwards. The correlations with RCI-FAT and RCI-VAS were .33 and -.32 respectively. All significant correlations with the *post*-measurement and with the change indices are shown in table 2.

Table 2.
Significant correlations (α < .05) between subjective fatigue (level and change scores) and workload during the simulator mission

	TL-MD	TL-PD	TL-TD	TL-PE	TL-EF	TL-FR	RTLX
FAT_{post}				.35	.27	.25	.29
VAS_{post}				-.38	-.32	-.24	-.24
RCI-FAT	.25			.42	.35		.33
RCI-VAS				-.37	-.33		-.32

Further analyses of the sleep logs revealed a positive relationship (α < .05) between the number of lapses in the first PVT measurement and time awake before the simulator (r = .30). Also, the accumulated sleep deficit, which really had a wide range (see figure 2), correlated .30 with $P\text{-}LAPS_{base}$. Sleep quality one night before the simulator correlated -.30 with FAT_{base}. Sleep quality three nights prior also showed significant correlation with subjective fatigue scores during the *base*-measurement of -.28 with FAT_{base} and .26 with VAS_{base}. Sleep quality of the second-to-last night correlated .27 with RCI-VAS, which meant less decrement of alertness with higher quality of sleep.

The strongest correlations with the objective PVT-scores of sustained attention were found with the rostering information and with the time of the simulator event itself as shown in table 3.

Table 3.
Significant correlations (α < .05) between change scores of subjective fatigue and workload during the simulator mission

	Base			*Pre*			*Post*		
	P-LAPS	P-RT10	P-MRT	P-LAPS	P-RT10	P-MRT	P-LAPS	P-RT10	P-MRT
Sleep deficit	.30								
Time awake	.30								
Duty days	.28	.42	.29		.34	.41	.26	.37	.45
Time zones		.29	.25						
Sim time	.26						.24	.30	.25

The time of the simulator event as a circadian factor was identified as the strongest predictor of change between the *base*- and the *post*-measurements. The correlations are with RCI-PVT .25, with RCI-FAT .48 and with RCI-VAS -.33.

Discussion

In summary, it can be confirmed by subjective as well as objective data that the pilots were moderately fatigued when they came to their simulator mission shortly before midnight. Fatigue further increased significantly during the three to four-hour experimental procedure. The level of fatigue could be predicted systematically by rostering information (e.g. number of duty days within the three days before the simulator) and some sleep history data (e.g. time awake and sleep quality).

As illustrated by the RTLX-data, the simulator mission was above average demanding with peaks for the task load factors of *Mental Demands* and *Effort*. However, the moderating effect of workload on increased fatigue could only be demonstrated for the subjective fatigue scores (FAT and VAS, table 2). While NASA RTLX scales were significantly correlated to several PVT scores of sustained attention, there was no significant interaction between the amount of workload and the amount of attention decrements from before to after the simulator. We did

not conduct explicit causal analysis here, but from the chronology of measurement it seems equally probable that workload causes attention decrements than that lack of sustained attention causes workload increments. It could be worth further investigating a common source of variance for workload and attention such as individual resources (training and basic abilities or simply the sleep history). Furthermore, individualized fitness-for-duty testing could become a promising option in this context (e.g. Elmenhorst et al., 2013).

The strongest predictor of change in levels of fatigue identified here was the time of the simulator event itself, which could illustrate the influence of circadian processes. However, with respect to our main question whether workload directly affects the levels of sustained attention, we did not find sufficient evidence. Only increases in subjective fatigue scores were significantly related to workload. An alternative explanation could be that the workload did not mount up high enough during the simulator or its effect appears with some time delay. To assess levels of individual fatigue risk in aviation, fitness-for-duty testing should complement FRMS recommendations.

Acknowledgements

Part of this work was funded by the German Federal Ministry of Economics and Technology via the Project Management Agency for Aeronautics Research within the Federal Aeronautical Research Program (LuFo IV-2). The authors acknowledge their thanks to Paul Kirchner for supporting this study and his contribution to the project.

References

Achermann, P. The two-process model of sleep regulation revisited. *Aviat Space Environ Med, 75(3*), 37–43.

Åkerstedt, T. & Folkard, S. (1990). A model of human sleepiness. In: J.A. Horne (Ed.) *Sleep '90*. Bochum: Pontenagel Press.

Borbély, A.A. (1982). A two process model of sleep regulation. *Human Neurobiology, 1*, 195–204.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Dinges D.F., Powell J.W. (1985). Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments & Computers, 17*, 652–655.

Elmenhorst, E.-M., Hoermann, H.-J., Oeltze, K., Pennig, S., Rolny, V., Verjvoda, M., Staubach, M. & Schießl, C. (2013). Validierung eines Fitness-for-Duty Tests zur Steigerung der Sicherheit in Luftfahrt und Verkehr. Research Report DLR-Project "FIT". DLR: Cologne/Germany.

Hart, S.G. (2006). NASA-Task Load Index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp 904–908). Santa Monica.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* . Amsterdam: North Holland Press.

Haslbeck, A., Eichinger, A., & Bengler, K. (2013). Pilot Decision Making: Modeling Choices in a Go-Around Situation. 17th International Symposium on Aviation Psychology, Dayton, Ohio.

Haslbeck, A., Kirchner, P., Schubert, E., & Bengler, K. (2014). A Flight Simulator Study to Evaluate Manual Flying Skills of Airline Pilots. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, pp 11–15). SAGE Journals.

Haslbeck, A., Schubert, E., Onnasch, L., Hüttig, G., Bubb, H., & Bengler, K. (2012). Manual flying skills under the influence of performance shaping factors. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *41*(Supplement 1/2012), 178–183. Retrieved from http://dx.doi.org/10.3233/WOR-2012-0153-178.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59 (1),* 12–19.

Mueller, S. T. (2011). PEBL: The Psychology Experiment Building Language. Retrieved on Nov 8, 2011 from http://pebl.sourceforge.net/.

Pfendler, C. & Pitrella, F.D. & Wiegand, D. (1994). *Workload measurement in human engineering testing and evaluation*. Wachtberg: Forschungsinstitut für Anthropotechnik. Report Nr. 109.

Samel, A., Wegmann, H.M., & Vejvoda, M. (1995). Jet lag and sleepiness in aircrew. *Journal of Sleep Research, 4*, 30-36.

Samn, S.W., Perelli, L.P., 1982. *Estimating aircrew fatigue: a technique with application to airlift operations*. Brooks AFB, USAF School of Aerospace Medicine. Technical report SAM-TR-82-21.

Tritschler, K. & Bond, S. (2010). The influence of workload factors on flight crew fatigue. 63rd Annual IASS, *Information at a Glance*. Milan, Italy.

Van Dongen, H.P.A., Maislin, G., Mullington, J.M., Dinges, D.F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep, 26*, 117–126.

Williamson, A., Lombardi, D.A., Folkard, S., Stutts, J., Courtney, T.K., & Connor, J.L. (2011). The links between fatigue and safety. *Accident Analysis & Prevention, 43*. 498-515.

THE EFFECTS OF BRIGHT LIGHT INTERVENTION ON FLIGHT CREW BEHAVIORAL ALERTNESS AND COGNITIVE FATIGUE

Lori Brown
Western Michigan University
Kalamazoo, Michigan USA
Geoffrey Whitehurst
Western Michigan University
Kalamazoo, Michigan USA

The study aimed to investigate the efficacy of bright light intervention to improve behavioral alertness and reduce cognitive fatigue in flight crew-members. During the four week study, crewmembers wore actigraph bands to monitor sleep behaviors. Self-assessed levels of sleepiness were recorded using the Karlosinska Sleepinees Scale (KSS), and self-assessed fatigue was measured using the Samn-Perelli (SP) fatigue scale. Participants completed psychomotor vigilance tests (PVT) to measure behavioral alertness. On the third and fourth weeks of the study, participants were exposed to short wavelength bright light (465nm blue) light intervention. The results show that there was a significant difference in alertness and cognitive fatigue between pre-intervention and post-intervention for each crew member and that 39.1% of the variance is explained by time (pre/post intervention). There is also a significant difference in alertness between flight crew and cabin crew and 49.4% of the variance is explained by position (flight/cabin crew).

Cognitive fatigue is a threat to aviation safety because of the impairments in alertness and performance it can create. It has significant physiological and performance consequences because it is essential that all flight crewmembers remain alert and contribute to flight safety by their actions, observations, and communications (Strauss, 2006). One poignant example which proved to be fatal for all onboard, was the crash of Colgan flight 3407, in Buffalo, New York, on February 12, 2009. According to the NTSB report (2010), "the pilot's performance was likely impaired because of fatigue". The first officer had traveled to work (commuted) all night from Seattle on a Fed-EX cargo aircraft and had been awake for 30 hours before the crash. The Captain had also traveled and slept in the crew room before the flight. This time traveling to work is not included in duty time, which increases the crewmembers 'time since awake.' Research shows that after 16 hours of 'time since awake' performance is similar to someone who is legally drunk. Additionally, there have been numerous occurrences of pilots falling asleep while on duty. These aircraft are operated by the autopilot, and are at risk or mid- air collisions, running low on fuel, impact into terrain, or being mistaken for hijacked aircraft and possible intercepted or shot down, if no communications are established. Western Michigan University, College of Aviation, Jeppesen (a Boeing Company), Nature Bright Company, Airline participants, and a leading sleep researcher Schoutens, A.M.C. of FluxPlus, BV, The Netherlands, collaborated to examine whether timed blue light could improve flight crewmember alertness and reduce cognitive fatigue.

