

EMBRY-RIDDLE
Aeronautical University™
SCHOLARLY COMMONS

Dissertations and Theses

Spring 2011

Effects of Online Training on Aircrew Monitoring Behaviors: A Field Study

Brian A. Potter
Embry-Riddle Aeronautical University - Daytona Beach

Follow this and additional works at: <https://commons.erau.edu/edt>



Part of the [Aviation Safety and Security Commons](#), and the [Higher Education Commons](#)

Scholarly Commons Citation

Potter, Brian A., "Effects of Online Training on Aircrew Monitoring Behaviors: A Field Study" (2011).
Dissertations and Theses. 117.
<https://commons.erau.edu/edt/117>

This Thesis - Open Access is brought to you for free and open access by Scholarly Commons. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

EFFECTS OF ONLINE TRAINING ON AIRCREW
MONITORING BEHAVIORS: A FIELD STUDY

by

Brian A. Potter
B.A., University of Delaware, 1999

A Thesis Submitted to the
Department of Human Factors and Systems
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Human Factors and Systems

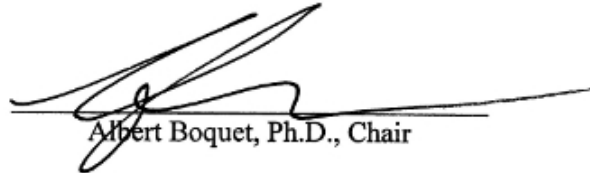
Embry-Riddle Aeronautical University
Daytona Beach, FL
Spring 2011

EFFECTS OF ONLINE TRAINING ON AIRCREW
MONITORING BEHAVIORS: A FIELD STUDY

By: Brian A. Potter

This thesis was prepared under the direction of the candidate's thesis committee chair, Dr. Albert Boquet, Ph.D., Department of Human Factors and Systems, and has been approved by members of the thesis committee. It was submitted to the Department of Human Factors and Systems and has been accepted in partial fulfillment of the requirements for the degree of Master of Science in Human Factors & Systems.

THESIS COMMITTEE:



Albert Boquet, Ph.D., Chair



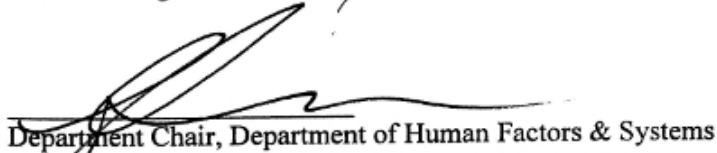
Elizabeth Blickensderfer, Ph.D., Member



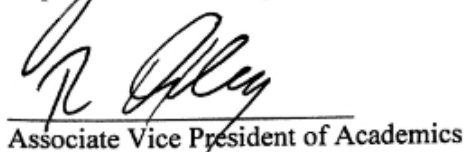
Nickolas D. Macchiarella, Ph.D., Member



MS HFS Program Coordinator



Department Chair, Department of Human Factors & Systems



Associate Vice President of Academics

Abstract

Data from aircraft accidents and line observation studies indicate that inadequate pilot monitoring is a growing safety concern. In the cockpit environment, pilots who fail to properly manage their workload commit more monitoring errors. Given the lack of training and educational programs available to pilots which emphasize improving their monitoring skills, more research is needed to assess the usefulness of types of training that can be used to improve pilots' monitoring. This research project sets out to determine if the potential exists to enhance pilots' monitoring skills through online training. For this study, 40 military helicopter pilots (participants) were divided into two training groups: (1) an online training group, which completed a 20-minute web-based training module, and (2) a control group, which read a 20-minute article on aviation safety. Within each group, the pilots were paired and completed two training events in a flight simulator as part of their normal duties. The effects of the training were evaluated using Kirkpatrick's multi-level framework: reactions, learning, behaviors, and results. First, all pilots receiving training were surveyed to capture the trainees' perceptions of satisfaction and utility of the training. Second, all pilots were given a multiple-choice test to assess the effect the training had on learners' knowledge of the training objectives. Next, the researchers observed, via video recording, both groups' behaviors during flight simulator events. The researchers recorded occurrences of four behavior markers as the crews flew multiple instrument approaches. The researchers used two of these markers to study transfer of training and two markers to examine positive vs. negative outcomes at critical tasks during the simulated flights. The results show positive changes in the reactions, learning, and behavior dimensions, lending support to the effectiveness of relatively inexpensive online training to teach monitoring skills.

Table of Contents

Abstract	iii
List of Tables	vi
List of Figures	vii
List of Abbreviations	viii
Introduction.....	1
Aircrew Monitoring	1
Psychology of Monitoring	3
Standard Operating Procedures.....	10
Phase of Flight	12
Monitoring and Assertiveness.....	13
Evaluating Monitoring Performance.....	17
Training Pilots to Monitor	18
e-Learning.....	23
Purpose of this Study	25
Statement of Hypotheses.....	25
Methods.....	28
Design	28
Participants.....	28
Apparatus	31
Informed consent	31
Demographic survey	31
Online training	32
Control group activity.....	34
Questionnaire	34
Post-training test	35
Reactions questionnaire	35
Attitudes questionnaire	35
Flight simulator and debriefing station	36
Flight scenarios	38
Debriefing Script.....	41
Procedure	42
Measures	45
Independent variable – training	46
Covariate – pilot experience	46
Dependent variables.....	47
Reactions scores.....	48
Post-training test scores	49
5-2-1 Compliance.....	49
Callout ratio	50
Task success rate.....	51
Overshoot.....	53
Results.....	55
Reaction outcomes	55

Learning outcomes.....	57
Behavior outcomes.....	58
5-2-1 rule compliance.....	58
Callout ratio.....	59
Task outcomes.....	60
Pilot experience.....	61
Attitudes questionnaire.....	62
Training Latency Analysis.....	64
Discussion.....	65
Overview.....	65
Hypotheses 1-2.....	65
Hypotheses 3-4.....	66
Hypothesis 5.....	70
Hypothesis 6.....	72
Helicopter “Areas of Vulnerability”.....	73
Limitations.....	75
Selection bias.....	75
Efficacy of online training.....	76
Training latency.....	77
Incomplete data.....	78
Instruction differences.....	79
Experimenter/measurement bias.....	79
Conclusion.....	80
Recommendations.....	81
References.....	83
Appendix A: Informed Consent.....	93
Appendix B: Demographic Questionnaire.....	94
Appendix C: Monitoring and Cross-checking Online Training.....	95
Appendix D: Post-training Test.....	103
Appendix E: Reactions Questionnaire.....	104
Appendix F: Attitudes Questionnaire.....	105
Appendix G: Participant Debriefing.....	106
Appendix H: Sample Rater Grade Sheet.....	107

List of Tables

Table 1. Pilot Experience.....	29
Table 2. Demographics for Participants in Flight Simulator	30
Table 3. USCG Training Schedule	31
Table 4. Instrument Approaches Flown by IFR Scenario.....	41
Table 5. Procedure	44
Table 6. Independent Variable and Covariates	47
Table 7. Dependent Variables.....	48
Table 8. Tasks-of-interest	54
Table 9. Post-training Reaction Scores	57
Table 10. Post-training Test Results	58
Table 11. Correlations Between Four In-flight Variables	60
Table 12. Monitoring Outcomes in Flight	61
Table 13. Callouts by Pilot Experience.....	62
Table 14. Attitudes Questionnaire Results.....	63

List of Figures

<i>Figure 1.</i> Sumwalt et al.'s (2002) Areas of Vulnerability in Commercial Aviation	13
<i>Figure 2.</i> USCG H-65 Helicopter Flight Simulator.....	37
<i>Figure 3.</i> CADS Replay Software Screen	38
<i>Figure 4.</i> Participant Attitudes towards Mandatory Monitoring Altitudes.....	64

List of Abbreviations

AC	Aircraft commander
ALAR	Approach and landing accident reduction
ASRS	Aviation Safety Reporting Service (asrs.arc.nasa.gov)
ATC	Air traffic control
CAD	Computer-aided debrief station
CBT	Computer-based training
CFIT	Controlled flight into terrain
CP	Copilot
e-Learning	Electronic learning
FAA	Federal Aviation Administration
FSF	Flight Safety Foundation (www.flightsafety.org)
IAP	Instrument Approach Procedure
ICAO	International Civil Aviation Organization
IFR	Instrument flight rules
IP	Instructor pilot
LOS	Line Operational Simulation
LOSA	Line Operations Safety Audit
NTSB	National Transportation Safety Board
PF	Flying pilot (pilot at the flight controls)
PIC	Pilot-in-command
PM	Monitoring pilot (synonymous with PNF and SP)
PNF	Non-flying pilot (synonymous with PM and SP)
SOP	Standard operating procedure
SP	Safety pilot (synonymous with PNF and PM)
USCG	U. S. Coast Guard
WM	Working memory

Introduction

In a typical two-pilot aircrew, the flying pilot's (PF) primary responsibility is to control the aircraft by physically manipulating the flight controls. The other pilot, referred to as the pilot not flying (PNF), conducts supporting responsibilities including navigation, communication, checklists, and monitoring the actions of the flying pilot (Tullo, 2010). These in-flight responsibilities are not indicative of authority (i.e., who is designated as the "captain" or "pilot in charge") as the PF and PNF roles may be alternated throughout the flight as needed (Dismukes & Berman, 2010; Prince, Salas, Brannick, & Prince, 2010). Recently, there has been a paradigm shift within the aviation industry regarding the role of the PNF. Particularly in the airlines, the preferred term for this role has changed from *non-flying pilot* into *monitoring pilot* (PM) to impress upon flight crews that (a) monitoring is the PNF's primary responsibility, and (b) the PNF has an active responsibility to detect errors in the PF's performance (Sumwalt, Thomas, & Dismukes, 2002; Orlady, 2010). Improving the PNF's performance in the monitoring role is the focus of this study.

Aircrew Monitoring

Data from accident reports and flight operation audits all confirm that errors in monitoring and cross-checking can be highly consequential. For example, a National Transportation Safety Board (NTSB) study found 84 percent of crew-caused air carrier accidents in the U.S. could be attributed to inadequate crew monitoring or challenging (NTSB, 1994). A Flight Safety Foundation Approach and Landing Accident Reduction (ALAR) study of 76 approach and landing accidents worldwide found that 63 percent were attributable to "poor monitoring and cross-checking" (Khatwa & Helmreich, 1998). Accident reports also indicate that breakdowns in monitoring are a factor in half of all controlled flight into terrain accidents

(International Civil Aviation Organization [ICAO], 1994). In 2010, the NTSB issued a safety recommendation (A-07-13) in response to the February 2009 Colgan Air Flight 3407 crash near Buffalo, NY. The two-pilot aircrew inadvertently allowed the aircraft to decelerate into an unrecoverable stall during an approach and subsequently crashed killing all 49 persons on board (NTSB, 2010). The NTSB recommendation required that “all pilot training programs be modified to contain modules that teach and emphasize monitoring skills and workload management and include opportunities to practice and demonstrate proficiency in these areas” (NTSB, 2010, p. 59). This was the second time a NTSB report from a multiple-fatality accident referred to this recommendation (NTSB, 2007; NTSB, 2010). The NTSB investigation of a 2005 crash of a Cessna Citation 560 in Pueblo, Colorado, had many similarities to that of Colgan Air Flight 3407. While conducting an instrument approach, the aircraft began to accumulate ice on its wings. The pilots initiated an airspeed adjustment as they began to diagnose the severity of the icing and decide on a course of action. Meanwhile, the pilots unknowingly allowed the aircraft to decelerate beyond safe airspeed limits into a stalled condition, crashing about four miles short of the runway (NTSB, 2007).

These reports and others have, in part, prompted several airlines to change their standard operating procedures (SOPs) to acknowledge monitoring as a primary and shared responsibility of both crewmembers (Sumwalt et al., 2002; Dismukes & Berman, 2010; Tullo, 2010). In 2003, the FAA published an advisory circular titled “Standard Operating Procedures for Flight Deck Crewmembers” with the aforementioned recommendation to aircraft operators to change the term “non flying pilot” in their SOPs to “monitoring pilot” due to government and industry preference (Federal Aviation Administration, 2003, p. 1). Further, the FAA (2004a) reinforced a similar sentiment in writing that:

Effective monitoring and cross-checking can be the last line of defense that prevents an accident because detecting an error or unsafe situation may break the chain of events leading to an accident. This monitoring function is always essential, and particularly so during approach and landing when controlled flight into terrain (CFIT) accidents are most common. (p. 14)

As Sumwalt (1999) explains, monitoring the situation within the cockpit environment is the first step towards recognizing deviations, which leads to challenging potential errors and ultimately increasing safety margins.

Psychology of Monitoring

Monitoring is a paradigm rooted in human cognition. Monitoring is defined as “to watch, keep track of, or check usually for a special purpose” according to www.merriam-webster.com (January 12, 2010). In aviation terms, monitoring refers to the “responsibility of pilots to keep track of the aircraft’s position, course, and configuration; the status of the aircraft’s systems and the actions of the other pilots in the cockpit” (Dismukes & Berman, 2010, p. 7). Sumwalt, Morrison, Watson, and Taube (1997) defined monitoring error as:

A failure to adequately watch, observe, keep track of, or cross check any or all of the following: (1) the aircraft’s trajectory, i.e., taxi and flight path, speed management, navigation; (2) automation systems and mode status, i.e., flight management system (FMS) entries, mode control panel (MCP) settings/ selections, awareness of automation mode; and (3) aircraft systems and components, i.e., fuel quantity, aircraft configuration, system status. (p. 1)

Monitoring then, by definition, requires pilots to pay attention to multiple concurrent tasks which, in turn, make them susceptible to errors simply because humans lack an infinite

capability to manage large quantities of tasks. Research indicates these errors are most commonly experienced in terms of one forgetting to perform an intended task (i.e., making an error of omission) (Sarter & Alexander, 2000; Dismukes, Loukopoulos, & Jobe, 2001; Dismukes, 2007). For a pilot to monitor both the aircraft and other crew members' actions they must mentally sort an assortment of tasks through a process known as *interleaving*. This process creates smaller cognitive tasks (mental placeholders) just to keep the progress of the tasks organized and then to later retrieve them, which leaves pilots increasingly vulnerable to making these types of errors (Loukopoulos, Dismukes, & Barshi, 2009). One study examined NASA's Aviation Safety Reporting System (ASRS) database to analyze self-reported safety incidents that occurred as a result of concurrent task demands for the crew's attention (Dismukes, Young, & Sumwalt, 1998). The study found that 69 percent of the mishandled activities involved a shortcoming in some type of monitoring responsibility (Dismukes et al., 1998). Further, the researchers surveyed a sample of pilots to identify existing techniques the pilots used to manage concurrent tasks and reduce their personal vulnerability to monitoring lapses. The techniques the participants suggested were consistent with a human cognition-based perspective – techniques such as “creating salient reminder cues, breaking concurrent tasks into subtasks, and identifying specific things to monitor” were the most common responses (Dismukes et al., 2001, p. 2). In a broader context, a study by Brandimonte and Passolunghi (1994) found that cues serving to remind individuals to complete a deferred task are more effective when (a) they are conspicuous, (b) they are strongly associated with the deferred task, and (c) they are positioned in a way the individual is likely to notice them at the appropriate time. Loukopoulos et al. (2009) explain the strategy for making these mental reminders is like that of creating a mental “sticky note” (p. 125).

The discussion of concurrent tasks reflects the profound influence of workload on monitoring performance (NTSB, 2010). Although definitions vary, workload refers to the mental and physical task demands on an individual at a particular time – a load that varies based on the environment, experience, and mental state of the crew member (Wickens, Gordon, & Liu, 2004; Orlady & Orlady, 1999). Studies in human performance have indicated that humans do not perform optimally under either extremely high or low workload conditions. Under low levels of workload, crews' vigilance tends to decrease and, especially if complacency sets in, performance deteriorates. The concern here is that monitoring a complex system for low frequency events for long periods of time requires a significant amount of vigilance, while other, less critical task demands in the cockpit may be providing pilots more salient cues (such as 3-D moving map displays). Dismukes and Berman (2010) explain that "humans are inherently poor at monitoring for infrequent events" (p. 24). By comparison, under high levels of workload, aircrew performance losses are also likely, but for different reasons. These situations force crews to prioritize tasks and diminish their cognitive resources available for other tasks like monitoring (Thomas & Petrilli, 2006). High workload situations often tunnel the attention of the PNF towards other tasks (Kontogiannis & Malakis, 2009). High workload conditions, such as the pre-takeoff taxi, and the approach-to-landing phase of flight, are generally where pilots make consequential monitoring errors (NTSB, 1994; Sumwalt et al., 1997; Sumwalt, 1999). Several commercial aviation landing accidents have occurred after the non-flying pilots failed to challenge or abort dangerous approaches under high workloads despite cockpit indications that strongly suggested such a course of action (Dismukes, Berman, & Loukopoulos, 2007).

Human factors research has contributed to our understanding of monitoring performance mainly in the context of pilots' reliability in detecting and managing errors in the cockpit

(Sumwalt, 1999; Sumwalt et al., 2002; Thomas, 2005). Sumwalt (1999) explained that effective monitoring is the first step towards recognizing deviations (i.e., errors), which in turn leads to pilots' challenging them and ultimately increasing safety margins in flight deck environments. An observational study of line operations in a simulator found that even experienced aircrews fail to catch nearly half of all errors committed in the cockpit, and went on to identify "vigilance" and "monitoring and cross-check" as the top two (of twelve) non-technical performance markers indicating whether threats and errors were effectively managed by the aircrews (Thomas, 2004). Vigilance and attention are key components of monitoring (Tullo, 2010) and have been identified as significant predictors of effective error management in flight operations, particularly during the vertical phases of flight (Thomas, 2004). Although the terms monitoring and vigilance are sometimes used interchangeably, they do have differences. Vigilance tasks generally refer to tasks that monitor exclusively for specific targets or deviations and require that the task is kept in the conscious, short-term memory for quicker retrieval (Brandimonte, Ferrante, Feresin, & Delbello, 2001; Guynn, 2008). Monitoring, in contrast, is the process of sharing cognitive resources between various types of tasks, including but not entirely focused on target detection tasks (Guynn, 2008). This suggests that monitoring is a better description of the larger mental processes (e.g., flying, system cross-checks, and navigation) often required of pilots in the cockpit environment.

The term "threat and error management" rose to prominence within aviation safety vernacular in the late 1990s as a result of extensive line operations studies into team performance and crew resource management behaviors as part of the University of Texas' Human Factors Research Project (Helmreich, Klinect, & Wilhelm, 1999). This research greatly advanced the notion that non-technical skill training is needed to help trap errors that infiltrate the cockpit or

flight deck environment (Sumwalt, 1999). Non-technical skills are defined as “the cognitive and social skills of flight crew members in the cockpit, not directly related to aircraft control, system management, and standard operating procedures” (Flin et al., 2003, p. 96). Among these non-technical skills, monitoring is considered a skill central to successful error management (Sumwalt, 1999; Sumwalt et al., 2002; Thomas, 2005; Dismukes & Berman, 2010). Thomas (2005) described monitoring as a group of cognitive skill dimensions “concerned with constantly watching and scrutinizing the operating state of the system” to manage errors (p. 19). The first of the skill dimensions is self-monitoring, which describes the metacognitive process of “maintaining an awareness of one’s own state” (Thomas, 2005, p. 19). This includes the ability to detect degradation in one’s own performance, such as is the case when one is overtasked or distracted. The second monitoring skill dimension Thomas describes is “scan and check”, a skill requiring conscious analysis of the status of the system. In the flight environment, scan and check requires special vigilance, as monitoring for a low-frequency, low-signal-to-noise ratio type of event is tedious. Many factors, including experience and planning, factor into how successful or vigilant the person is at scan and check performance. According to Thomas, the third dimension, divergence detection, is simply the act of identifying deviations from the plan within the system. This aspect of monitoring is dependent on one’s mental model and how the real, current situation is compared against the ideal models. The success of monitoring is directly proportional to one’s attention and vigilance, and plays into an individual’s overall situation awareness (SA).

The link between monitoring and situation awareness is expressed throughout research (Endsley, 1995; Jones & Endsley, 1995; Thomas, 2005; Fioratou, Flin, Glavin, & Patey, 2010). Situation awareness consists of three levels, (a) perceiving elements within the environment, (b)

comprehending their meaning, and (c) projecting their status in the near future (Endsley, 1988). Jones and Endsley (1995) developed and applied a “situation awareness error taxonomy” to identify causal factors leading to SA errors in the cockpit. The most frequent error causal factor was a “failure to monitor” which accounted for 37 percent of the errors. Failure to monitor was categorized as a Level 1 SA (i.e., perception) error and was defined as an instance when the information was directly available to the pilots, but for various reasons, was not observed or included in the crew member’s scan pattern (Endsley, 1995). Level 2 SA is important as it describes the role that experience and mental models have on monitoring performance. Given the large and complex amount of data that a pilot must monitor for potential changes, one’s ability to detect a potential error is directly related to how developed and advanced their mental model is (Kontogiannis & Malakis, 2009). A highly developed mental model defines an individual’s expectations in terms of how quickly they match what they are seeing with what they expect to see (Endsley, 1995). This research suggests flight experience should improve a pilot’s monitoring skills.

The use of “prospective memory” is another cognitive performance dimension which impacts aircrew monitoring. Prospective memory refers to a person’s ability to perform an intended action at an appropriate moment in the future without a direct reminder to do so (Ellis, 1996). For example, when an air traffic controller tells a pilot to descend to a certain altitude, the pilot immediately starts the descent, but must use prospective memory processes to retrieve the required action (leveling off) at the appropriate time (upon reaching the appropriate altitude). Prospective memory tasks are generally either event-based or time-based, referring to which criterion is used as a mental cue or reference to retrieve the prospective task from memory at the right time. Generally speaking, the more a sequence is practiced and the more salient the

environmental cues, the better humans are at performing the prospective memory task (Ellis, 1996). Prospective memory theories allow researchers to explain and evaluate the outcome of deferred, yet intended actions. As another example, pilots must routinely complete a landing checklist to ensure the aircraft is configured properly for landing – e.g., landing gear is down and wing flaps are extended. Since this task is highly practiced and reliable, it is referred to as a habitual task (Dismukes, 2007). The aircraft itself also provides cues, like warning horns, to remind pilots to complete the landing checks prior to landing, and the system is generally very effective. If, instead, the pilots intentionally decided to delay a step of the landing checklist, e.g., keep the landing gear and flaps up to enable a faster speed on approach to stay ahead of an approaching thunderstorm, the delayed task would become an episodic prospective task.

Compared to a habitual task, humans are much more prone to forget an episodic task (Dismukes, 2007). Key to this discussion of prospective memory is the understanding that at any particular time, pilot errors are largely a function of the interplay between the concurrent tasks associated with managing the flight (radios, configuring the aircraft, etc.) and the encoding, storage, and retrieving (prospective memory) processes the tasks inevitably produce.

To elaborate an application of prospective memory a bit further, federal regulations and air traffic control (ATC) operating procedures are designed to ensure aircraft are efficiently and safely routed from departure to destination. Pilots begin the flight by gaining a very specific clearance (i.e., flight path “routing” guidance via expected airways and altitudes to fly) from an ATC dispatcher, but then have to be responsive to any changes that amend this original plan from that point forward. ATC may amend these directions periodically in the form of new altitudes, airspeeds, or headings. Changes are especially difficult and critical for pilots to effectively manage during periods of high workload. For example, as an aircraft nears its

destination airport, the pilots expect to adhere to a published flight path procedure called an instrument approach procedure (IAP). IAPs contain precise directions with regard to altitudes and headings and a printed copy of these directions is kept in the aircraft's cockpit. As an aircraft nears the airport, the air traffic controller will usually provide a course and altitude and an approach clearance, such as, "[aircraft call sign], descend to 3,000 [altitude], fly heading 180 [degrees], cleared for the ILS [type] approach to runway 22." This creates several prospective memory tasks for the aircrew. First, they start the "turn-to-heading" assignment, with the prospective memory task to stop the turn when heading gauge reads 180 degrees (i.e., south). Second, they must start a descent with the prospective task to level off at 3,000 feet above mean sea level. Third, they must refer to the "ILS runway 22" published approach procedure, which assigns headings and altitude requirements to complete a safe descent to the airport. This procedure can be fairly simple or complex based on the airport's instrument approach procedure itself. During these tasks, the increased prospective memory tasks contribute to an increased probability of making an error. Consequently, there is an increased dependence on the non-flying pilot to back up the flying pilot.