Countermeasures

Safety being the most important objective in aviation can be improved with effective countermeasures to reduce fatigue related errors. Countermeasures can be classified in two categories, preventative and operational. Preventive strategies are used before flying or between flights to reduce the effects of fatigue, sleep loss, and circadian disruption. Proper sleep hygiene, a nap (no longer than 45 minutes) before a flight schedule, hydration, nutrition, exercise, and in some cases -treatment for sleep apnea have all been cited as effective preventative strategies for flight crew. These techniques can help to decrease the likelihood of the crewmember starting the trip with a sleep deficit. Whereas, 0perational strategiesare used during flights to maintain alertness and performance include controlled timed napping, hydration, bright light, strategic use of caffeine, proper nutrition, short walks when able (flight attendants only),and in seat stretches for pilots. The need of a combination of napping and other countermeasures, to improve alertness has been demonstrated by, Garbarino et al., 2004; Gronfier et al., 2007. While operational countermeasures can be most effective when combined with other operational or preventative countermeasures, this study focused on the bright light intervention.

Light Interventions

Alerting effects of light are tied to its suppression of melatonin, which is ordinarily released in the evening and night. Researchers Cajochen et al. (1999) showed measurable increases in subjective alertness and reductions in slow eye movements, with short wavelength (blue) light appearing to have the greatest alerting effect. There is also evidence that the alerting effects of light are independent of the time of day, leading to the possibility of employing light during the daytime to improve alertness and performance in individuals impaired due to prior sleep deprivation (Cajochen, 2007). Capitalizing on the immediate, direct alerting effects of light for flightcrew is particularly useful because the flight-deck environment with its high automation level, "limited opportunities for physical activity or social interaction, steady background noise, and low nighttime light levels creates a setting ripe for boredom, complacency, attention lapses, sleepiness, and performance decrement" (Caldwell, 1997). As shown in research conducted by Leger (2004), "bright light could be an effective countermeasure" and warrants further study. Despite the widespread use of light intervention in competitive sporting environments, the possible impact of blue or bright light therapy (ocular therapy) in aviation has received little attention, compared to caffeine.

A review of literature on the acute alerting effect of light shows it to be a potentially useful countermeasure where conditions allow its use. Light treatments may be easy to apply in real aviation occupational settings such as : crew check-in rooms, at home before flight schedule, hotel layovers, air-traffic control break rooms, and possibly aircraft galleys. Several studies have noted that a combination of countermeasures have a more pronounced affect compared to a single countermeasure (Wright JR, K., Badia, P., Meyers, B., and Plenzler , 1997). Based on this hypothesis, Léger et al., (2004) designed a preliminary study to test the effects of combination of napping and bright light pulses in a pilot group of shift workers. This study used short pulses (10 min) of 5000 lux white light, combined with naps (Leger et al., 2004). Both the number and the duration of the episodes of sleepiness were reduced by the intervention" (Leger et al., 2004). Beaven and Ekstrom reported (2013) that both the caffeine only and blue light only conditions enhanced accuracy in a visual reaction test requiring a decision and an additive effect was observed with respect to the fastest reaction times. Research also suggests that natural light has the same beneficial affects providing the crewmembers have the ability to receive natural light treatment in the operational setting-which makes the use of small lightweight portable artificial light units appealing. Adjusting the light level and color temperature is one of a limited number of possible environmental manipulations. A recent study (Brown, 2014) funded by Western Michigan University (Kalamazoo, Michigan, USA) FRAACA award was the first to look at the effects of blue light (460nm) in the occupational setting with flight crew members. The study aimed to investigate the efficacy of blue light therapy to improve alertness in flight crew-members. Western Michigan University, College of Aviation, Jeppesen (a Boeing Company), Nature Bright Company, Airline participants, and a leading sleep researcher Schoutens, A.M.C. of FluxPlus, BV, The Netherlands, collaborated to examine whether timed blue light could improve flight crewmember alertness, and mitigate cognitive fatigue- as seen with gold medal Olympic athletes to improve performance.

Methods

Fourteen flight crew members, males (n=9) and females (n=5), working as pilots or flight attendants, participated in the study under the Western Michigan University IRB approved protocol. All participants were nonsmoking, active flight crewmembers. The crewmembers were based in Sweden and maintained flight schedules to the Mediterranean and the Canary Islands, as well as long-haul flights to Thailand, India and Vietnam. Each participant was provided an informed consent document and attended a two hour training session on the use of the light and actigraphy band. Each participant was provided with a confidential code and completed the Morningness-Eveningness Questionnaire (MEQ), a self-assessment questionnaire (Horne & Ostberg, 1976), to measure their peak sleepiness and alertness time (diurnal type). The MEQ was used once at the beginning of the testing period to assess the habitual and preferred weekday and weekend clock times of the participants. The MEQ is a 19 item, self-report instrument that consists of questions in which the participant indicated their preference using a 4-point Likert Scale. During the 30 day study, the crewmembers wore actigraph wrist bands to record sleep/wake behaviors, and recorded self-assessed levels of sleepiness with the Karolinska Sleepiness Scale (KSS). Self-assessed fatigue was recorded using the Samn-Perelli Fatigue Scale (SP), and completed daily psychomotor vigilance tests (PVT). On the third and fourth weeks, the flight crew-members were exposed to blue light (BL) in field-based treatment with short wavelength (460nm) light therapy. Data collection was through the (iOS) Boeing Alertness Model (BAM) application called Jeppesen CrewAlert Lite, which can be used anywhere in the world with an iPhone, iPod, or iPad device.

Equipment

Nature Bright Company provided 20 Square One® rechargeable portable lightweight wake-up lights which weighed less than 2 lbs. The Square One light provides blue (λmax = 465 nm, 84.8 µW/cm2, 39.5 lux, 1.74 × 1014 photons/cm2/s) light intervention and is one the smallest light therapy devices on the market, with an advanced optical lens, and a wakeup light alarm. The Square One (figure 1.) was selected due to the small portable size, ideal for crewmembers, as it was easy to place in a flight bag, handbag, or luggage.

Figure 1. Nature Bright Square One® rechargeable portable light[http://www.naturebright.com/]

CamNTech MotionWatch 8 actigraphy wrist band with a tri-axial digital accelerometer were worn for 30 days by all participants. Actigraphy has been used in studies to measure sleep/wake patterns for over 20 years (AASM, 2010). The advantage of actigraphy over traditional polysomnography (PSG) is that actigraphy is non-invasive (a water proof watch band) and can conveniently record continuously for 24-hours a day for days, weeks or even longer. The waterproof Motionwatch 8 provide usb downloaded activity plots coupled with specialized software used to quantify the intensity and duration of daily physical activity. These data was analyzed to identify irregular activity patterns for assessment of sleep quality. Individual daily sleep efficiency and sleep bouts were used to look for correlations with the KSS, SP, and PVT results. The band also measured the amount of lux the participant was exposed to.

Results

A repeated measures multivariate analysis of variance (MANOVA) was conducted, using IBM SPSS Statistics 20 software, to test the intervention effect of blue light (IV) on both flight and cabin crew alertness, measured by the 4 DVs; KSS, SP, PVTR, and PVTL. A one-way MANOVA revealed a significant multivariate within-subject main effect for time (pre and post light intervention), Wilks' λ = .609, F (4,55) = 8.843, p <. 001, partial eta squared = .391, and the power to detect the effect was .999. The analysis also revealed a significant multivariate between-subject main effect for position (pilot/flight attendant), Wilks' λ = .506, F (4,55) = 13.429, p <. 001, partial eta squared = .494, and the power to detect the effect was 1.000.

The results show that there was a significant difference in alertness between pre-intervention and post-intervention for each crew member, and that 39.1% of the variance is explained by time (pre/post intervention). There is also a significant difference in alertness between flight crew and cabin crew, and 49.4% of the variance is explained by position (flight/cabin crew).

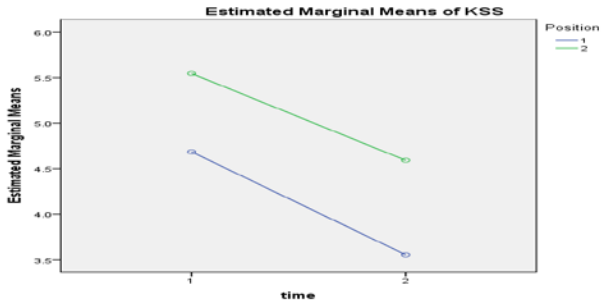


*Figure 2***.** KSS Sleepiness Pre/Post Intervention Marginal Means by Position

Figure 2 above shows that for the measure Karolinska Sleepiness nine point Scale (KSS), there is a similar intervention effect for both pilots and cabin crew, but there is a difference in the estimated marginal means related to crew position (1 = pilot and 2 = cabin crew). The Karolinska Sleepiness Scale (KSS) is a 9-point Likert scale based on a self-reported, subjective assessment of the subject's level of drowsiness at the time where 1 = extremely alert and 9 = extremely sleepy/fighting sleep. The independent measure derived from the KSS Checklist was self-rated sleepiness. Higher scores indicated a higher level of subjective sleepiness. KSS has been used widely, particularly for describing changes over time within subjects (Gillberg et al., 1994). It is clear that both pilots and flight attendants had a decreased self-assessed sleepiness; however, the reason for the difference in the estimated margin of means based on crew position was not evident.

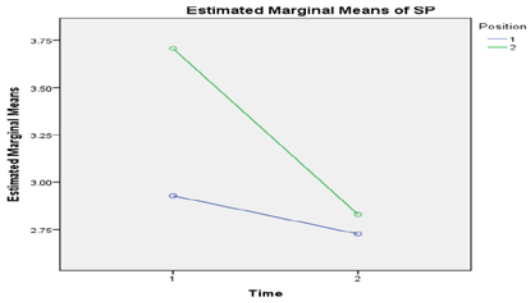


Figure 3. SP Pre/Post Intervention Marginal Means by Position

Figure 3. above shows that for the measure PVTR there is a positive intervention effect (reduced reaction time) for cabin crew, but not for flight crew. However, there is still a difference in the estimated marginal means related to crew position (1 = flight crew and 2 = cabin crew). Subjective fatigue was assessed using the Samn-Perelli Fatigue Checklist [30]. The Samn-Perelli is a 7-point Likert scale, where 1 = fully alert/wide awake and 9 = completely exhausted, unable to function effectively. Higher scores indicated a higher level of subjective fatigue (Samn and Perelli, 1982). Vigilance was assessed with the Psychomotor Vigilance Test (PVT), a 5-minute iOS visual reaction-time task which evaluates sustained attention [4]. Participants were instructed to respond to the appearance of a visual stimulus by tapping a black bulls-eye target on the iOS screen as quickly as possible. During each 5-min session, visual stimuli appeared at variable intervals of 2–10 s. From each PVT trial, reaction times (RTs) were collected and 2 performance variables, average response time and number of lapses (i.e. failure to respond or RT > 500 msec) were extracted by Jeppesen Crew Alert.