Standard Operating Procedures

Checklists, SOPs, and flight manual guidance are among the best tools aircrews currently have in the multi-pilot cockpit to prevent errors of omission, as they promote cross-checking and monitoring certain tasks at the most appropriate times to prevent the most likely errors of omission (Dismukes & Berman, 2010). One example of how airlines accomplish this is by specifying acceptable ranges for flightpath deviations during approaches – a phase of flight particularly susceptible to errors (Loukopoulos et al., 2009). Nonetheless, an in-depth analysis of flight operations manuals and SOPs at two major airlines found that these publications

described monitoring duties with insufficient guidance – particularly in terms of how frequently to check the status of the various items and how to appropriately divide attention between monitoring and other tasks performed simultaneously (Loukopoulos et al., 2009).

Leaders and managers are responsible to develop well-thought-out SOPs, by making monitoring guidance explicit in the way flight checklists and other flight procedures are written specific to each company and aircraft type (Sumwalt et al., 2002). In theory, the earlier deviations in altitude, heading, airspeed, or other systems are detected by the crew, the less likely the results will be consequential. Recently, researchers have leveraged explicit monitoring guidance in SOPs as an opportunity to measure the frequency of monitoring deviations using observational techniques. For example, during busy flight sequences like climbouts and approaches, PNFs are required to callout “1,000 feet to go” in advance of arriving at the target altitude (Loukopoulos et al., 2009; Dismukes & Berman, 2010). These types of procedures rely upon prospective memory tasks and are consequently more prone to error. One line observation study found that of the six to seven monitoring deviations aircrews make on average during a flight, late or omitted callouts (i.e., mandatory ones) account for over half of all monitoring deviations (Dismukes & Berman, 2010). Other, expected-but-voluntary pilot monitoring responsibilities include speaking up when a system malfunctions, ensuring controller directions are followed and challenging flight path deviations deemed outside a “normal” range. Organizations are increasingly making these normal ranges explicit, but standardizing is tricky as allowable deviations largely depend on the flight regime. An example of this would be a recent change in SOPs to mandate the PNF to call for a mandatory wave off if, at 500 feet above ground, the aircraft’s flight path met certain “unstabilized” approach criteria (e.g., too steep or too fast; Sumwalt et al., 1999; Dismukes et al., 2007; Dismukes & Berman, 2010).

Phase of Flight

Klinect, Wilhelm, and Helmreich's (1999) study of one airline's line operations identified that the "descent/approach/landing" phase contained the most aircrew errors and these errors tended to result in the most consequential outcomes. In another study accounting for 300 observations of line operations to capture various behavioral markers associated with threat and error management in the aircraft, Dismukes and Berman (2010) found that monitoring deviations (defined as omitted or late callouts, omitted verification, or not monitoring aircraft state or position) occurred most frequently during the climb (41.9 percent) and descent (26.5 percent) phases of flight. Both sets of data were consistent with a 1997 review of the NASA's ASRS database of incident reports which found 75 percent of monitoring errors took place during climbing, descending, or approach phases (Sumwalt et al., 1997). In fact, Sumwalt and colleagues (2002) compiled a diagram giving a profile view of a typical flight from the point of departure to its destination. Certain phases of flight are depicted in yellow indicating these are considered "areas of vulnerability" (p. 181, see Figure 1). The areas of vulnerability include taxi-out, climbout below 10,000 feet, transition altitude (i.e., the cockpit adjustment from referencing sea level altitude to flight level altitudes), within 1,000 feet of level off, the cruise-descent transition, descent, approach, and landing, and taxi-in (Sumwalt et al., 2002). There are several aviation safety implications of the research-based findings which link consequential monitoring errors and vertical phases of flight. The first is the understanding that these areas impose the greatest workload on aircrews, so the aircrews are most error-prone during these phases of flight. The second is the idea that errors committed in these regimes have fewer defenses in place to trap them, which means that monitoring is, in some cases, the last defense preventing a consequential outcome or one in which flight safety is compromised (Sumwalt et

al., 2002). The third takeaway is that AOV awareness should be incorporated into aviation education programs and lectures. In response to the frequency and criticality of these AOV-related errors, the vertical phase of flight will be the primary focus of monitoring performance in this study.

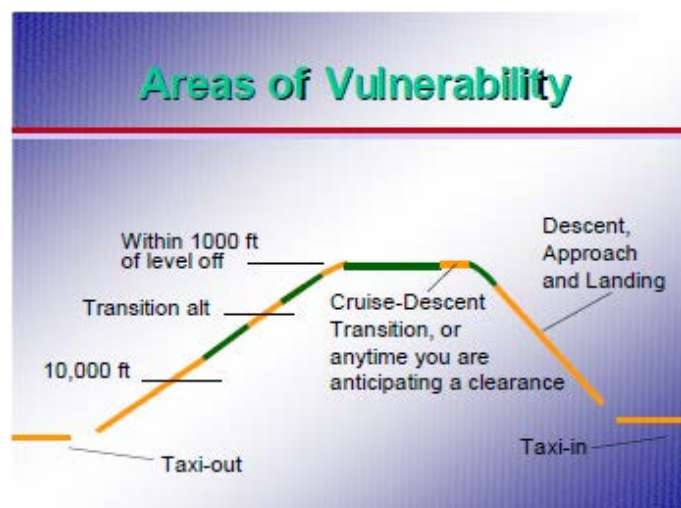


Figure 1. Sumwalt et al.'s (2002) Areas of Vulnerability in Commercial Aviation

Monitoring and Assertiveness

Crew resource management (CRM) has grown to become the premier human factors-skill training program for airlines, military, and most other aviation organizations (Kearns, 2010). CRM is used to minimize human error by teaching aircrews cognitive and social skills to improve their behaviors related to: decision-making, crew coordination, leadership, effective communications, risk management, situation awareness, and others (Salas, Wilson, Burke, & Wightman, 2006; Tullo, 2010). CRM training is usually classroom-based training, but in recent years there have been a few examples of computer-based CRM training formats (Kearns, 2009). As the content, instructor qualifications, and method of delivery for CRM training is highly variable across organizations and types of operation, this has made it very difficult to draw

conclusions regarding the effectiveness of CRM at the industry-wide level (Salas, Burke, Bowers, & Wilson, 2001).

One discussion in many CRM training programs emphasizes that pilots to be both vigilant and assertive, i.e., sustain one's attention and, regardless of rank or experience, be prepared to quickly "speak up" when unsafe flight conditions arise (Orasanu, Murray, Rodvold, & Tyzzer, 1999; Salas et al., 2006). The better CRM programs go further and provide junior pilots some examples of professional, non-confrontational methods to challenge authority in the flight environment should the need arise. Assertiveness is relevant because many apparent "monitoring" errors, as identified in mishaps, can actually be instances of poorly-applied assertiveness, e.g., the non-flying pilot saw something wrong but did not speak up because of professional courtesy or latent fears of repercussions (Jentsch, Martin, & Bowers, 1997). Studies on the role assertiveness plays in the multi-pilot cockpit provide insight into some of the individual pilot differences associated with our understanding of monitoring performance. For one, there is mounting evidence that errors committed by pilots-in-command are less likely to be caught and checked by the co-pilots than vice versa (Jentsch et al., 1997; Jentsch, Barnett, Bowers, & Salas, 1999; Sumwalt et al., 2002). In fact, 81 percent of commercial accidents from 1978 to 1990 occurred with the captain at the flight controls and similarly, 79 percent of accidents from 1991 to 2001 occurred with the captain at the controls (NTSB, 1994; Dismukes et al., 2007). This is consistent with research by Kontogiannis and Malakis (2009) that suggested that less experienced pilots would have less advanced mental models and therefore may function as less reliable monitors. However, combining these two NTSB studies and excluding specific cases where the captain rapidly took the controls from the first officer just prior to the accident, the percent of the time the captain was at the flight controls drops to 66, which is not

significantly greater than if this factor were random (Dismukes et al., 2007, p. 284). Helmreich and Foushee (1993) suggest that the less experienced pilots commit more monitoring errors because these pilots are (a) more error prone because of their inexperience, and (b) more hesitant to be assertive with a more senior pilot as they fear it will create additional tension between the pilots. Another study exploring these causes (Jentsch et al., 1997) identified three mutually exclusive explanations for why junior pilot fail at monitoring tasks. First, the junior pilots were not managing their workload well, and did not detect the errors. Second, in many cases the junior pilots were detecting the errors, but were not speaking up as they were not recognizing the problem required attention. Finally, similar to the Helmreich and Foushee (1993) finding, the junior pilots were sometimes afraid to speak up because of the expected negative consequences in terms of the senior pilot's reaction (see discussion in Besco, 1995; Jentsch et al., 1997). In summary, while the data does point toward increased pilot experience contributing to effective monitoring, this data tends to come from relatively small samples of accident data showing less-experienced pilots (who tend to be less assertive) not intervening in time to prevent an incident. This means that conclusive evidence whether pilots develop into better monitors with more flight experience is still lacking.

The continuing need to train pilots to assertively challenge in-flight inconsistencies is well documented in aviation research (Jentsch et al., 1997; Orasanu et al., 1999; Fischer & Orasanu, 2000; Dismukes & Berman, 2010). This training plays an important role particularly in ensuring that junior and less-experienced pilots are active participants and do not allow something bad to happen that they may have seen coming because they were afraid to speak up to a more experienced pilot. One such structured intervention is the P.A.C.E. (i.e., probe, alert, challenge, and emergency model; Besco, 1995). Other organizations (e.g., Flight Safety

Foundation and Department of Defense) have institutionalized other methods for intra-cockpit challenges, methods such as the “two-challenge rule” (Lindblom, 2007; Flight Safety Foundation, 2010). The problem, however, is that like the Jentsch et al. (1997) study found, the assertiveness role is contingent upon pilots first detecting these unsafe flight conditions. There is still a shortcoming in CRM curricula providing explicit guidance or training on how pilots can more effectively monitor, and therefore detect, these deviations or potentially unsafe conditions (Khatwa & Helmreich, 1998; Sumwalt & Lemos, 2010). On one hand, pilots are expected to have the technical proficiency to detect any and all errors and deviations, but on the other hand, they may not fully understand their performance limitations as they pertain to this skill. Sumwalt (1999) explained:

CRM courses deal with improving the ability of crewmembers to challenge others when a situation appears unsafe or unwise, many of these courses provide little explicit guidance on how to improve monitoring. We feel that carefully developed procedures and guidelines to enhance flight-crew monitoring can make a significant contribution toward improving aviation safety. (p. 2)

This is not to suggest that simply having organizations add a CRM module on the subject of monitoring is the best solution – at this moment, not enough is known in order to speculate on the best method to train monitoring skills (e.g., classroom, simulation, etc.; Sumwalt, 1999). While research has identified a few effective methods for training some categories of CRM skills such as teamwork (Brannick, Prince, & Salas, 2005), there is no empirical research specifically evaluating training programs (or parts of programs) targeted at improving pilots’ monitoring skills. Guidance for more effectively designing such training, however, is relatively abundant (Dismukes & Berman, 2010; Sumwalt et al., 2002; Thomas, 2004; Flight Safety Foundation,

2010). Dismukes and Berman (2010) suggest CRM training should include a module training pilots to counter their vulnerability to committing monitoring errors through practical techniques, e.g., “be slow and deliberate” and “respect human cognitive limitations” (p. 32). According to one aviation safety professional, training directed at improving pilots’ monitoring skills needs to be more interactive and engaging: “good monitoring skills are not inherent in pilots as they progress...effective monitoring techniques must be trained and rewarded” (Tullo, 2001, p. 106).