Figure 4. below shows that for the measure PVTL there is a similar intervention effect for both flight crew and cabin crew, but there is a difference in the estimated marginal means related to crew position (1 = flight crew and 2 = cabin crew).

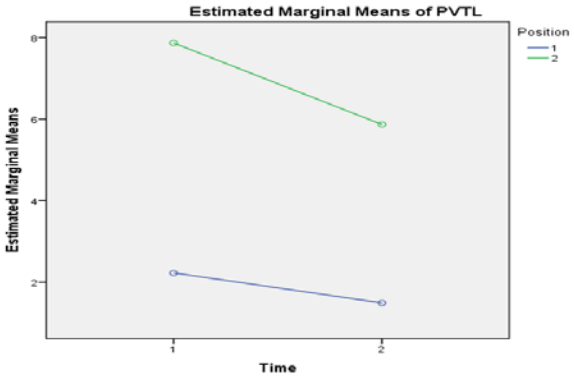


Figure 4. PVTL Pre/Post Intervention Marginal Means by Position

The results reveal that crewmembers may be able to improve behavioral alertness with the use of bright light interventions as a fatigue countermeasure to improve occupational safety in transportation. A review of literature and results of this study shows the acute alerting effect of blue light to be a potentially useful countermeasure to reduce physiological, perceived, and cognitive fatigue, where conditions allow its use. Results garnered can be used to develop innovative light therapies and preventive strategies for industries with shift workers such as aviation, maritime, rail, nuclear, and medical.

Recommendations

The benefits of light therapy extend well beyond aviation, and are often used with depression, dermatology, psychiatry, neurology and gerontology and work related issues such as, shiftwork and sports medicine (Dutch Olympic Swimming Team, TVM Ice -skating team, Dutch Olympic Committee). In addition to individual crewmembers using portable light units as a fatigue countermeasure, crewmembers can benefit from simple 'light stations' in crew check-in areas and light effect can be integrated into alertness models. Expanding the limited body of scientific knowledge about light effect on alertness may allow us to integrate light/dark effect into alertness models, to improve fatigue management systems. With algorithms indicating peek and low times in the schedule we may be able to determine when and how long the crewmember should seek natural or artificial light beyond adjusting circadian rhythms.

Although natural light is not always practical in the aviation setting, portable small light weight units for flight bags and desktop light boxes placed in crew break areas, (10,000 lux, 17,000 Kelvin UV-Free lights which mimic a blue sky) could be effective. Operators can work with their flight surgeons and health departments to discuss these options. Educating crewmembers on the acute effect of light on their sleep, mood, and alertness is crucial, particularly as we look closer at the quality of life and sleep issues with crewmembers —such as sleep apnea. In addition to improved alertness, relief from seasonal affective disorder could also be a benefit- particularly in the dark winter months in areas such as Seattle, Norway, Sweden, Canada and Michigan.

Clearly, there is still a need for further research on the best ways to integrate specific timed light in the occupational setting, perhaps drawing on some of the innovate mood lighting for passengers in modern aircraft, and evaluations of the most appropriate spectrums. One area which may deserve further research is red light intervention —which may conducive to night flight deck operations. In a study conducted by researchers Levent, et al., (2013), at Rensselaer Polytechnic Institute shows that exposure to red wavelengths and levels of light has the potential to increase alertness. Providing future research looks into the spectral sensitivity of alertness and how if it changes over the course of 24 hours, this would be helpful for building light into bio-mathematical alertness models. Light therapy is not new to Western Michigan University (WMU) or the aviation industry. Currently the WMU health center (sindecuse) offers light therapy for students and faculty aimed at the treatment of seasonal affective disorder. Additionally, sky effect lighting has replaced florescent tubes in a WMU classroom to improve concentration for students, along with an innovative 'light bar' was installed in the student lounge area at the College of Aviation. Students can bask under light therapy while studying or relaxing in between classes. The light bar is also used to educate students about the relationship of fatigue and aviation accidents, countermeasures and alertness strategies. This is an opportunity to apply the same concepts to transform countermeasures in the flight deck. Recently, we have seen a flux of light therapy innovations aimed at passengers to improve mood, decrease effects of jet lag and minimize fatigue, include commuter train installations in the UK. Next generation aircraft such as the B787, and A380 have mood lighting installed to help passengers adjust to new time zones. Paris Charles de Gaulle airport installed three light therapy 'spaces' where passengers can enjoy light therapy to fight their jet lag. The light can be used before, during, or after the flight. The Nature Bright Company Sun Touch 'light boxes' are currently used in light bars at WMU, medical facilities, and at airlines such as Novair.

Acknowledgements

The authors would like to express their gratitude for all of the airline participants who volunteered their time to participate in this study; participating Airlines and staff; Jeppesen, A Boeing Company; Tomas Klemets; Gregory A. Pinnell MD, senior AME, senior Flight Surgeon USAFR; Undergaduate Research Assistants Troy Booker, Travis Davis, Industry Aviation Human Factors Consultants, Jeanne Kenkel, Sherry Saehlenou, and Captain John Gadzinski; Light therapy researcher Toine Schoutens; Nature Bright Company, Western Michigan University and CamNtech. This study would not have been possible without your collaboration. Funding/Disclosure: The study

was funded by Western Michigan University, Faculty Research and Creative Activities Award (FRACAA). Equipment for the study was provided by Nature Bright Company. Lori Brown serves as scientific advisor for Nature Bright Company. KSS, SP and PVT data were collected by Jeppesen CrewAlert iOS APP by Tomas Klemets from Jeppesen Boeing Company. Data analyzed be Dr. Geoff Whitehurst, Western Michigan University.

References

Beaven CM, Ekström J (2013) A Comparison of Blue Light and Caffeine Effects on Cognitive Function and Alertness in Humans. PLoS ONE 8(10): e76707. doi:10.1371/journal.pone.0076707.

Brown, L., Schoutens, A.M.C., Whitehurst, G., Booker, T., Davis. Losinski., Diehl, R., (2014) The Effect of Light Therapy on Flight Crew-members Behavioral Alertness to Militate Fatigue and Improve Performance. *Social Science Research Network.* Retrieved from http://ssrn.com/abstract= 2402409.

Cajochen C. (2007). Alerting effects of light. *Sleep Medical Review*. 11 : 453 – 64.

Caldwell, J.A. (1997). Fatigue in the Aviation Environment: An Overview of the Causes and Effects As Well As Recommended Countermeasures, *Aviation Space and Environmental Medicine* 1997, 68:932-8.

Garbarino S, Mascialino B, Penco MA, Squarcia S, De Carli F, Nobili L, Beelke M, Cuomo G, Ferrillo F. (2004). Professional shift-work drivers who adopt prophylactic naps can reduce the risk of car accidents during night work. Sleep. 2004;27(2):1295–1302.

Gronfier C., Wright K.P., Jr., Kronauer R.E., and Czeisler C.A. (2007). Entrainment of the human circadian pacemaker to longer-than-24h days. *Proc. Nationa. Academy Science*. 104: 9081.

Gillberg, M., Kecklund, G., and Akerstedt, T. (1994). Relations between performance and subjective ratings of sleepiness during a night awake. *Sleep, 17(3), 236-41.*

Horne, J. A., & Ostberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology, 4(2), 97-110.*

Jeppesen (2013). Jeppesen Introduces CrewAlert Pro for iPad to Manage Fatigue Risk for Airline Crew. Englewood, Colo. 09 Sep 2013, Jeppesen. Retrieved from http://ww1.jeppesen.com/company/newsroom/articles.jsp?newsURL=news/newsroom/2013/CrewAlertPro_NR.jsp.

Leger, Damien, Philip, Pierre, Jarriault, Philippe, Metlaine, Arnaud, Choudat, and Dominique (2008). Effects of a combination of napping and bright light pulses on shift workers. *Sleepiness at the wheel: a pilot study*. European Sleep Research Society.

Rensselaer Polytechnic Institute (RPI). (2013, April 22). Red light increases alertness during 'post-lunch dip'. *ScienceDaily*. Retrieved from ww.sciencedaily.com/releases/2013/04/130422100801.htm.

Samn, S. W., & Perelli, L. P. (1982). Estimating aircraft fatigue: a technique with application to airline operations *(Technical Report No. SAM-TR-82-21)*: USAF School of Medicine, Brooks AFB, TX.

Strauss, S. (2006). Pilot fatigue. *Aerospace Medicine NASA/Johnson Space Center*, Houston, Texas. http://aeromedical. org/Articles/Pilot_Fatigue. html.

Wright Jr, K.., Badia, P., Myers, B., & Plenzler, S. (1997). Combination of bright light and caffeine as a countermeasure for impaired alertness and performance during extended sleep deprivation. *Journal of sleep research*, 6(1), 26-35.

TOWARD HEAD-UP AND HEAD-WORN DISPLAYS FOR EQUIVALENT VISUAL OPERATIONS

Lawrence (Lance) J. Prinzel III, Jarvis J. (Trey) Arthur, Randall E. Bailey, Kevin J. Shelton, Lynda J. Kramer, Denise R. Jones, Steven P. Williams, Stephanie J. Harrison, Kyle K. Ellis

NASA Langley Research Center
Hampton, VA

A key capability envisioned for the future air transportation system is the concept of equivalent visual operations (EVO). EVO is the capability to achieve the safety of current-day Visual Flight Rules (VFR) operations and maintain the operational tempos of VFR irrespective of the weather and visibility conditions. Enhanced Flight Vision Systems (EFVS) offer a path to achieve EVO. NASA has successfully tested EFVS for commercial flight operations that has helped establish the technical merits of EFVS, without reliance on natural vision, to runways without category II/III ground-based navigation and lighting requirements. The research has tested EFVS for operations with both Head-Up Displays (HUDs) and "HUD equivalent" Head-Worn Displays (HWDs). The paper describes the EVO concept and representative NASA EFVS research that demonstrate the potential of these technologies to safely conduct operations in visibilities as low as 1000 feet Runway Visual Range (RVR). Future directions are described including efforts to enable low-visibility approach, landing, and roll-outs using EFVS under conditions as low as 300 feet RVR.