Evaluating Monitoring Performance

Any evaluation of a crew’s monitoring performance requires a distinction between the duties of the pilot flying (PF) and the pilot not flying (PNF) within the flight deck environment. The flying pilot’s primary responsibility, from the moment the aircraft begins to taxi, assumes a high level of active and conscious monitoring, particularly of the aircraft’s flight path or trajectory. For the PF, flight training and experience promote a mindset which suggests monitoring the primary flight instruments (i.e., altitude, airspeed, heading, vertical speed, and turn needle) is the most important factor within his or her overall responsibilities. Still, the PF’s job requires vigilance to counter complacency during low workload periods, and channeled attention to effectively manage concurrent tasks during periods of high workload (Jentsch et al., 1999). Data indicates that the workload involved with actually flying the aircraft can certainly reach a point that peripheral situation cues are likely to be missed (Jentsch et al., 1999). Since monitoring skills involve having to divide attention between keeping track of several aspects of the aircraft and what the other pilot is doing, the most important skill any pilot possesses is knowing where to focus his or her attention, i.e., awareness of what the most important things to monitor within their environment are. One early finding from simulator studies of cockpit management (CRM) was that the PFs may not be as vigilant monitoring the right equipment

when automation is being used. As Sarter, Mumaw, and Wickens (2007) reported, “monitoring failures constitute a major contributor to breakdowns in pilot-automation interaction” (p. 355).

Studies examining the limits of pilots’ cognitive abilities suggest that in multi-crew aircraft, the PNF’s level of active monitoring and cross-checking is critical to flight safety. As previously mentioned, the PNF is the pilot responsible for normal radio communications, navigation, and checklist duties. Observational research has in fact shown that the PNF actually has better overall situation awareness than the PF during critical phases of the flight (Jentsch et al., 1999; Thomas, 2005). This does not necessarily imply that the PNF has a less active job, but instead that the PNF is in a better overall position to allocate mental resources to manage and monitor the bigger picture, such as to catch a subtle change in a weather report or a miscommunication over the radio. During much of the flight, the PNF focuses attention on several tasks that take priority over the concurrent responsibility to monitor the actions of the other pilot. However, during the final few minutes of an approach, the PNF would have completed most navigation and communication tasks and should be actively involved in sustained monitoring and backing up the actions of the flying pilot. This is particularly important as it demonstrates that the PNF, with the benefit of their reduced workload during the approach, is in an excellent position to monitor the flight path, the flying pilot’s actions, and to serve as the final defense against any other threats and errors. The in-flight monitoring performance of the PNF is of particular interest to this study.

Training Pilots to Monitor

Developing on the idea that little research has tapped into the development of specific non-technical aviation skills (O'Connor, Hörmann, Flin, Lodge, Goeters, the JARTEL Group, 2002), this research intends to produce a literature-driven approach to the design, development,

implementation, and evaluation of a cognitive skill training program focused on aircrew monitoring. An effective training program is built on learning theory and lessons learned from previous designs. Learning is defined as “a relatively permanent change in human capabilities that is not a result of growth processes” (Noe, 2004, p. 107). Three learning theories in particular play a role in how pilots learn and each has a direct application to the design of future monitoring skill training programs. First, information processing theory is important because it describes a specific arrangement of internal processes (e.g. perceiving, storage, and feedback) that are a part of trainees’ ability to learn and retain information they have been taught. Within this model, feedback is emphasized as a critical factor which reinforces behaviors that are appropriate to store in the long-term memory (Noe, 2004). Sumwalt et al. (2002) specifically expressed the importance of feedback with respect to training pilots to be better monitoring pilots. Second, the cognitive load theory (CLT) builds on the processes as described by the information processing theory (Kearns, 2010). The CLT is based on the premise that a person learns through an interactive process in which active “chunks” of information in the limited-capacity working memory (WM, or short-term memory) become stored in the long-term memory based on the amount of rehearsal, elaboration, and practice they are given. Learning occurs at the moment the information gets stored in the long-term memory (Wickens, Gordon, & Liu, 2004). The CLT focuses on describing three key factors – intrinsic, extraneous, and germane – which affect the cognitive load the learner must endure to process and learn new information. Intrinsic cognitive load refers to the simplicity or difficulty level of the information itself. It is impossible for the training designer to control for the intrinsic aspect of cognitive load. Extraneous cognitive load refers to unnecessary information fed to learners outside of the to-be-learned material. A training design goal is to minimize extraneous cognitive load for the learner

through a clear presentation style (Kearns, 2010). Most important to this study is the germane element, which suggests that learning should be tailored to how learners process new information and construct schemas related to the training content. From a practical standpoint, the germane cognitive load on learners can be reduced by using step-by-step examples, self-explanatory activities, and practice. These techniques all promote the development and automation of schemas, reducing the overall cognitive load on an individual (Pass, Renkl, & Sweller, 2003). A third learning theory, goal theory, is important as it suggests that understanding what motivates the learners – intrinsic and extrinsic factors – drive learners' achievement of learning objectives. It is common to find intrinsic factors such as seeking to master a skill, ability, or competence in a task is a large motivator for pilots. Other extrinsically motivated trainees are motivated by the value they place in out-performing their peers at a task (Noe, 2004). Goal theory suggests it is important to first ensure training is designed to increase pilots' competence at a specific task, then to clearly express that objective to the learners. This objective must be ultimately fulfilled in the training presentation by giving all training elements this same goal-oriented context (Noe, 2004).

Given their abundance in aviation, CRM training programs provide a general blueprint to draw upon in the development of similar non-technical skill training programs. Given the mixed results regarding the overall effectiveness of CRM training to perceptibly improve in-flight behaviors (Salas et al., 2001, 2006), there are opportunities to draw upon some of its documented successes in the future design of other non-technical skill training programs. In the case of developing guidelines to improve pilot monitoring, the desired learning outcome is to train a cognitive strategy: in this case, the learner must know how to scan and process information and other signals coming from the aircraft and the other pilot and then decide how to allocate his or

her attentional resources. Training a cognitive strategy should include (a) a verbal description of the strategy, (b) strategy demonstration, (c) practice with feedback, and (d) use of a variety of tasks that provide opportunity to apply the strategy (Noe, 2004). Jensen (1995) contends the three learning principles of “practice, feedback, and ‘follow ideas with examples’” are the most reliable conventions for generating behavior change (pp. 177-182). The use of situational training or specific examples, followed by presenting opportunities for the learner to respond, is especially reliable methods for teaching pilots (Jensen, 1995). According to Salas and colleagues (2006), existing CRM training programs apply similar strategies, but lack opportunities for learners to immediately practice what they have learned with feedback. Consequently, to be effective, any new training program should incorporate these lessons.

An often overlooked, but critically important final step in training development is planning for training evaluation (Noe, 2004). According to Kirkpatrick’s (1976) framework, training evaluation is important as a tool to (a) decide whether to continue the program, and (b) gain insights into how to improve the training program. Ideally, Kirkpatrick suggests, training evaluation should be designed to consider outcomes at four levels: reactions, learning, behavior, and results (1976). Reaction outcomes refer to the attitudes and perceptions of the learners toward the training and whether it satisfied their needs, i.e., how they liked the training and how useful they found it. Kirkpatrick recommends that organizations measure the trainees’ reactions immediately and, when a multiple-choice questionnaire is used, to provide a range of quantitative responses to each question. The second evaluation method, learning, refers to extent which learners improved their knowledge of the subject matter of the training. Usually, this measure consists of providing the learners a post-training test to identify if the learners acquired the knowledge and/or cognitive skills of interest. To measure the learning effects of a training

program, Kirkpatrick recommends using pre- and post-training tests and comparing results between an experimental and control group. Also, Kirkpatrick recommends the test's measures should be objective and quantifiable. The third level, behaviors, examines whether the trainees' behavior changed measurably as a result of the training. In aviation, such an evaluation for a flight training program would consist of measuring indications that the training has changed pilot behaviors in the flight environment. Organizations would accomplish this by taking appropriate measurements before and after training or comparing performance between a control and experimental group. Kirkpatrick suggests four conditions must exist in the learner for behaviors to change. The learner must: (a) have a desire to change, (b) they must know what and how to change, (c) the right climate must exist, and (d) they must be rewarded for changing (Kirkpatrick, 1976). The fourth training evaluation level, results, refers to whether the training has had an impact on some aspect of the organization's bottom line. Kirkpatrick points out that the focus and driving force behind developing the training program should be a results-level goal. An organization's results goals could be increased productivity, safety, cost savings, or employee retention (Kirkpatrick & Kirkpatrick, 2006). To an aviation organization, for example, the results of CRM training could be measured in terms of fewer human-error related accidents. CRM training programs have, in fact, been evaluated from the perspective of Kirkpatrick's framework (Salas et al., 2001). Research into these programs identified that most CRM-related training evaluations are more likely to consider either the reactions and learning outcomes than the behavior and results outcomes as the latter two can be more difficult and time-consuming to measure (Salas et al., 2001; Salas & Cannon-Bowers, 2001; Salas et al., 2006).

E-Learning

Technology plays an increasingly powerful role in today's training systems and offers many advantages compared to more traditional training methods (Noe, 2004). In an exploratory study, Raisinghani and colleagues (2005) found that pilots generally feel comfortable with technology and confident in their personal ability to use e-learning. This research considers the effectiveness of online training to train non-technical aviation skills. Online training (also referred to as web-based training) refers to an instructional strategy in which a web-enabled computer provides the learning stimulus, the trainee must respond to the stimulus, and the computer analyzes the responses and provides feedback to the trainee (Rosenberg, 2001). Online training can be delivered to learners by workstation, by smartphone, or by any other web-enabled device. By comparison, e-learning is an umbrella term referring to delivery of any informational content (does not necessarily have to be of an instructional nature) through networked electronic technology (Rosenberg, 2001). E-learning is a term used in literature to contrast this type of training from classroom-based learning and blended learning (i.e., combined use of classroom and e-learning), the two other categories of learning environments. From a training design and evaluation perspective, the effectiveness of any e-learning method greatly depends on applying good learning principles (Noe, 2004, p. 233). One major advantage of online training is that it can be configured to provide the learner with a more personalized training experience, providing trainees with feedback on their performance using interactive games, simulations, quizzes, and exercises to determine if they need additional practice (Kearns, 2009). Another advantage of online training is that the programs tend to be highly accessible and self-paced, giving the learner more control and flexibility to complete the training at a time convenient for them. Still, e-learning (and hence online training) are not ideal for all types of training and upfront

development times and costs are generally higher. Also, motor skills and attitudes train poorly through e-learning, and generally these programs lack interactions with the instructors and other learners which, in a lecture setting, for example, can add to the value of the experience (Kearns, 2009).

Studies into the effectiveness of online training can generally be found in meta-analyses of comparative studies between classroom-based and e-learning (Bernard et al., 2004; Clark & Mayer, 2006). Bernard and colleagues (2004) found no differences in effectiveness between e-learning courses and classroom training, but did find wider variability in learning improvements for e-learning courses. Specifically, they found asynchronous training outperformed classroom training, while synchronous training did not (pp. 408-409). Asynchronous refers to self-paced training that allows learners to complete the training at a time convenient for them. Synchronous training, by comparison, refers to training in which the learners complete online training simultaneously from various locations. This variability indicates that the more important factor is the quality of the training, specifically, whether e-learning leverages the appropriate amount of technology and tailors the course material to incorporate interaction and feedback into the program design (Clark & Mayer, 2006). Designers are encouraged to adopt a constructive architecture approach, where learners are provided opportunities to elaborate on concepts that actively involve them and stimulate prior knowledge (Kearns, 2009). The bottom line is that while e-learning has been shown to be more effective than classroom-based learning in some cases, what is most important is that the designer understand issues of course content, instructional design, and learner characteristics in order to maximize the effectiveness of e-learning (Rosenberg, 2001; Kearns, 2010).