Commercial aviation accident statistics evince the hazards associated with the approach and landing phase of flight. Boeing (2013) reported that 41% of all fatal accidents (2003-2012) occurred during the final approach and landing phase of flight, but approach and landing phases represents only 4% of flight time exposure. Low visibility is often reported as the contributing factor in as much as 90% of controlled flight into terrain (CFIT) landing accidents wherein less than 60% involve high terrain.

In 2003, the U.S. Government established the Next Generation Air Transportation System (NextGen) to transform the national air transportation system. An emerging NextGen concept, termed "Equivalent Visual Operations" (EVO), strives to replicate the airport capacity and safety now achieved under visual flight rules (VFR) in all weather conditions to mitigate, even eliminate, low visibility as an etiology (see Bailey, Prinzel, Kramer, and Young, 2011 for an alternative concept termed, "Better Than Visual"). Today, an alternative, intuitive means of conducting low visibility operations and possibly achieving EVO, is available. EFVS offers an "all-weather" capability, independent of the weather or vision obscurant, without significant aircraft or airport investment that creates real world-like visibility. The use of EFVS supports the Federal Aviation Administration (FAA) 2014 NextGen Implementation plan for "Improved Approaches and Low-Visibility Operations" (FAA, 2014).

Enhanced Flight Vision System

"Enhanced Vision" (EV) refers to an electronic means to provide a display of the external scene by use of an imaging sensor. The FAA defined, "Enhanced Flight Vision System" (EFVS), as, "... an installed aircraft system which uses an electronic means to provide a display of the forward external scene topography (the applicable natural or manmade features of a place or region especially in a way to show their relative positions and elevation) through the use of imaging sensors, ..." An EFVS uses a head-up display (HUD), or equivalent display to provide flight information, navigational guidance, and real-time imagery of external scene via imaging sensors. On January 9, 2004, a final rule, Enhanced Flight Vision Systems, was published in Federal Register (69 FR 1620) that allows an EFVS to be used in lieu of natural vision to descend below the decision altitude/height (DA/DH) or minimum descent altitude (MDA) down to 100 feet above the touchdown zone elevation (TDZE) of intended landing runway.

Proposed EFVS Rulemaking

On December 16, 2010, RTCA SC-213/EUROCAE Working Group 79 (established December 2006) published DO-315A which developed minimum aviation system performance standards (MASPS) which extended the operational credit established under CFR 91.175 (l) and (m) enabling EFVS operations below the 100 feet TDZE to touchdown and rollout (without the requirement for a natural visual segment) down to visibilities as low as 1000 feet RVR (RTCA, 2010). On June 11, 2013, the FAA published a Notice of Proposed Rulemaking to enable EFVS-equipped aircraft to conduct operations down to touchdown and rollout under visibilities as low as 1000 feet RVR.

The benefits would significantly expand EFVS operations, which should increase efficiency, allowing access to more runways, allowing for new EFVS operations, and minimizing the need for go-arounds and missed approaches during low visibility approach and landing operations (see FAA 2010; 2013). As the FAA observed (FAA-2013-0485-0001), however, there does not exist sufficient historical data to quantify these benefits.

NASA has conducted numerous high-fidelity simulation and aircraft flight test research to provide the requisite data to inform the proposed rulemaking to extend §91.175 operating rules to enable EFVS operations with lower visibility minimums. Bailey, Kramer, and Williams (2010) provide a review of NASA research that describes the efforts that helped to make "operational credit" EFVS HUD operations a reality.

EFVS Equivalent Displays

With many operational credits being provided by HUD operations (e.g., AC-120-28D; FAA Order 8400.13), one possible avenue of HWD adoption across the NextGen fleet is by providing a "HUD-equivalent capability." The requirements for a HWD to meet a HUD-equivalent capability may be derived from FAA guidance material and these "essential features" are described in Bailey, Kramer, & Williams, 2010. NASA has conducted research to evaluate prototype HWD systems as a potential replacement for a HUD as an EFVS. If this equivalence can be shown, then the unique capabilities of the HWD - that is, unlimited field-of-regard head-up operations for low visibility flight operations - can be capitalized. The design challenge (and certification challenge) is to create this equivalent capability without increasing pilot workload, or encumbrance, or obscuration of their normal vision.

Recent NASA HUD/HWD EFVS Research

The following describes three representative examples of simulation and flight test research that have examined the use of EFVS of HUDs and HWDs for the revised §91.175 and RTCA SC-213 proposed extensions for EFVS operations to 1000 ft. RVR. Abbreviated descriptions of methodology and experimental results are provided with references to obtain more detailed information.

HUD EFVS High-Fidelity Simulation

A fixed-based experiment was conducted to evaluate the operational feasibility, pilot workload, and acceptability of conducting straight-in instrument approach procedures with published vertical guidance using EFVS for the approach, landing, roll-out, and turn-off in simulated visibility as low as 1000 ft. RVR (see Kramer, Bailey, et al., 2013).

Pilot Participants

Twenty-four pilots served as participants for the research. The pilots were paired by airline and role (Captain, First Officer) to ensure crew coordination and cohesion with regard to terminal and surface standard operational procedures. All pilots were required to hold an Airline Transport Pilot rating and average pilot experience was over 12,000 flight hours.

Simulation Facility

The research was conducted in the Research Flight Deck (RFD) at NASA LaRC, which is a high-fidelity, 6 degrees-of-freedom motion-based large commercial aircraft simulator with full-mission capability and advanced glass flight deck displays. The out-the-window (OTW) scene was generated by an Evans and Sutherland Image Generator graphics systems providing approximately 200° H by 40° V field-of-view (FOV) at 26 pixels per degree. All standard audio call-outs were generated. The HUD was a Rockwell-Collins HGS-4000 HUD.

Enhanced Vision Simulation

The EV real-time simulation is created by the Evans and Sutherland EPX physics-based sensor simulation. The EV simulation mimicked the performance of a short-wave/mid-wave forward looking IR (FLIR) sensor, using a ~1.0 to 5.0 micron wavelength detector. The nominal enhanced visibility was approximately 2400 feet for this experiment. The EV eye point reference/parallax error was 2.5 milliradian (mrad) to a point located 2000 feet away (DO-315 specifies 5 mrad max).

Evaluation Task

Approaches were flown only to runways with Medium intensity Approach Lighting System with Runway alignment indicator lights (MALSR) installed. ORD Runways 4R, 9R, 22L, or 22R were used. All runways had available high intensity runway lights and serviceable centerline and surface markings. Airport lighting was drawn using calligraphics. The evaluation task was a straight-in Instrument Landing System (ILS) approach that started three nautical miles (nm) from assigned runway threshold with a three degree descent angle. The weather consisted of low to moderate winds with either ten knot headwind, ten knot tailwind, 7.5 knot crosswind, or 15 knot crosswind, light turbulence (root-mean-square (rms) of 1 ft/sec), and varying OTW visibility levels (1800 feet, 1400 feet, or 1000 feet RVR). Auto-throttles were used for all approaches.

Experimental Results

Landing criteria of Joint Aviation Authorities All Weather Operations (JAR-AWO) and AC-120-28D (Appendix 3, section 6.3.1) was adopted from CAT III requirements for the purpose here to evaluate EFVS landings. Overall, the touchdown statistics evinced to be within the "desired" range for both longitudinal and lateral position and "adequate" for sink rate at touchdown. No go-arounds were conducted for trials with the EFVS HUD and the positional performance was excellent. Pilots reported "moderate, easily managed" (Ames & George, 1993) workload.

HUD EFVS Flight Test

The flight test evaluated synthetic and enhanced vision systems in partnership with Honeywell and Gulfstream with the objectives to determine (see Shelton, Kramer, Ellis, & Rehfeld, 2012) operational feasibility, pilot workload, and pilot acceptability of conducting a straight-in instrument approach with published vertical guidance using EFVS during approach, landing, roll-out, and runway exit in visibility of at least reported 1000 RVR.

Pilot Participants

Six pilots participated in the flight evaluations representing a cross-section from commercial, military, corporate, and the FAA (FAA test pilot). Average total flight time was 9108 hours with a max/min of 16250 and 4800 hours, respectively. Average commercial pilot experience was 28 years (range of 35 to 19 years). All pilots had flight experience with EFVS (379 average hours).

Test Aircraft

The flight test was conducted using Gulfstream's G450 flight test aircraft N401SR, S/N 4001. The test aircraft was equipped with certified avionics and software including the Honeywell SV-Primary Flight Display (PFD) and monochromatic EFVS HUD with display of conformal symbolic information, flight information, and FLIR imagery. The aircraft's certified avionics are described in Shelton, Kramer, Ellis, and Rehfeld (2012).

Enhanced Flight Vision System

The certified EFVS onboard consisted of a Rockwell-Collins' model HGS 6250 and Kollsman Enhanced Vision System (EVS) II infra-red camera (FLIR) and approved to conduct EFVS operations, based on electronic flight visibility, to descent below published minima to 100' HAT (14 CFR §91.175(l), (m)). The Kollsman II EVS has a FLIR sensitivity of less than 5mK, IR spectrum 1 to 5 Micron, and 30°H x 22.5°V FOV.