Purpose of this Study

Data from aircraft accidents and line observation studies indicate that inadequate and insufficient pilot monitoring and cross-checking is a growing safety concern within the industry. Within the cockpit environment, there is ample evidence that pilots who fail to properly manage their workload commit more monitoring errors. Given the lack of monitoring training programs available to pilots, more research is needed to investigate the effectiveness of types of training that can be used to improve their cross-checking and monitoring skills. The present study seeks to determine the effectiveness of an interactive, online program to improve pilots' monitoring behaviors. The focus of this research was to measure voluntary changes in non-flying pilots' monitoring practices, particularly within the framework of high-workload environments. The researchers designed and administered an online training program with a strong focus on providing opportunities for the learners to practice and receive performance feedback during in-flight scenarios. Our objective was to show positive transfer of monitoring behaviors trained in the online program to the cockpit of a high fidelity flight simulator. A multivariate analysis was used to evaluate the training's effectiveness by comparing specific, voluntary behaviors of trained teams versus control teams.

Statement of Hypotheses

The primary research question is to determine if online training would affect the pilots' performance at in-flight monitoring-related tasks. It is expected that the training will improve both pilots' understanding of monitoring and their performance of monitoring tasks during vertical phases of flight. The following hypotheses were examined:

Hypothesis 1. Trained participants will react more positively to the training activity than the control group reacts to the control activity. Specifically, according to a composite reactions

score, trained participants will rate the training's usefulness and instructional value higher than participants in the control group.

Hypothesis 2. Trained participants will be able to demonstrate knowledge and application of the training program's learning objectives in a written test. Specifically, participants completing the online monitoring skill training will score higher on the post-training multiple-choice test than the pilots completing the control activity.

Hypothesis 3. There will be a transfer effect of the online training as evidenced by use of the "5-2-1" rule during a training event in the flight simulator. Specifically, aircrews in the training group will use the 5-2-1 callouts more frequently than the untrained pilots in the flight simulator.

Hypothesis 4. There will be an effect of the online training on the number of voluntary callouts pilots make during a training event in the flight simulator. Specifically, non-flying pilots in the training group will make more altitude callouts per minute (i.e., callout ratio) than non-flying pilots in the control group in the flight simulator.

Hypothesis 5. There will be an effect of the online training on the aircrews' success at climb and descent tasks during a training event in a flight simulator. Specifically, the trained group will make fewer errors in successfully completing climb and descent tasks than the control group in the flight simulator.

This study also includes a possible covariate, aviation experience, as previous studies suggest that senior pilots tend to be more proactive monitors than junior pilots (Foushee & Helmreich, 1988, Jentsch et al., 1999). This study predicts:

Hypothesis 6. There will be a relationship between aircrew experience and performance gains from the online training. Specifically, aircrews containing a junior pilot (i.e., one that is

designated a copilot or first pilot) will experience greater performance gains across the dependent variables in comparison to the aircrews with more senior pilots.

Methods

Design

This field study used a posttest-only design with a comparison group, comparing behavioral differences between an experimental (training received) and a control (no training received) group. Monitoring skill training was the between-subjects independent variable; participants in the experimental group completed a 30-minute computer-based monitoring skills training program while participants in the control group completed a control activity. Following the training activity, participants were paired into teams and flew a 2.5-hour event in a high-fidelity flight simulator, followed by a similar 2.5-hour simulator event the next day. The pilots used for the study were all fully-qualified in the aircraft, although experience levels and designations varied somewhat between pilots.

Monitoring skills were measured according to D. L. Kirkpatrick's 4-level framework (1976). Researchers first measured variables regarding participants' reactions and learning outcomes using a questionnaire and post-training test. Next, the researchers measured aircrews' behaviors and task outcomes along four dependent variables by observing recorded data from a flight simulator.

Participants

Participants in this study were U.S. Coast Guard (USCG) pilots qualified to fly the H-65 helicopter ($N = 55$). The ages of the participants ranged from 26 to 52 years ($M = 33.5$). Fifty were males and five were females. Total flight hours ranged from 430 to 4,700 ($M = 2,014$, $Mdn = 1,786$). At a minimum, each pilot had completed initial military flight training at least a year prior and was currently instrument-rated. Based on a combination of experience with the type of helicopter (H-65) and USCG operations, each had achieved one of four pilot designations within

the service. Ranked from lowest to highest, these designations are: copilot (CP), first pilot (FP), aircraft commander (AC), and instructor pilot (IP). At least an AC designation is required for a pilot to be put in charge of an actual USCG mission. This designation is aircraft-specific, so, for example, a 5,000-hour pilot who is newly qualified in this type of helicopter could still be designated as a CP, even if they had achieved an AC rating in another type of aircraft. There were no significant differences between the groups in terms of age, pilot designation, total flight hours, flight hours in type aircraft, or hours flown in the last 60 days. A summary of pilot experience is provided in Table 1.

Table 1

Pilot Experience

	Trained (<i>n</i> = 25)	Control (<i>n</i> = 30)	Total (<i>N</i> = 55)
Gender	2 female, 23 male	3 female, 27 male	5 female, 50 male
Mean age	34.9 (6.54)	32.3 (3.86)	33.5 (5.39)
Mean total flight time	2,150 (1,260)	1,896 (966)	2,014 (1,103)
Mean flight time in type	1,466 (1,065)	1,275 (939)	1,364 (992)
Mean flight time past 60 days	38.9 (16.8)	40.7 (15.6)	39.9 (15.4)
Current designation	6 CPs 2 FPs 12 ACs 5 IPs	6 CP 2 FP 17 AC 5 IP	12 CPs 4 FPs 29 ACs 10 IPs

Note: Mean results are shown with standard deviations in parentheses (where applicable).

All participants completed the training activity and the master questionnaire containing the *reaction*, *test*, and *attitudes* measures. Of these 55 participants, 44 participants were paired into two-person teams whose performance was then evaluated in the flight simulator ($N = 22$). Eleven of these participants (3 trained and 8 control participants) were not included in the results from the simulator observations discussed in hypotheses 3 through 6. The two reasons participants were not included in the simulator observations were because either: (a) the participant's simulator event was not recorded (e.g., the simulator event instructor/supervisor

forgot to press record button) or (b) only one of the two paired participants agreed to participate in the study (e.g., one participant opted out of the research at some point). The researchers decided to keep the reactions and learning measures from these participants as this data still provided relevant information to help validate the effectiveness of the online training and contributed to the analysis of monitoring-related attitudes. Demographic and flight experience information for the participants who were included in the simulator portion of the study is included in Table 2. Because these 44 participants were observed in pairs during the simulator portion of the study, the number of total observed “teams” was cut in half to 22, i.e., an *n* of 11 teams observed in each of the two conditions.

Table 2

Demographics for Participants in Flight Simulator

	Trained (<i>n</i> = 22)	Control (<i>n</i> = 22)
Mean total flight time	1,970 (1,148)	2,106 (949)
Mean flight time in type	1,392 (1,074)	1,389 (1,015)
Mean flight time past 60 days	41.1 (16.3)	42.7 (12.2)
Current designation	5 CPs 1 FPs 12 ACs 4 IPs	3 CP 2 FP 12 AC 5 IP
Gender	1 female, 21 male	2 female, 20 male
Mean age	33.7 (5.19)	32.9 (3.77)

Note: Mean results are shown with standard deviations in parentheses (where applicable).

To recruit participants, the USCG allowed the researchers to use pilots attending a one-week annual proficiency course in Mobile, AL. Participants normally complete 8 to 10 ½ hours of classroom training, participate in four 2- to 3-hour events in a full-motion flight simulator, and take a written exam as part of their normal duties during the week (see Table 3). This study added a 30-minute training activity, completed individually, to augment the week’s training, and examined performance on two of the pilots’ simulator events. Participants were paired for flight

simulator training by the USCG at least a month in advance the course. The requirements for a crew pairing were that the pilots cannot be from the same unit and that at least one pilot holds an AC or IP designation. Given the approximately 500 H-65 helicopter pilots in the USCG fleet, it would be unusual that the pilots paired for the proficiency course had previously flown together.

Table 3

USCG Training Schedule during 1-week Proficiency Training Course

Name	Type	Day of week
Instrument flight rules training	Lecture	Monday
Emergency procedures training	Simulator	Monday or Tuesday
Flight planning training	Lecture/Demo	Monday
Instructional methods (optional)	Lecture	Monday
IFR scenario #1*	Simulator	Tuesday or Wednesday
Ship-helo operations	Simulator	Wednesday
USCG SOP/regs training	Lecture	Wednesday
CRM training	Lecture	Wednesday
IFR scenario #2*	Simulator	Wednesday or Thursday
Operational procedures check	Simulator	Thursday or Friday
IFR procedures exam	Online Exam	Due NLT Friday

*Source of data captured for this study

Apparatus

Informed consent. The researchers used an informed consent form, based on the American Psychological Association's standard, procedures, and protocols. The consent form was approved by both Embry-Riddle Aeronautical University's institutional review board and the USCG's scientific review committee (SRC) and will be stored with project records for a period of six years. In the consent form, each participant acknowledged they understood their participation was voluntary and that their identity would be kept confidential. A copy of the informed consent form is included in Appendix A.

Demographic survey. Researchers administered an online demographic survey to collect general information regarding the participants. The survey collected specific information which could have an effect on measurement outcomes. Info included: age, gender, total flight

time, flight time in type aircraft, flight time in the past 60 days, and crew designation (CP, FP, AC, or IP; see Table 1). A copy of the demographic questions is included in Appendix B.

Online training. The training course was an original, asynchronous, computer-based training program titled: *crew monitoring and cross-checking skill training*. The training was designed following the ADDIE (analysis, design, develop, implement, evaluation) framework for instructional design (Kearns, 2010). The first step was to analyze the pilots' need for monitoring skill training and to plan a tailored e-learning instructional approach. The researchers chose to deliver the course via computer to allow the learning to be a highly interactive experience for the learners, providing them with performance feedback spaced throughout the lesson. The electronic, web-based format also allowed the researchers the ability to track learner performance. After the basic architecture was designed, the course was developed using Microsoft PowerPoint. Three subject-matter experts (SMEs) reviewed the training and provided feedback. Finally, the training was published for use in Adobe PDF format. The PDF was uploaded to the website containing the online questionnaire and so learners could only link to the training after finishing the demographic survey. The instructions explained that the course had to be completed in one sitting and would take approximately 30 minutes.

Based both on Sumwalt et al. (2002) and Loukopoulos, Dismukes, and Barshi's (2003) recommendations for aircrew-specific monitoring training, the course followed a five-step outline using the following headings: *importance of monitoring*, *what to monitor*, *how to monitor*, *practice*, and *end-of-course test*. Upon completion of each of the first three sections, the program required the learners answer 2 or 3 multiple-choice questions to evaluate their understanding of the material presented in the preceding section. The learners received feedback on whether a response was correct or incorrect, and the program's functionality allowed learners

to repeatedly select incorrect responses until the correct answer was chosen, at which point the learner could move to the next section.

The following is a description of each section. *Importance of monitoring* provided learners with a vivid real-world example of the consequences of poor monitoring and familiarized the learners with monitoring concepts. The section *what to monitor* focused on teaching the concepts behind the “areas of vulnerability” chart (see Figure 1). The section *how to monitor* taught the learners how to apply the “5-2-1 rule” to prevent in-flight level-off errors. The 5-2-1 rule was developed by the researchers specifically for this study as a memory device to promote having non-flying pilots monitor and callout specific altitudes to alert the flying pilots that they are approaching a required level off altitude (or decision height). During vertical phases of flight (i.e., climbs and descents), the rule required the PNF callout at 500, 200, and 100 feet prior to reaching all target altitudes. The rule was loosely derived from an airline SOP requirement that states the PNF shall make a callout at 1,000 feet prior to arriving at any target altitude and then both pilots must exclusively monitor altitude until the level off is complete (Sumwalt et al., 2002; Dismukes & Berman, 2010). Since helicopters make smaller-increment altitude transitions and use climb rates of a smaller magnitude, the researchers chose to use a smaller scale (relative to commercial jets operations) of 500 feet and below for this training program. The next section, *practice*, gave the learners a chance to apply the 5-2-1 rule in several scenarios. Since this rule required some simple mental math, each scenario presented the aircraft in a climb or descent and had the participants respond where the next callout would occur. The researchers believed that having learners practice scenario-driven mental math in these exercises was an effective method for facilitating development of the same mental processes required for the in-flight 5-2-1 task and increase the transfer of training. Following the practice session, the

participants were given a link to a 10-item multiple choice test to assess the participants' knowledge and application of the monitoring strategies.