Evaluation Task

Shelton, Kramer, Ellis, and Rehfeld (2012) describe the training, airport and runway selection criteria, and crew procedures. Nine test flights were flown in Gulfstream's G450 flight test aircraft and pilots flew 108 approaches (SVS, EFVS, and baseline displays) in low visibility weather conditions (600 feet to 3600 feet reported visibility) under different obscurants (mist, fog, drizzle fog, frozen fog) and sky cover (broken, overcast). A total of 73 useable EFVS approach evaluations were conducted with 53 touchdowns, and 20 (27%) missed approaches; the 20 go-arounds were all conducted safely based on decision criteria established for the FAA exemption waiver (FAA "Certificate of Waiver" was issued April 1, 2011 thru March 31, 2012) to conduct the approaches below published DH/DA/MDA to landing using an EFVS.

Experimental Results

Out of the 80 EFVS approaches, seven were culled out of the data analysis for various extraneous reasons such as: Approach Lightning System (ALS) automatically turning off, or the evaluation pilot mistakenly left autopilot on during much of the approach, etc. These events were anomalous and caused significant deviations from the nominal operation and therefore, were not representative of the other approaches.

Of the 73 useable EFVS approach evaluations, 53 (73%) resulted in a touchdown and 20 (27%) resulted in missed approach. Eight of the EFVS approaches were to an offset runway. The 20 missed EFVS approaches were all conducted safely with the go-around decision correctly determined based on conditions. All approaches were within Category II approach minima, as outlined in AC120-29A, for the glideslope vertical CAT II minima (0.46 dots) and localizer lateral CAT II minima (0.33 dots), with the exception of one approach (lateral deviation = 0.37 dots), in a challenging crosswind, that resulted in a safe successful touchdown. RMS EFVS Landing Decision Altitude call-out for touchdowns was 126 feet radar altitude versus. 163 feet for missed approaches. The touch-down means reported were for longitudinal (2058 feet, δ = 501 feet) and lateral (3.47 feet, δ = 3.28 feet). Pilot workload ratings (Ames & George, 1993) ranged from “easily managed” during landing (2.5 rating) and go-around (2.9 rating).

HUD/HWD EFVS High-Fidelity Simulation

The NASA HUD/HWD EFVS RFD simulation study was conducted to evaluate “equivalent displays” of head-worn displays (HWD) for manually flown approach and landing EFVS operations under simulated visibilities as low as 1000 feet RVR (see Arthur et al., 2014).

Pilot Participants

Twenty-four commercial airline transport pilot-rated pilots participated in the research and had familiarity with the Memphis International Airport (FAA identifier: MEM). All pilots were required to have significant HUD experience (>100 hours) and preference was given to those with EV/EFVS training. Pilots were paired by airline and role, as in previous studies, forming twelve flight crews.

Head-Up Display

The HGS6700 commercial HUD is collimated and subtends 46°H by 34.5°V FOV with a 1400 x 1050 display resolution and greater than 4,000fL display brightness. The HUD system was measured to be 14 kg in weight. The HUD provided stroke FLIR imagery.

Head-Worn Display

A prototype head tracker was used to provide head orientation and was mounted on the left side of a pair of Lumus© DK-32 glasses. The head tracker was a hybrid-inertial tracker with image processing to correct for inertial drift and standard methods were used for ensuring accurate head tracking. The HWD is see-through, full color (green monochrome only used to be consistent with HUD) which utilizes patented Light-guide Optical Element (LOE) technology. The HWD was collimated and subtends 35°H by 20°V FOV with a 1280 x 720 display resolution and greater than 1,000fL display brightness (these specs are markedly lower than the HUD used). The image focal plane matched the HUD at infinity (using LOE). The measured weight was 0.20 kg.

Enhanced Vision Simulation

The same Evans and Sutherland EV real-time, physics-based sensor simulation was used as in the HUD EFVS simulation experiment described earlier, which is capable of modeling a wide range of sensors (image intensification, low-light, and infrared) and wavelengths. The MEM database was instantiated with material code properties. From this database, an IR sensor simulation, interacting with this material-coded database and the simulated weather conditions, created the desired test experimental conditions. As in previous experiments, the EV simulation mimicked the performance of a short-wave/mid-wave FLIR, using a ~1.0 to 5.0 micron wavelength detector. The nominal enhanced visibility was approximately 2400 feet for this experiment with a 2.5mrad eye reference/parallax error.

Evaluation Task

Flight crews conducted manually flown approach and landing operations to MEM runways (36L, 36C, 36R) starting at 1000 feet HAT. The EFVS crew procedures were trained and utilized for all HUD EFVS approach trials. The experiment conditions replicated actual operating conditions, lighting systems, operational procedures, required call-outs, and air traffic controller-pilot communications. All pilots reported that the simulation emulated real-world operations and workload typically experienced during low-visibility operations.

Experimental Results

An Analysis of Variance (ANOVA) was conducted on Flight Technical Error between HUD and HWD displays for localizer dot error and glideslope dot error tracking performance from an altitude of 1000 feet to 50 feet AGL. The results found no significant effects for RMS localizer, glideslope, or sink rate. The same dependent measures were analyzed via ANOVA to examine the effect of the display concepts at published decision height (200 feet) to threshold crossing height (50 feet HAT); this is the "equivalent visual segment." The statistical results showed that the display concepts were not significantly different from each other, during the equivalent visual segment. For the landing phase, the results on touchdown statistics further showed no significant differences between the HUD and HWD for longitudinal distance from threshold, lateral distance, or sink rate. The landing results evince that all landings using either the HUD or HWD were within the AC 120-28D CAT III minima criteria of "desired" (albeit these criteria are based on auto-land performance). The qualitative data also showed that pilots rated the HUD and HWD equivalents in terms of situation awareness and workload measures.

Conclusions

The research on HWDs and HUDs extend beyond the need to evaluate the efficacy of these technologies to achieve EVO. The experiments described are representative examples of NASA efforts to enhance the flight deck to revolutionize how low-visibility approach operations, using an EFVS, are conducted today and in the future. If successful, these works will establish the precedence that an electronic means of visibility can be used in lieu of a pilot's natural vision – a *first* that will open up many new capabilities. The research delineated here evince that a head-up (HUD or HWD) EFVS can safely enable 1000 feet RVR approaches without need of all the many expensive ground-based requirements and significantly reducing airport costs and expanding the number of runways operational under low-visibility conditions. The research establishes the advance of HWD technology that is fast approaching HUD EFVS "display equivalency" while also substantiating the advantages afforded these unlimited field-of-regard displays.

Since 1929, Instrument Flight Rules (IFR) has been conducted by pilots using abstract cockpit instrumentation and navigational aids to allow penetration of the weather until a pilot can see to land. For extremely low visibility conditions, auto-land systems were developed in the 1960s for use when a pilot's vision out-the-windows was almost completely obscured during the landing. However, these auto-land systems cost millions of dollars per airplane, and require millions of dollars in annual maintenance and pilot/crew training costs. Further, only 144 airports are equipped world-wide with expensive landing and lighting systems that enable safe operations at less than 1000 feet visibility.

The value of the EFVS research can be traced to the substantial promise of these head-up display concepts (HUDs, HWDs) to reduce reliance on expensive ground-based landing and lighting systems and significantly increase the number of possible operational runways in use when the weather reduces visibility. Both the HUD and HWDs have been demonstrated to permit low-visibility flight operations in conditions as low as 1000 feet RVR. Recently completed research (December 2014) have extended the HUD EFVS application, using a multi-sensor EVS, to 300 feet RVR approaches. Today, the HUD enjoys operational credit to allow manual approaches (700 feet RVR) and departures (300 feet RVR). Enhanced vision (an EFVS) may further that credit to allow CAT IIIb approaches and departures without need of certified CAT III auto-lands, landing (e.g., CAT III ILS) and lighting systems (e.g., ALSF-2). Further, other "vision technologies", in particular SVS, may complement EFVS to potentially permit EVO to all phases of flight (SVS provides database-based imagery of the flight environment independent of real-time imaging sensors). Taken together, the research may pave the way toward true "all weather" operations and revolutionize future low-visibility operations. Indeed, the EFVS concept may actually best the EVO NextGen idea; and, rather than "equivalent visual operations," may allow instead "better-than-visual-operations" (Bailey, Prinzel, et al., 2011) as the standard for Next and Future Generation Air Transportation Systems.

The path toward “better-than-visual” operations shall require many changes, and there remains significant hurdles to realities. Although EFVS has been certified and today allows manually flown approaches to continue below published DA/DH to a required visual segment at 100 feet HAT and current regulatory efforts likely will permit no visual segment landings to 1000 feet RVR, there are many challenges that remain. These include the quality of the enhanced vision sensor; the weight and costs of these systems; the use of head-down EVS as an EFVS “equivalent display”; to name a few. Further, the transformation requires solution to issues of restricted flight visibility in other operational phases, such as issues of high runway occupancy time and need for expensive surface movement guidance and control systems and surface operational procedures. However, given the tremendous potential of the EFVS and combined vision system (e.g., EFVS + SVS), envisioned applications abound and with continued research and practice, the distinctions between IFR and VFR may become a moot distinction. Examples include operational requirements that exist today to preserve level of safety under instrument meteorological conditions (IMC), such as need for airport alternates and emergency fuel; IFR procedures, such as IMC traffic spacing or precision instrument approaches; or certain avionics, such as auto-land systems, may no longer be necessary. Much work remains but the existing body of work and continued advancement in the technologies evince the tremendous potential capability of these vision-based technologies toward a singular operational concept of “equivalent visual flight rules”.

References

Ames, L.L., & George, E.J. (1993). "Revision and verification of a seven-point workload estimation scale," Air Force Flight Test Center: AFFTC-TIM-93-01.

Arthur, J.J., Prinzel, L.J., Barnes, J.R., Williams, S.P., Jones, D.R., Harrison, S.J., & Bailey, R.E. (2014). Performance comparison between a head-worn display system and a head-up display for low visibility commercial operations. Baltimore, MD: International Society for Optical Engineering (SPIE).

Bailey, R. E., Kramer, L. J., and Williams, S. P. (2010). Enhanced vision for all-weather operations under NextGen. Orlando, FL: International Society for Optical Engineering (SPIE).

Bailey, R.E., Prinzel, L.J., Kramer, L.J., & Young, S.D. (2011). Concept of Operations for Integrated Intelligent Flight Deck Displays and Decision Support Technologies. NASA/TM-2011-217081, Hampton, VA.

Boeing (2013). Statistical summary of commercial jet airplane accidents (1959-2012). Seattle, WA: Boeing.

Federal Aviation Administration (2010). Enhanced flight vision systems. Advisory Circular 90-106. Washington, D.C.: FAA.

Federal Aviation Administration (2013). Operational Requirements for the Use of Enhanced Flight Vision Systems (EFVS) and to Pilot Compartment View Requirements for Vision Systems. Federal Register Number: 2013-13454. Washington, D.C.: FAA.

Federal Aviation Administration (2014). NextGen Implementation Plan. Washington, D.C.: FAA.

Kramer, L.J., Bailey, R.E., Ellis, K.K., Williams, S.P., Arthur, J.J., Prinzel, & Shelton, K.J. (2013). Enhanced flight vision systems and synthetic vision systems for NextGen Approach and Landing Operations. NASA Langley Research Center, NASA/ TP-2013-218054, Hampton, VA.

Radio Technical Commission for Aeronautics (RTCA) Minimum Aviation System Performance Standards (MASPS) for Enhanced Vision Systems, Synthetic Vision Systems, Combined Vision Systems and Enhanced Flight Vision Systems (DO-315A). Issued September 15, 2010. Washington, D.C.: RTCA.

Shelton, K.J., Kramer, L.J,, Ellis, K.K., Rehfeld, S.A. (2012). Synthetic and enhanced vision systems for NextGen (SEVS) simulation and flight test performance evaluation. Williamsburg, VA: Digital Avionics Systems Conference (DASC).

SIMULATOR-BASED ASSESSMENT OF FLIGHT-SPECIFIC APTITUDES IN GERMAN ARMED FORCES' AIRCREW SELECTION

Meierfrankenfeld, Katrin
German Air Force Center for Aerospace Medicine
Fürstenfeldbruck, Germany
Greß, Werner
German Air Force Center for Aerospace Medicine
Fürstenfeldbruck, Germany
Vorbach, Tina
German Air Force Center for Aerospace Medicine
Fürstenfeldbruck, Germany

This paper outlines German Armed Forces' (GAF) approach to predict future success in flight training of applicants for becoming aircrew member. GAF's aircrew selection procedure consists of three phases. Phase I and II include the assessment of basic aptitudes and the aviation-medical examination. Phase III (fixed wing) is more complex. It consists of one week simulator-based screening in a typical training scenario: Candidates prove their skills both in 4 simulator-flight missions with increasing workload and in academic training. As in real flight training, a briefing, a demonstration and a practice phase and subsequent debriefings prepare candidates for their check phases. The aim is to evaluate aptitudes and to propose specific cockpit assignments (e.g. jet pilot, transport pilot, weapon system officer/ navigator) and to minimize attrition rate during basic flight training. GAF's aircrew selection is primarily conducted before applicants decide to join German Armed Forces. The aircrew selection process works quite well, as long term evaluation shows: Attrition rates during flight training are very low (e.g. in ENJJPT: 2007 to 2012: less than 10% total and less than 5% due to flying deficiencies). Approximately 200 applicants are tested at Phase III fixed wing per year.
This paper will describe the flight simulator as well as the scenario including players involved in the screening process and missions in use.

Flight Simulator used in Phase III fixed-wing: A test device

The FPS/F (Aviation Psychological Pilot Selection System/ Fixed Wing) is a flight simulator consisting of 4 cockpits with canopies, a spherical projection dome with 200° horizontal and 45° vertical field of view, a 5-channel high resolution projection system, a multi-functional display with all basic flight instruments plus a master caution panel for malfunctions and a radio panel (Figure 1). The instructor's consoles enable monitoring the applicant's activities and performance. Digital video protocols as well as mission logs are used for debriefing purposes. Data can be analyzed at an evaluation station.

The flight simulator is no training device, but a test device. A generic single seated single engine prop aircraft retractable landing gear is simulated. An automatic trim feature is implemented to ease aircraft control: If a certain flight attitude is set and no flight control inputs are made, the aircraft tries to maintain this attitude. Thereby, complexity is reduced (no trim is

necessary). Furthermore, there is no torque-effect, and the weather is always fine. The aim is to keep the simulation simple enough for the candidates: “Pedestrians” should be able to fly complex missions during one week. This is for sure a time frame too short to learn how to fly an aircraft – or would one assume that a “pedestrian” will get his/her driver’s licence that fast? In addition, missions in Phase III are designed to test how applicants deal with high workload and maneuvers above basic flying capabilities. Therefore, complexity has to be reduced, to avoid floor effects. For screening purposes, standardized missions are used. Mission flow and standardized test conditions are ensured by LUA. With the script language LUA new tasks, missions and evaluation matrixes can be designed.

Figure 1. The flight simulator used in Phase III/ fixed wing (FPS/F) consists of cockpits with a high quality screen comprising the field of view (200° horizontal, 45° vertical) (left), and the multifunctional display showing expanded instrumentation as well as touchscreen and radio (right).

Description of scenarios and players

Scenario: Structure and mission contents

The simulator-based screening takes one week. Six applicants are tested per week. Applicants are pre-selected “pedestrians”, mainly just about to graduate from college with an average age at about 19 to 20 years with no flight experience. Few applicants are active duty soldiers or civilians holding a licence. In Phase III applicants go through four simulator flight missions and two academic sessions. In both fields pace of progress has to be high and written tests are conducted. The combination of studying and learning to fly both at a time is a challenge regarding time management and again reflects demands in real military flight training.

As to the academic part, aerodynamics and navigation are main topics. Proper preparation is required (using a handout that is available at least 2 weeks before applicants arrive in Phase III). Further, the ability to understand and apply brand-new and more complex topics in limited time is tested.

Mission structure is as follows: Mission 1 allows for familiarization with the simulator. Mission 2 consists of traffic pattern procedures from taxiing to full stop landing testing rather procedural skills. Missions 3 and 4 are tactical missions requiring proper information management, fast decision making and task management. Maneuvering, including recoveries from unusual attitudes and trail formation, are elements in Mission 3. Mission 4 is also dynamic, but nevertheless different from Mission 3: A low level navigation route with additional tactical tasks that occur unpredictable for the candidate during the mission task saturates the applicants. Each of the above mentioned missions consists of specific requirements on the one hand and common parts that remain the same from Mission 1 to Mission 4 (e.g., take-off) on the other

hand. The last mentioned allows evaluation of training progress and automation. During each mission, radio transmissions and standard checks are required.

The schedule duplicates real flight training demands in German Armed Forces' flight training, as well as its structure. For example, learning and applying procedures are essential. Contents and demands of each mission are explained in a briefing (conducted by experienced former military Jet Instructor Pilots or Navigators) prior to flying. Afterwards, there is some time left for preparation. Further, each missions consists of three parts: At first, a demonstration phase shows what is going to happen – no action is required. Second, the applicant tries to fly the mission and is being supported by an Instructor Pilot, who is giving helpful advice during the mission and in a short debriefing. Third, the applicants conduct a solo flight – a test flight without help. Each test mission is followed by a debriefing using flight data (including visual system, instruments, maps, audio and video files) to show improvement opportunities.

The aim is to simulate real flight training, and to test the applicants' trainability.

Players in Phase III

Phase III is conducted using an interdisciplinary approach. Each test mission is graded by an Aviation Psychologist and an experienced military Jet Instructor Pilot or Navigator. Two observers grade independently from one another, as means of standardization: Inter-rater-reliability is high in Phase III. A Military Training Staff Officer teaches academics, observes and grades behavior in academic lessons (e.g. team-work, cooperation, participation), and evaluates tests.

At the end of Phase III, there are 3 equal votes (1 Instructor Pilot, 1 Aviation Psychologist, 1 Military Training Staff Officer) to decide if the applicant passes or fails the selection process.

The following section will describe ratings and conditions for passing or failing the selection process in Phase III.

Ratings in Phase III

Expert ratings and the decision process

Experts grade each maneuver, each pattern, as well as each Mission on scales from 1, "excellent", to 7, "unsatisfactory". Behaviorally based rating scales (Standard Operating Procedures) define standards for each and every maneuver and are used to ensure reliable expert ratings. Further, as already mentioned, 2 experts observe each test mission (1 Aviation Psychologist, 1 Instructor Pilot). The simulator also delivers objective criteria (e.g. procedure-evaluation, monitoring of minimum criteria, etc.).

Aptitudes assessed in Phase III are distribution of attention, situational awareness, multi-tasking, aggressiveness, task saturation, concentration, speed of automation, stress resistance, coordination, tolerance towards failure, mission preparation, radio transmissions, training progress and will to perform. Aptitudes are graded for each mission and average values are computed (based on 4 missions). Again, these aptitudes reflect aptitudes assessed during real flight training, easing comparison of results in Phase III with results in real flight training.

According to aptitude profiles, best suitable future flying assignments are determined (that is jet pilot, transport pilot or navigator). The pass/fail decision as well as grading (7-scaled, 1 = excellent and 7 = unsatisfactory) and proposals on cockpit assignments are based on aptitudes, performance and progress during the week and performance in academics. To pass

through Phase III, the average grade (based on performance in 4 missions) must be better than 6 or 7 (on scales from 1, “excellent”, to 7, “unsatisfactory”). Furthermore, no aptitude should be below “5”.

The quality management system is a crucial part of GAF’s aircrew selection system: Each (future) pilot or navigator is graded during the course of his/her flight training several times. Results are reported to Phase III to compare Phase III’s proposals and predictions with training outcomes. Feedback from training squadrons is also used to compute empirically based aptitude profiles, which help to determine assignments for future candidates: Aptitude profiles are computed for each flying assignment. Profiles are based on performance of those pilots or navigators who completed their flight training successfully at an at least average level. Nevertheless, decisions in Phase III do not fully depend on the above mentioned empirical data: Experts also consider training progress, peculiarities of missions, individual weaknesses and strengths in terms of trainability, potential compensation and maturation, to name only a few. Nevertheless, feedback from training squadrons is used to adjust the decision processes in Phase III to changing requirements in military flight training.