Extensive effort was devoted to create a novel, yet practical training program. The author developed this training curriculum after having taken a graduate-level course in training systems and performance measurement. The author also consulted several articles (Jensen, 1995; Sumwalt et al., 2002), books (Noe, 2004; Kearns, 2010), safety training presentations (Dismukes & Berman, 2007; Sumwalt, 2009; Curzio & Arroyo, 2010), and on-line resources in the design and development of this training program. The researcher also consulted aviation SMEs within the USCG to ensure the content and methods align with the organizational goals (R. Donnell, personal communication, January 30, 2011). A complete reproduction of the lesson slides used for the training program is included in Appendix C.

Control group activity. From a research design perspective, providing control group participants with a control activity is useful in keeping the research experience of the experimental and control groups as identical as possible. In this study, the researchers gave the control participants a six-page article to read from Flight Safety Foundation's *AeroSafety World* magazine on the subject of aircrew multitasking (Loukopoulos, Dismukes, & Barshi, 2009, August). Following the activity, a 10-item multiple choice test was administered to participants identical to the one given to those receiving the computer-based monitoring training.

Questionnaire. The sections described below – containing a test, a reactions questionnaire, and an opinions questionnaire – were all presented in a single master questionnaire constructed using the web-based *SurveyGizmo* program, a product of Widgix, LCC (www.surveygizmo.com). Two master questionnaires were designed and used for this study: one was administered to the training group and one to the control group.

Post-training test. The researchers designed an online, 10-item multiple-choice test to assess the participants' learning, i.e., examine whether the learning objectives were met. The researchers gave the same test to both groups of participants – the trained participants completed it upon completion of the online training and the control group completed the test following their control activity. Four questions tested the participants' ability to apply the 5-2-1 rule and six questions tested their declarative knowledge of monitoring concepts, such as the *areas of vulnerability* chart. Appendix D lists all 10 questions.

Reactions questionnaire. Following the training, the researchers gave participants in the training group a seven-item, online reactions questionnaire to assess their attitudes (i.e., satisfaction and perceived utility level) toward the training. The researchers provided the control participants with a modified questionnaire containing only four of the seven questions that were provided to the training group. The researchers agreed that three of the reactions questions were critical to the trainees' evaluation of the utility of the computer-based training program, but were distracting to the control group participants' evaluation of the control activity. The questions are presented in Appendix E.

For purposes of presenting the results of the reactions questionnaire, three questions were chosen to group into a composite reactions score for each participant and allow for a between-groups comparison of reactions to either the training or control activity. The reliability of this composite reactions score (based on averaging identically-scaled questionnaire responses) was determined to be high, as Cronbach's $\alpha = 0.79$.

Attitudes questionnaire. Gagne (1984) held the belief that attitudes should be considered a learning outcome, explaining that attitudes affect individuals' behaviors, actions, and performance. Kraiger, Ford, and Salas (1993) also contended that Kirkpatrick's (1976)

framework ignores affectively-based measures of learning, opting instead for reactions measures. Kirkpatrick and Kirkpatrick (2006) have since elaborated that changes in trainees' attitudes are a dimension that may be used to infer "learning" outcomes, which include knowledge, skills, or attitudes. Given that so little is known about pilots' monitoring habits and priorities, the researchers thought it useful to solicit attitude data pertaining to pilot's attitudes towards specific monitoring roles and responsibilities. This data would contribute to a needs assessment for the development of future training programs related to monitoring. The researchers designed a 23-item assessment of participants' attitudes and perceptions toward monitoring and other non-flying pilot duties. The questions here were directed at determining how the participants ranked monitoring among their other non-flying pilot duties, what flight instruments they considered most important to monitor, and what phases of flight they considered monitoring to be most important. None of the questions were directly related to any of the online training program's content. All participants completed this section of the questionnaire as the last section in the questionnaire (Appendix F).

Flight simulator and debriefing station. Participants' monitoring performance was observed using their recorded performance in an H-65 helicopter (military version of the Eurocopter AS365) simulator. This is a high-fidelity representation of an aircraft each participant was fully qualified to fly. The flight simulator rests on a motion platform which moves in six degrees of freedom and provides the pilots with more than 225 degrees of horizontal visual display of the outside world (see Figure 2).



Figure 2. USCG H-65 Helicopter Flight Simulator.

Each simulator event began in a briefing room with a single flight instructor briefing the two-person crew on the content of the session. The crew boarded the simulator with the two pilots (participants) sitting side-by-side in the pilot seats and the instructor sitting behind them at a control console, with the ability to manipulate the simulation (e.g., start, stop, introduce poor weather, reposition the aircraft, introduce simulated emergencies, etc.). Additionally, the instructor played the role of the air traffic controller (ATC) and communicated with the pilots to provide flight clearances and other required messages. The flight simulator was equipped with flight data and voice recording capabilities which were uploaded to a single-purpose workstation (computer) with three monitors called the computer-aided debriefing station (CAD), which was located in an adjacent briefing room. When the event was over, the computer used specially-designed software (called the “CADS replay client”) as an interface to enable the instructors to debrief the pilots by examining recorded data from the flight instruments and radio communications. Like any digital recording, the instructors had the option to play the entire

event or jump to a specific task to view. The CAD display also provided an over-the-shoulder (OTS) video recording of the event (upper-right corner, Figure 3). For this study, the experimenter exclusively used the CAD software to review and collect data from simulator events, using a process which will be described in the next section. Note: The CAD software did not enable the researchers to view raw or exact simulator data, e.g., to-the-exact-foot altitude readings. This data was instead read from a needle on a computer rendering of the same analog gauge that is in the aircraft (see primary instrument displays in lower right corner of Figure 3).

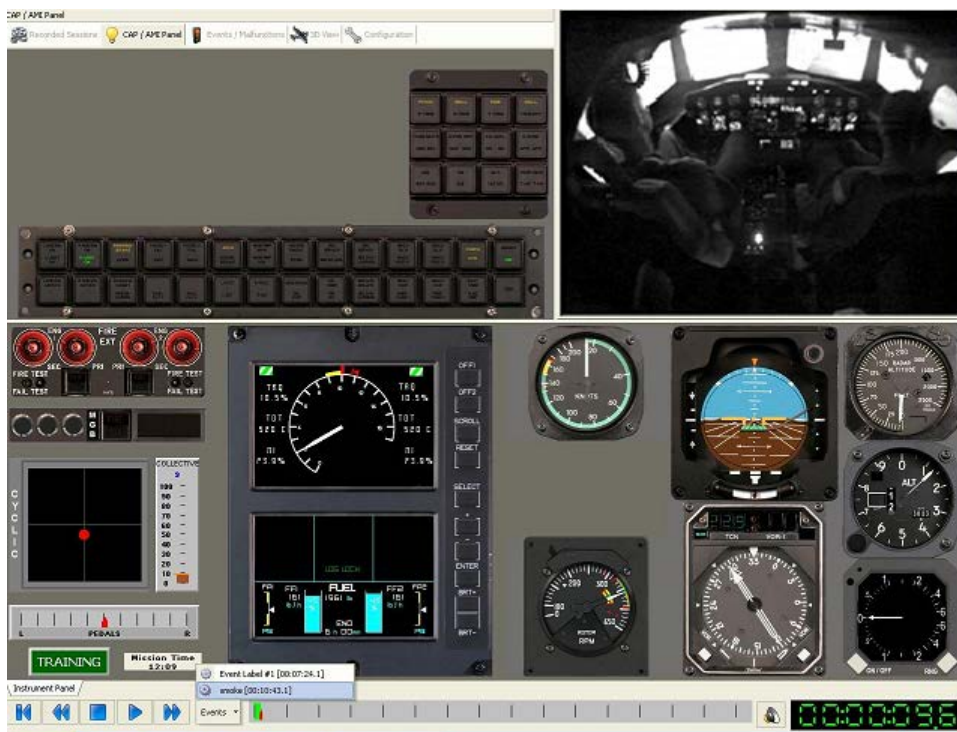


Figure 3. CADS Replay Software Screen

Flight scenarios. Recent research has made use of Line Operations Safety Audits (LOSA) and Line Operational Simulations (LOS) to enable organizations to gather robust performance data exploring the types of errors flight crews make during line operations and scenario-based training events (Sumwalt et al., 2002; FAA, 2004b, 2006). For example, this data has enabled managers to identify error-prone situations and then to adjust the relevant

operational procedures and checklists to make them safer (Flight Safety Foundation, 2005; FAA, 2006; Loukopoulos et al., 2009). The use of data gathered in real or simulated line environments marks an important step towards proactively improving aviation safety, i.e., without having to wait for something to go wrong to realize a problem exists.

For this study, the researchers used the CAD recordings and adopted a similar observational strategy to evaluate aircrews' monitoring behaviors during routine simulator training events. A group of USCG flight instructors (i.e., SMEs) develop standardized simulator scripts for an assortment of pilot skill training and mission training needs. The scenario scripts are updated annually, therefore, the types of tasks, stressors, and decision-making scenarios presented to the aircrews during these events should be up to date, challenging, and representative of the line environment (J. Cabell, personal communication, November 1, 2010).

The researchers examined crew performance in a flight simulator during two specific "scenario" events designated as *IFR scenario #1* and *IFR scenario #2*. The maneuvers and context of these scenarios were scripted based on the tasks required to complete an annual USCG instrument proficiency check flight, which was consistent with the requirements of a FAA instrument proficiency check (FAA, 2004c). In terms of task similarity, the two IFR scenario events were very similar (Table 4).

During either *IFR scenario* event, the instructor evaluated the two pilots from a control station as the crew conducted a planned flight from a departure point to a destination airport with four practice instrument approaches and other tasks along the way. Particularly of interest to this study were the crews' monitoring behaviors during vertical flight (i.e., climbs and descents). There are two general categories of instrument approaches – precision and non-precision – of which crews must fly two of each. An instrument approach is a procedure associated with a

specific airport to allow a crew, solely using their aircraft instrumentation, to fly from a predetermined point in space down to a position close to the runway at an airport where the crew can visually land the aircraft. The point at which the crew must be able to land the aircraft visually is called either the *missed approach point* (MAP) or *decision height* (DH), at which time, if the crew cannot proceed visually, must instead abort the approach and climb away from the airport in a procedure called the *missed approach*. Precision approaches require more accurate navigational aids and follow more restrictive flight path parameters which allow the aircraft to descend a few hundred feet closer to the runway environment without having it in sight. The pilots have planned the route of flight in advance, so the sequence of three of the approaches was known prior to the flight. The final approach was added to the scenario by the instructor manipulating the weather conditions at one of the airports so that an alternative approach must be attempted. From the USCG's perspective, each event was designed to evaluate the pilot in the role of the pilot in-command (PIC).

The main difference between the two IFR scenarios was that the two flights use different airports, meaning that the altitudes, headings, and timing requirements built into each of the approaches are different. The instructors generally followed an instructional script kept secret from the crews under evaluation. The instructor's script ensured a high level of standardization between flights – it prompts the instructors to manipulate certain in-flight variables to give the flight more task realism, pose certain challenges to the crews, and force the crew to adapt to changes from the original plan – all while ensuring the basic and necessary tasks of the flight were completed. The instructors had the authority to customize the training to an extent (e.g., vary holding instructions, increase frequency of altitude changes), and initiate three or four

simulated in-flight emergencies of their choosing per flight (J. Cabell, personal communication, November 1, 2010).