Among the applicants who successfully completed Phase III only the best should be selected by the Human Resource Department to join a specific flight training track according to hiring needs. However, the population of successful applicants decreases, e.g. due to demographic changes.

Results

Phase III/ fixed wing replaced flying screening in 1998. The simulator system described here has been established 2008. The main idea is to reduce attrition rates in flight training and training costs. The aircrew selection process works quite well, as long term evaluation shows: Attrition rates during flight training are very low, that is below 10 % due to flying deficiencies. In ENJJPT 2007 to 2012: Less than 10% total and less than 5% due to flying deficiencies. Besides, qualitative analyses show: Predictions in Phase III fit very well with feedback from training squadrons.

Conclusion and Discussion

Although Phase III is working well, there is an ongoing effort to use feedback from training squadrons – have there been changes in flight training; are there possibilities to improve predictions? Besides, there are attempts to establish methodological improvements, for example implementation of further objective criteria, especially for aptitude gradings, and development of new procedure-evaluators.

COGNITIVE ENGINEERING: WHAT'S OLD IS NEW AGAIN

Ronald John Lofaro, PhD
Embry-Riddle Aeronautical University Worldwide, adjunct Associate Professor; (FAA ret.)
Orange Beach, AL 36561

The views and opinions expressed in this paper are solely those of the author. They do not necessarily represent the positions or policies of any private, public or governmental agencies

This paper presents what began as a specific task analysis methodology developed in the context of what then was called knowledge engineering. The resultant model was based on Fleishmann's concept of underlying abilities coupled Delphi techniques and small group dynamics. Core features were the use of small groups of Subject Matter Experts (SMEs) and, a highly structured workshop environment. The model was termed the Small Group Delphi Paradigm (SGDP). As time past, its usage in a variety of aviation venues, ranging from selection to training proficiency, resulted in modifications and refinements. Thus, it became more than just a task analysis even being used, e.g., in identifying civilian managerial and employee core competencies. However, it seemed that, while in the literature multiple times, there was not a measure of general useage. This is not the case today, as will be shown, plus ways to technologically up-date the SGDP.

Knowledge engineering (KE) is defined as follows: "... an engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise." (Feigenbaum and McCorduck, 1983). For a succinct overview of KE see Studer, Benjamins and Fensel (1998). Some of the possible uses and functions of KE: articulation and assessment of an issue/problem; development of a knowledge-based system structure for dealing with issues/problems; obtaining and structuring relevant information and knowledge; developing tests for validation of the obtained information/knowledge. Since the mid-1980's, KE has grown in use and importance concomitant with the advances in computer memory, capabilities and useage. Additionally, KE is often an iterative process with many challenges. Thus, KE can be seen as somewhat more art than engineering. There are no neat boundary lines as to what constitutes KE. Knowledge engineering is also linked to cognitive science and socio-cognitive engineering where the knowledge is produced by socio-cognitive aggregates (mainly humans); this was one rationale for the SGDP. Cognitive engineering (CE) areas include mental workload, decision-making, skilled performance, human-computer interaction, human reliability, work stress and training as these may relate to human-system design. Therefore, CE has mainly replaced KE as the term used in such efforts.

A subset of CE/KE is the Delphi technique/process. Traditional Delphi techniques include anonymity of response, multiple iterations, convergence of the distribution of answers and, a statistical group response (Judd, 1972). A seminal paper on the Delphi process was written by a then-Rand Corporation employee (Brown, 1968) and may be available from Rand or from American Society of Tool and Manufacturing Engineers (ASTME), now known as Society of Manufacturing Engineers.

A modification to Delphi processes is the small group Delpi paradigm (SGDP). The SGDP took the Delphi process in another direction by modifying it via merger with elements of group dynamics in order to have interactive (face-to-face) Delphi workshops. This modification resulted in a paradigm for using small groups of subject matter experts (SMEs). The SGDP can be used for any project that requires that a set of SMEs be used to identify, evaluate, and criticality rank tasks (an enhanced task analysis), identify core needs/skills, recommend modifications to equipment, procedures and training. Finally, the SGDP can be used to sharpen, modify and revise existing methodologies. As Meister (1985) had noted, “The (Delphi) methodology is by no means fixed…[it] is still evolving and being researched." This is as true

now as it was when Meister stated it. In point of fact, with the leaps in communication methods and related technology, even more so.

The Initial SGDP

The development of this modified Delphi, the SGDP, involved the merger of a specific knowledge engineering technique (Delphi), with Fleischmann's theories of underlying abilities (Fleishmann and Quaintance, 1984; revised 2000) and some principles of group dynamics. It was the result of a specific issue and difficult problem: to provide US Army Aviation Command with a unified aviator candidate selection test that also indicated which of the current rotorcraft would be the optimum operational aircraft for the candidate upon completion of initial training. The SGDP methodology was used in four workshops-one for each of the then-operational U.S. Army rotorcraft. These workshops had small groups of aviator SMEs, carefully selected and brought in from both the continental United States (CONUS) and overseas Army bases, in a highly structured set of face-to-face workshops sessions.

A major consideration in the SGDP design was the possible negative impact of using face-to-face groups for ratings and evaluations. Pill (1970) said that this may dilute the opinions of the real expert. This seemed a strange objection as the subject matter experts (SMEs) selected ARE the real experts. However, if what is meant is that one or more persons in a group may have more expertise in a specific area that is being worked on and the group would defer to them, then the reality (based on the author's conducting seven or so of these) is that the other SMEs recognize, welcome and use that expertise–as their goal is the best result/product possible.

Another objection is that the group dynamics may force ratings and analyses towards a mean or middle ground that does not fully reflect all the SME's views. There are two responses to this: the first is that true SMEs will not allow that to happen because they see themselves (and, are) THE experts.They want the SGDP products to demonstrate that expertise. Pride will forego them from "going along to get along." The second is that the instruction in group work, the trained facilitator and the iterative methodology used in accomplishing the sub-objectives/objectives are all structures in the SGDP process designed to ensure that this does **not** happen.

Therefore, at that time, the use of small groups of SMEs in a non-anonymous Delphi setting seemed to the author to offer strong points and benefits. Thereupon, this paradigm was first used in the development and fielding of a computerized test battery and algorithm that would both select U.S. Army aviator candidates for initial training and, indicate which of the types of operational helicopter they should go into for transition training. The U.S. Army Aviation Center successfully used this test battery and set of algorithms ("Multitrack") in selecting its rotorcraft aviator candidates for over 5 years.(Lofaro and Intano,1989; Lofaro, Intano and Howse,1990; Intano, Howse and Lofaro, 1991). As the author was told by United States Coast Guard (USCG) pilots who had come over from Army Aviation, while he was teaching for Embry-Riddle Aeronautical University (circa 2008), that some? all? of Multitrack was again in use by the Army.

In sum: In first developing and using the SGDP, circa 1985, (Lofaro, 1992a), these aspects of the traditional Delphi were maintained: specified objectives; iterative process; SMEs; consensus. Added to these were the use of small face-to-face groups and group dynamics training/exercises, a large read-ahead package for each SGDP participant, the use of a facilitator, strict protocols for the participants and sessions, as well as a sequential, step-wise plan of attack on the sub-objectives and objectives. Thus, traditional Delphi processes were modified into a new paradigm for small-group projects

The SDGP: Over The Years

This initial effort and the subsequent use of the resultant paradigm, SGDP, in other and varied venues have produced both highly accurate data (that were operationally implemented) and modifications to the SGDP. Every use of the basic SGDP model resulted in some modifications as the objectives are defined, the SMEs are selected and time limits are set. These many SGDP efforts resulted in sharpening,

modifying and revising the original methodology. Over the years, the SGDP...and revisions... have produced eight operational products in aviation, such as air traffic controller (ATC) and x-ray baggage screener selection tests, training criterion for rotocraft maneuver proficiency, crew resource management (CRM) performance evaluation, task analyses and, for the FAA, sets of highly specific managerial/employee core competencies. (e.g., Lofaro,1999; Lofaro, 1998; Lofaro, Gibb and Garland, 1994; Gibb and Lofaro, 1994; Gibb, Lofaro, et al. 1993; Lofaro 1992b). The SGDP has been used in many environments demonstrating a robust flexibility and generalizability of the paradigm. The extensions of the paradigm indicate that it has an applicability over many domains. For a fairly detailed exposition of each aspect of the SGDP, see Lofaro, R.J. and Maliko-Abraham, Helene.
Of particular note is that, circa 2009, the use of face-to-face groups in a Delphi has now become accepted. This is called the Mini-Delphi/Estimate-Talk-Estimate (ETE) with many variations. Some twenty-five years after the SGDP was devised, used and appeared in multiple publications, it has been re-discovered, as it were. More on this later.