Table 4

Instrument Approaches Flown by IFR Scenario

IFR scenario #1	IFR scenario #2
1. PAR (precision approach radar) approach to runway 22 at NAS New Orleans	1. ILS approach to runway 6 at Columbus Metro Airport
2. NDB (non-directional beacon) approach to runway 18 at Hammond Regional Airport	2. NDB approach to runway 6 at Columbus Metro Airport
3. VOR (VHF omni range) approach to runway 22 at Baton Rouge Metro	3. VOR approach to runway 15 at Lawson AAF
4. ILS (instrument landing system) approach to runway 22 at Baton Rouge Metro	4. PAR approach to runway 33 at Lawson AAF

The other difference between the flights was that for evaluation purposes, the role of PIC was exchanged between the pilots. For all flying events, one pilot is assigned the authority role of pilot-in-command (PIC), and the other is assigned the role of copilot (CP). These roles are equivalent to the *captain* and *first officer* roles used in commercial aviation, respectively, in terms of authority and responsibility. The PIC role is assigned to the pilot with the higher designation and if the designations are the same, to the pilot with more seniority within the organization. So, for purposes of training and evaluation, instead of a normal crew assignment, the junior pilot would be assigned as PIC over the more-experienced pilot for one of the simulator events.

Debriefing script. As an ethical precaution, the researchers debriefed each of the participants at the completion of the experiment. A copy of the debriefing script is presented in Appendix G.

Procedure

Prior to attending the USCG flight training course, all pilots scheduled for training during a given week were assigned into two-person teams (i.e., individual flight crews) by the USCG. The researchers assigned the two-person teams to their condition – training or control – based on a predetermined alternating week-to-week schedule. This meant that all teams arriving on a given week would be assigned to the same condition. (Since the pilots attending the course on any given week spend a significant amount of time together as a larger group [classroom training, meals, berthing, etc.] it was determined that splitting groups within a week opened up the possibility that the individuals would discuss the two training activities during breaks and social periods of the week and introduce a potential confounding variable.) Based on this assignment, the researchers ensured that once the pilots arrived at the course both paired pilots followed the set of instructions in accordance with the correct sequence for their assigned condition, as outlined in Table 5. The pilots performed all steps of the sequence individually except for the flight simulator portion, which they completed as a crew of two.

Upon arrival at the first training event of the week, the researchers provided all participants with the informed consent form to review and sign (Appendix A). At the bottom of the consent form was a website link to the online training (or control activity) and questionnaires. Each participant received one of two links, depending if they were in the training group or the control group. The researchers directed pilots to individual workstations with instructions to login and follow the program's directions. Once logged in, the participants first entered their demographic information (Appendix B). At this point, the questionnaire directed the participants in one of two ways based on the training condition. Pilots in the training/experimental group completed the "aircrew monitoring and cross-checking" training session (Appendix C). The

pilots in the control group completed the control activity. Having completed either activity, the participants next completed a 10-item multiple choice test to evaluate their learning of the material (Appendix D). The participants' responses on this test were recorded. After the test, the participants continued using the online questionnaire to provide their reactions to the training activity (Appendix E). Here, the questionnaires differed slightly by condition: the training group responded to seven (7) questions with regards to the training while the control group answered just four (4) questions. Following the reactions portion of the questionnaire, each participant answered a 23-item survey exploring their attitudes and perceptions towards the monitoring roles and responsibilities of pilots (Appendix F). This activity completed the training and questionnaire portion of the experiment and the participants were informed the experimental training was complete and they should follow their normal USCG-directed schedule to complete the rest of the week's training events (see Table 3).

The course's week-long flight simulator schedule normally consisted of each aircrew completing one simulator event per day, according to a prescribed sequence. In each case, the teams all flew a 2.5 hour "emergency procedures training" event in the flight simulator before commencing IFR scenario #1. This "warm-up" event, in effect, helped the participants to adjust to the simulator and reduced any unfamiliarity with the simulator that would affect their performance on the IFR scenarios (the events of interest to this study). All of these events were recorded using the CAD system, yet neither the pilot participants nor their instructors had knowledge of which flights or aspects of the simulator flights were of interest to the study, outside the information provided in the participant consent form. At the end of the week, the researchers provided a debriefing summary to each participant.

Table 5

Procedure

<i>Experimental Group</i>	<i>Control Group</i>
Consent form: Participants read and sign consent form.	Consent form: Participants read and sign consent form.
Demographics: Participants answer six demographic questions including: <i>total flight time, flight time in type aircraft, 60-day flight time, pilot designation, age, and gender.</i>	Demographics: Participants answer six demographic questions including: <i>total flight time, flight time in type aircraft, 60-day flight time, pilot designation, age, and gender.</i>
Training: Participants in this group complete a 30-minute online training program.	Control activity: Participants read a five-page article on multitasking (15-20 minutes)
Test (learning): Participants were administered a 10-item test to evaluate whether they learned the information contained in the training program.	Test (learning): Participants were administered a 10-item test to evaluate their knowledge of the information contained in the training program. (This group was not trained the material covered in the test, but was included to establish a baseline level of pilot knowledge on the subject matter.)
Reactions: Participants answer seven (7) questions regarding their attitudes toward the online training program.	Reactions: Participants answer four (4) questions regarding their attitudes toward the control activity.
Monitoring habits and attitudes questionnaire: The participants answer 23 questions expressing their monitoring habits and perceptions of monitoring importance.	Monitoring habits and attitudes questionnaire: The participants answer 23 questions expressing their monitoring habits and perceptions of monitoring importance.
Flight simulator: Over the next two days, participants complete two events in a flight simulator (IFR scenario 1 then IFR scenario 2) as part of their scheduled (normal) annual training. Both simulator events are instrument check rides and last 2 to 2.5 hours.	Flight simulator: Over the next two days, participants complete two events in a flight simulator (IFR scenario 1 then IFR scenario 2) as part of their scheduled (normal) annual training. Both simulator events are instrument check rides and last 2 to 2.5 hours.
Debriefing: Researchers send an email to debrief each participant on the intent of the study.	Debriefing: Researchers send an email to debrief each participant on the intent of the study.

After all simulator events were done for a given week, a USCG representative would download the week's CAD data for all IFR scenario events, save them to an external memory device, and provide them to the researchers. Each event computer file was only labeled with a time stamp to keep researchers blind to each team's names and training condition.

Two H-65 pilots served as SME raters/evaluators to observe the teams' recorded performance. Both SMEs had over 2,500 flying hours in the H-65 and 3+ years of experience as flight instructors. Both raters received a two-hour block of training in the use of a structured performance evaluation methodology (see *Measures* section). After the raters rated an event, a USCG representative provided the researchers with a master schedule to allow them to correlate missing participant data to each event. This data included the dates of the events, the participant identification codes for the events (to allow the researcher to correlate the pilots' experience and demographic information), and the crews' training group assignment.

In summary, the participants individually completed a training activity, a test, and two questionnaires – all online. Then, paired into teams, the participants completed their scheduled flight training in the simulator. The researchers observed tapes of the aircrews' performance, but were kept blind to which group the aircrews were in until after conducting the review. After the review, the USCG representative shared records showing the pilots and their simulator schedule for the week so the researchers were able to correlate this observational data with the participants' data (training condition, test scores, and questionnaire responses).

Measures

The training was evaluated on all four levels of Kirkpatrick's typology (1976). For reactions outcomes, all pilots receiving the training were surveyed to elicit feedback on the usefulness and likeability of the training. For learning outcomes, a 10-item test was given to both trained and untrained participants. The differences between groups were analyzed. Next, the researchers observed and compared the trained group of pilots (experimental group) and an untrained group (control group) on their performance in a flight simulator. The researchers used this setting to measure whether the learners applied the trained skills, i.e., do or do not the trained

behaviors transfer to performance measures in an on-the-job setting, a measure of Kirkpatrick's *behavior* outcomes (1976). Finally, the researchers recorded whether the crew had or had not met a specific performance standard at the critical point for the climb or descent task – a measure directed at evaluating the task from the perspective of Kirkpatrick's *results* category.

Independent variable - training. The independent variable for this study was the monitoring skill training. This was a between-groups measure in that paired participants (i.e., a single aircrew) were assigned to either the trained (experimental) or the untrained (control) group (Table 6).

Covariate – pilot experience. While not the primary focus, this study provided an opportunity to examine the role of pilot experience in monitoring performance. As previously mentioned, the flying role is transferred frequently between pilots during a flight, so there is not a single pilot who is the PNF for each event. In fact, the time spent in this role was highly variable across flights and is by-in-large directed by the pilot-in-command and their personal “cockpit management” style (J. Cabell, personal communication, November 1, 2010). However, since both pilots are assigned the same training or control condition, team performance was compared by using their pilot designations. For purposes of this study, CPs and FPs were considered to be junior pilots and ACs and IPs to be senior pilots. The research accounted for experience by reporting whether a paired flight crew did or did not contain a junior pilot. Generally, a pilot upgrades to the AC designation after two to three years of experience in the type of aircraft. Additional information was recorded pertaining to the simulator events regarding training latency, which will be explained in the *Discussion* section (Table 6).

Table 6

Independent Variables and Covariates

Variable	How Measured/Levels
Group	1 – No training received (control) 2 – Online training received (experimental)
Designation of junior pilot	1 – CP; 2 – FP; 3 – AC; 4 – IP
Designation of senior pilot	1 – CP; 2 – FP; 3 – AC; 4 – IP
Junior pilot (CP or FP) aboard?	1 – Yes; 2 – No
Training latency – senior pilot	Time between training completion and first simulator event (reported in hours)
Training latency – junior pilot	Time between training completion and first simulator event (reported in hours)

Dependent variables. The researchers developed seven outcome measures to evaluate the effects of the training: *reactions composite score*, *reaction change score*, *post-training test score*, *5-2-1 compliance*, *callout ratio*, *task success rate*, and *overshoot magnitude* (see Table 7). The first three measures were taken from the questionnaire data and the last four used rater observations. Two raters, both SMEs within the USCG, were used for both the 5-2-1 compliance and the callout ratio measures. Before conducting any observations for the research, the raters met with some general guidelines and reviewed two simulator event recordings (the events used for rater training were not included in this study). The raters watched the recordings together with the goal to develop a feasible and consistent method to rate the events. The researcher and the SME worked together to develop two grade sheets, one for each of the two IFR scenarios, to standardize the observation process (see Appendix H). After this process was complete, the raters utilized the grade sheet to rate events independently. Once both sets of observations were complete, the raters compared results. The raters' reliability index (i.e., probability of agreement) is reported in the next section. The raters re-evaluated all results that did not match

and came to agreement on the best score to give. This resulted in a third set of scores which were used in the final analysis. This performance measurement was fairly straightforward in the sense that it is a non-evaluative description of a behavior and the opportunity for subjectivity is remote; however, using two evaluators allowed the researchers to establish a measure of inter-rater reliability and alleviate concerns of experimenter bias.

Table 7

Dependent Variables

Dependent variable	Units of measure
Reactions composite	1 to 5 (participant rating)
Reaction change	-4 to 4 (participant rating)
Test score	0-100%
5-2-1 compliance (2 raters)	0-100%
Non-flying pilot callout ratio (2 raters)	Callouts per minute
Task success rate	0-100%
Overshoot	Feet

Reactions score. Participants rated their reactions to the training across several specific items on a 5-point Likert scale ranging from low (score 1) to high (score 5). The researchers averaged the participants' responses to three of the questions and used this composite score to compare the experimental groups' reactions to the training. The questions included in this measure included: (a) "rate the extent you expect this training will make a difference in the way you do your job," (b) "rate your overall knowledge of monitoring after the training activity," and (c) "rate this training course overall."

The researchers developed a second "change in reactions" measure by taking each participant's rating (from 1 to 5) of their own "overall knowledge of monitoring" after the training and subtracting their response to the same question (using the same scale) referring to

their self-assessed knowledge *before* the training. Scores for this measure could theoretically range from -4 to +4, but the researchers anticipated it would be rare for a learner to report a decrease in their overall knowledge and a score of zero would be the lowest score a respondent would normally give.