The SGDP In 2014 and Beyond

On a personal level, the author is heartened to see his seminal concepts (the coupling of traditional Delphi methods with group dynamics and face-to-face sessions) seem to have become accepted, questions arose: What now? What are some current 2014 aviation issues (as well as prior but unresolved ones) which are both important and amenable to some form of the SGDP? What can modern technolgy offer in 2014 and beyond to the SGDP and, vice-versa? Some current issues,as well as some in the past that seem to re-emerge, are as follows:
Aircarrier upset training, training that the Colgan Air accident brought to the fore. In direct response to the Colgan crash, Congress passed the Airline Safety and FAA Extension Act of 2010, which mandated that the Federal Aviation Administration require pilots to complete 1,500 flight hours before they're allowed to fly commercially, up from just 250 hours before the act. While this new rule may do little to improve safety, it is exacerbating an already severe pilot shortage. Too few pilots are now available to replace the ones who are retiring. The pilot shortage is beginning even faster than expected. In that context, the new 1,500-flight-hour requirement is a particular problem. Both pilots involved in the Colgan crash had far surpassed 1,500 hours of flight time, so that requirment probably had little to no impact on the accident.
A historically low number of people are training to become pilots and, of those, only half are seeking a career with commercial airlines. For many would-be pilots, a main consideration is financial: while flight training costs between $60,000 and $70,000, entry-level pilot positions typically pay $25,000 a year or less. Furthermore, the financial turbulence that has plagued the airline industry since September 11, 2001, makes the profession somewhat less attractive to aspiring aviators. The existing workforce has been stretched even thinner by new anti-fatigue rules. Pilots were once required to have eight hours of time off between shifts; but now they must be given no less than ten hours. This particular anti-fatigue rule (see below) was empirically justifiable and it may well improve safety, but it also results in airlines' needing between 3 and 7 percent more pilots available ("on the clock") at any given time.
The FAA, while not yet issuing an Advisory Circular (AC) or a federal aviation regulation (FAR), has issued a document called Airline Upset Recovery Training Aid, version 2. The issues seem to be use of a full motion flight simulator (FS) that will be part of an expected FAA pilot training rule by 2018 (Croft, John. 2014a); in-aircraft training and, swept wing jet aircraft specialized training . The American Airlines UPRT ground school with FS training, called advance aircraft manuevering program (AAMP), was seen by The National Transportation Board (NTSB) as possibly a contributing factor in the American Airlines flight 587/A300-600 crash in November, 2001 (Croft, John. 2014b).
We now return to that long-time and often researched area: crew fatigue. It surfaced again with the United Parcel Service (UPS) flight 1354 crash in Birmingham, AL in 2013. At this time, it seems that the UPS pilots, NTSB and Airbus (the aircraft involved was an Airbus A300-600) are in "disagreement." Remember that UPS has an FAA-managed fatigue risk management program (Croft, John. 2014c).

If memory serves (author was with FAA from 1989 into 2004), the FAA was involved in a NASA/United AirLines study about long-haul/TransPac flight and sleep/rest. As one result, in 1991, the FAA proposed a draft AC called Controlled Rest on the Flight Deck, which was opposed by industry. Here we are in 2014 and aviation is still working this issue while still more lives have been lost. Admittedly, the issues cited above all come from a small sampling of Aviation Week & Space Technology magazines (AW&ST), but the attempt has been made to select both current and somewhat safety oriented problems. (*Full disclosure*: while writing this paper, the author had plans to show the "how" of a revised SGDP in dealing with two (2) of the above aviation problems. He soon discovered that the page limitation made this an impossibility. He will submit a second paper with more detail on the structure of the SGDP, possible structures for a revised SGDP that incorporates technological advances and, the procedures for addressing at least two of the current aviation problems indicated in this paper. It is hoped that the second paper will, as well as this one, be accepted and possibly form a basis for a 2015 ISAP Symposium session. However, the second paper will be self-contained and will not rely, for comprehension, on a reading of this paper.)

SGDP: Its Time Has Come Again

The Delphi is based on the principle that forecasts (or decisions) from a structured group of individuals are more accurate than those from unstructured groups. As has been said, the use of face-to-face Delphi techniques has been re-discovered. New technologies have resulted in what are generically referred to as mini-Delphi or Estimate-Talk-Estimate (ETE). Other innovations come from the use of computer-based (and later web-based) Delphi conferences. One example of a difference in a type of ETE (a computer-based Delphi) versus either a tradtional or SGDP Delphi is the iteration structure used in the tradtional or SGDP Delphis, which is divided into three or more discrete rounds, can be replaced by a process of continuous (roundless) interaction, enabling SMEs to change their evaluations at any time. In view of technological advances, it is posited that the SGDP structure and processes are still relevant but need integration with ECE. It is further posited that this revision of the SGDP will produce the same level, if not a higher level, of accurate information and products in the aviation arena.

Integrating SGDP With ETE

Here is a brief review of the core structure of a SGDP with indications of where aspects of an ETE can be used. The core on a SGDP is: careful selection of a limited number of SME; the use of an extensive read-ahead package for the SMEs; the use of some facilitation and group dynamics instruction, combined with some type face-to-face sessions. A new ETE/SGDP model would be computer-based; the reader is referred to the work of Turoff and Hiltz (1996) on computer-based Delphis. Integration of the SGDP with a ETE approach can be achieved thusly: all participants can be logged on simultaneously, each participant can briefly state their name and credentials, the group dynamics instruction can be done by the facititator to all simultaneously (aside: it would seem that a linked network of all SMEs is possible and even *de rigueur*. This will allow for instanteous feedback by any SME during a session, as well as discussions). The iteration structure used in SGDP, which is divided into as many discrete rounds as needed for consensus, can be replaced by a process of continuous (roundless) interaction. This will enable participants/SMEs to change their evaluations at any time and give a rationale with ensuing discussion in real-time. Finally, the statistical group response can be updated in real-time and shown whenever a SME or a group provides a new evaluation.

It is clear that "face-to-face" discussion will be virtual. This is both a real and significant loss. However, the speed, multiple iterations, real-time and other aspects to be gained cannot be ignored. Another possible modification is a multi-tiered SGDP/ETE in which the use of two or more SGDP/ETE groups with different issues/expertise can be convened and given objectives based on these issues/expertise. As these groups come to consensus on their objectives, these new data can be integrated, built into a new re-ahead package and made available to a new SGDP/ETE set with new or prior SMEs.

What Is Next?

Future research can revolve around comparison of the accuracy of results using a traditional Delphi, the SGDP, various ETE Delphis (computer and/or web-based). A recent Delphi technique is a web-based communication structure involving a large number of participants.These web-based variable communication structures are designed to make Delphi efforts more fluid and adapted to the hypertextual and interactive nature of digital communication. As above, comparisons can be made among various ECE results and those of other Delphi techniques. Finally, new and perhaps blended Delphi techniques may emerge from such research and comparisions.

References

Brown, B.B. (1968). Delphi Process: A Methodology for the Elicitation of Opinions of Experts. Santa Monica, CA: The Rand Corporation.(possible:ASMTE Vectors, February, 1968).

Croft, John. (2014a). Unambigous Upset. *Aviation Week & Space Technology*, Vol.176, No.20, 52.

Croft, John. (2014b). Unambigous Upset. *Aviation Week & Space Technology*, Vol.176, No.20, 53.

Croft, John. (2014c). Rest Assured. *Aviation Week & Space Technology*, Vol.176, No.18, 38.

Feigenbaum, Edward A, McCorduck, Pamela (1983). *The Fifth Generation (1st ed.).* Reading, MA: Addison-Wesley.

Fleishman, Edwin A. and Quaintance, Marilyn K. (1984, reissued 2000). *Taxonomies of human performance: The description of human tasks.* Management Research Institute, Inc.: Potomac, Maryland.

Gibb, G.D., Lofaro, R.J. et al. (1993). Computer-based Assessment of Air Traffic Controller Abilities: Experimental Selection Tasks in *Proceedings of 1993 Aerospace Medical Association Annual Symposium.* Alexandria, VA.

Gibb, G.M., Lofaro, R.J. (1994). Airport Security Screener and Checkpoint Security Supervisor Training Workshop. DOT/FAA/CT-94/109. National Technical Information Service. Springfield, VA.

Intano, G.P., Howse, W.R., and Lofaro, R.J. (1991). The Selection of an Experimental Test Battery for Aviator Cognitive, Psychomotor Abilities and Personal Traits. U.S. Army Institute for the Behavioral and Social Sciences RN 91-21: Alexandria, VA.

Judd, R.L. (1972). Use of Delphi Methods in *Higher Education in Technological Forecasting and Social Change, 4.* NV: Reed Elsevier

Lofaro, R.J., Gibb, G.M., and Garland, D.(1994). A Protocol for Selecting Airline Passenger Baggage Screeners. DOT/FAA/CT-94/110. Springfield, VA: National Technical Information Service.

Lofaro, R.J. and Intano, G.P. (1989). Exploratory Research and Development: Army Aviator Candidate Classification By Specific Helicopter in *Proceedings of 5th International Symposium of Aviation Psychologists.* ISAP: Colunbus, OH.

Lofaro, R.J., Intano, G.P., and Howse, W.R. (1990). Final Validation of the Army Aviator Multitrack Algorithm in *Proceedings of the 1990 Annual Military Testing Association Conference.*

Lofaro, R.J. and Maliko-Abraham, Helene, (2002). Techniques for Developing High Predictive Validity Selection and Screening Test Batteries in *Proceedings* of the 2002 World Aviation Congress; Phoenix, AZ.

Lofaro, R.J. (1992a). A Small Group Delphi Paradigm in *The Human Factors Society Bulletin, 35 (2),* 1-4. Santa Monica, CA.

Lofaro, R.J. (1992b). Workshop on Integrated Crew Resource Management (CRM). DOT/FAA/RD-95/5. National Technical Information Service: Springfield, VA.

Lofaro, R.J., (1998). Identifying and Developing Managerial and Technical Competencies: Differing Methods and Differing Results in *Proceedings of 42nd annual Human Factors and Ergonomics Society:*

Santa Monica, CA.

Lofaro, R.J.,(1999). Core Competencies and Individual Differences: Chasm or Linkage? in *Human Factors and Ergonomics Society/HF&ES Individual Differences in Performance News, 12 (2).* Santa Monica, CA.

Meister, David (1985). *Behavioral Analysis and Measurement.* NYC: Wiley Press.

Murray Turoff, Starr Roxanne Hiltz (1996). Computer-based Delphi processes in Michael Adler, Erio Ziglio (eds.). *Gazing Into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health,* 56-88. London: Kingsley Publishers.

Pill, J. (1970) The Delphi Method: Substance, Context: Critique and an Annotated Bibliography. *Case Western Reserve University Technical Memorandum 183*: Cleveland, OH.

Studer, Rudi, Benjamins, V. and Fensel, Dieter (1993). Knowledge Engineering, Principles and Methods in *Data & Knowledge Engineering 25,* 161-197. North Holland: Elsevier.