Post-training test score. Each participant received an individual score representing the number of correct responses provided on the 10-item post-training test. This number of correct responses was converted into a percentage and reported on a scale from 0 to 100 percent.

5-2-1 compliance. The raters first referenced the instructor guide (i.e., standardized script) for both simulator events of interest to the study and created a list of the vertical flight tasks the crew would be required to perform. For example, after the team takes off (in the simulator), the script might call for the instructor to give the crew a climb to 4,000, followed by a descent to an airport that would require a descent to 500 feet. The researchers would categorize this as two separate vertical tasks. For IFR scenario #1, the researchers measured variables associated with a total of four (4) climb tasks and four (4) descent tasks. For IFR scenario #2, the researchers also measured variables associated with four (4) climbs and four (4) descents. In summary, over two events, each crew pairing would be evaluated on a total of 16 climb or descent tasks, i.e., 16 tasks-of-interest (see Table 8). The *5-2-1 compliance* variable was measured by recording the safety pilots' level of conformance to the 5-2-1 rule taught during the online training course. To calculate this measure, the raters would use the CAD recording to review the final 500 feet of each of the 16 tasks-of-interest. The raters would use a check-off sheet to record whether the PNF made altitude advisory-type callouts at 500 feet, 200 feet, and/or 100-feet prior to arriving at the target altitude, while watching a replay of the event in real time. The total number of 5-2-1 callouts was divided by the total number of opportunities (three per

climb or descent) to determine an overall compliance level for each crew. For example, if the PNF reported 500 and 100 feet prior to level off for a climb task, and later reported 100 feet prior to the minimum descent altitude (MDA) on the subsequent descent, the reported compliance rate for the two maneuvers would be 50% (3 out of a possible 6). The raters agreed that any reference to either the number of feet prior (relative altitude, e.g., “200 to go”) or a report of the actual altitude (e.g., “700 feet for 500 feet”) would constitute an appropriate 5-2-1 callout. (Note: Any SP-initiated callouts during a climb or descent were a voluntary behavior by the SP.) Each crew pairing would receive a score representing their overall percentage of possible 5-2-1 callouts made on the 16 climbs and descents, i.e., the percentage of 5-2-1 callouts the PNF made out of 48 possible callouts (see Table 7). In the case the PF did not reach an anticipated altitude callout checkpoint during a task-of-interest because of an emergency or other problem with the task, the raters would make note of which callout opportunities were lost and subtract this number from the team’s total possible 5-2-1 compliance calls.

In this study, the two raters initially agreed on 93 percent (275 of 296) of their 5-2-1 callout tallies. The raters re-examined the callouts in dispute and were able to come to 100% agreement, and the revised tallies were used for these comparisons.

Callout ratio. The callout ratio measure, by comparison, captured the frequency which non-flying pilot made any altitude callouts during the last 1,000 feet of each task of interest. Using the same 16 specific tasks-of-interest, the raters would, independently, take the CAD recordings from the simulator to the point which the crew was climbing or descending through an altitude 1,000 feet prior to the target altitude, record the time, and press play. At this point, any PNF reference to the aircraft’s altitude until reaching the target altitude would be tallied as a callout. Upon reaching the target altitude, the time would again be recorded, and the total number

of PNF callouts would be recorded for that specific task. The raters would continue to record the number of callouts during the same 1,000-foot window of each of the crew's 15 remaining tasks-of-interest. The total number of callouts would be divided by the total elapsed time. This statistical result would represent the crew's *callout ratio* (reported in callouts-per-minute). Any PNF-initiated callouts during a climb or descent were a voluntary action by the PNF, and are not required per any formal USCG procedure. These callouts provided explicit indications that the PNF was actively monitoring and checking the aircraft's flight path and, at the very least, periodically integrating this responsibility with any other flight-related tasks. For example, if ATC requested a climb to 2,000 feet, every instance the PNF verbalized a reference to the aircraft's altitude to the PF would be recorded. For example, this would include PNF callouts such as "climbing through 1,000 feet for 2,000," "300 feet to go," or "approaching level off altitude." Callouts that were not counted would include those that occurred within 50 feet of level off in cases which the PF had already significantly slowed the climb or descent and the PNF seemed to be helping the PF to fine tune the level off altitude (vice providing an altitude alert callout). Also, a standard call (callout) is required by the PNF upon reaching the lowest altitude of an approach announcing to the PF whether to continue or abort the approach. This call was not tallied as an altitude callout for the callout ratio measure.

After initial review of the simulator data, the raters agreed on 95% (580 of 611) of their callout ratio tallies. The raters re-watched all of the callouts in dispute and were able to come to 100% agreement, and these revised tallies were used to make the above comparisons.

Task success rate. Task success rate was measured by reporting the two-person teams' overall success rate at leveling off at assigned altitudes and at making on-time decisions at the bottom of approach descents. This measure was unique in that while the previous two

measurements were voluntary (i.e., callouts were not required per USCG or other flight doctrine), this measure reported compliance with a mandatory USCG operational procedure. The requirements for crews to meet the “success” criteria during climb tasks and descent tasks were different. All descent tasks-of-interest were already integrated into formal instrument approach procedures at an airport, meaning the crews flying the approach would first descend to a certain altitude then make a decision at that point whether to continue to land (if the pilots have obtained certain visual references with the airport) or to abort the approach. As the aircraft reaches this altitude, depending on what the PNF sees, the PNF calls out a command to either abort the approach (a “missed approach”) or to continue the descent to land (by saying “continue”). To account for this, the raters rated approaches (descents) as successful if: (a) the PNF made the appropriate callout within ± 50 feet of the target altitude, and (b) the PF did not descend more than 50 feet beyond the target altitude if the callout required a missed approach (i.e., aborted approach).

Measuring success at the climb tasks used much more straightforward criterion – raters scored these as successful if the PF did not extend the climb more than 50 feet beyond the target altitude. Given the graded nature of the simulated event and the pilot experience levels, substantial altitude exceedences (called “busts”) were highly unlikely. In terms of federal regulations, FAR 91.123 is the most tolerant aviation standard in stating that an altitude violation does not occur until an aircraft is ± 300 feet of an assigned altitude (Compliance with ATC Clearances and Instructions, 1995). The FAA’s practical test standard for an instrument rating, however, requires that pilots stay within ± 100 feet of all assigned altitudes (FAA, 2004c). With this in mind, the researchers decided a 50-foot tolerance would more effectively differentiate between teams’ performance at these tasks.

For each scenario, the researchers again evaluated *task success* on four climbs and four approaches, for a total of 16 tasks for the two scenarios combined. For each two-person team, the researchers recorded a single percentage score representing the number of successful tasks divided by the total number of tasks.

Overshoot. For each of the 16 tasks, the researchers recorded the number of feet the pilots extended the climb or descent beyond the target altitude. This measure was designed to identify if teams in the trained group were more conscious of not flying beyond the target altitude than the teams in the control group. These overshoots were recorded after each of the tasks-of-interest, then averaged across the total number of tasks the crew performed (max of 16). Overshoots were recorded in increments of five feet according to the CAD display of an analog gauge cockpit altimeter. The total number of tasks-of-interest for the overshoot measure did not include any approach descents in which the aircrew did not have to arrest its descent as they continued to the runway. For example, if the aircrew was supposed to descend to 300 feet for the approach, but the weather was clear once the aircraft descended through 600 feet, the aircrew would most likely elect to continue the descent to the runway without pausing to level the aircraft off at 300 feet. This event would not have been included in the overshoot measure. Fortunately, the way the IFR scenarios were scripted, each team of participants (i.e., aircrew) would experience the same weather at the bottom of each approach, so typically four of the eight total approaches would not include a requirement to level off. Only one rater was used to record this measure, as the aircraft's altitude gauge was easy to read and the measure was considered straightforward.

Table 8

Tasks-of-interest

IFR Scenario 1

- Task 1.1 – ATC instructs climb to 2,000'
- Task 1.2 – PAR approach requires descent to 250'
- Task 1.3 – ATC instructs climb to 4,000'
- Task 1.4 – NDB approach requires descent to 480'
- Task 1.5 – ATC instructs climb to 2,000'
- Task 1.6 – VOR approach requires descent to 460'
- Task 1.7 – ATC instructs climb to 2,000'
- Task 1.8 – ILS approach requires descent to 270'

IFR Scenario 2

- Task 2.1 – ATC instructs climb to 3,000'
 - Task 2.2 – ILS approach requires descent to 560'
 - Task 2.3 – ATC instructs climb to 2,400'
 - Task 2.4 – NDB approach requires descent to 1020'
 - Task 2.5 – ATC instructs climb to 2,400'
 - Task 2.6 – VOR approach requires descent to 820'
 - Task 2.7 – ATC instructs climb to 2,000'
 - Task 2.8 – PAR approach requires descent to 426'
-

Results

Reaction Outcomes

Between-group differences were first examined with respect to the participants' reactions, i.e., self-ratings, of the usefulness and overall impact of the training activity (Table 9). On average, trained participants reacted to the training activity more positively ($M = 4.04$) than the control group ($M = 3.17$). According to a Mann-Whitney test, this difference was significant, $U = 117.5$, $z = -4.40$, $p < .001$, and represented a large-size effect, $r = -.59$.

Further, participants self-rated their knowledge of monitoring concepts both before and after the training (or control) activity on a scale from 1 (low) to 5 (high). Nearly identical results were found between the groups in their ratings of their own monitoring knowledge before the training activity (means of 3.30 and 3.28 for the control and trained groups, respectively). A Mann-Whitney test confirmed the pre-training knowledge scores were not significantly different, $U = 359$, $z = -.29$, $p = .77$. The post-training knowledge scores, however, did vary to a larger extent (mean score of 3.53 for the control group compared to 4.36 for the trained group), and a Mann-Whitney test confirmed the post-training difference was significant, $U = 180$, $z = -3.67$, $p < .01$. To explore these pre- versus post-training differences, two distinct analyses were performed. First, a Wilcoxon's signed-rank test was run to compare the before and after conditions within each group. For the control group, participants' self-ratings of monitoring knowledge were significantly higher after the control activity ($M = 3.53$) than before the activity ($M = 3.30$), $z = -2.33$, $p = .02$. Additionally, for the trained group, self-ratings of monitoring knowledge were significantly higher after the training activity ($M = 4.36$) than before ($M = 3.28$), $z = -4.09$, $p < .001$. A Wilcoxon's test indicated that the effect size was medium ($r = -.43$) for the control activity, but much larger ($r = -.82$) for the trained group. To explore this difference

further, a measure of change in monitoring knowledge was developed by subtracting each participant's "before" knowledge rating from their "after" knowledge rating. The average perceived increase in knowledge was greater in the training group ($M = +1.08$, $SD = 0.76$) than the control group ($M = +0.23$, $SD = 0.50$). This difference represented a significantly greater increase in perceived learning among participants completing the online training group compared to those who completed the control activity, $U = 141.5$, $z = -4.38$, $p < .001$, $r = -.59$, and the effect size was large, (because $r > .5$, per Cohen's benchmarks, 1988).

Participants in the trained group responded to three additional questions pertaining to the quality of content and delivery of the online training program. The participants' responses ranged from 3 to 5 (again on a scale from 1 [low] to 5 [high]) and the average score across these three questions was very high ($M = 4.53$). Open-ended, written questionnaire feedback from the trained group included positive comments including: "that was one of the best put-together surveys I've seen in a long time," and "I recommend all pilots take this survey to heighten their [monitoring] awareness in different regimes of flight." Feedback from the control group included the comments: "survey addresses a very valid topic" and "would make for a good pilot training session."

Table 9

Post-training Reaction Scores

	Condition	
	Control (<i>n</i> = 30)	Experimental (<i>n</i> = 25)
Composite reactions score, <i>from 1 (low) to 5 (high)</i>	3.17 (.81)	4.04 (.44)**
Self-rating of monitoring knowledge before training	3.30	3.28
Self-rating of monitoring knowledge after training	3.53*	4.36*
Reported change in monitoring knowledge	+0.23 (.50)	+1.08 (.76)**
Reported utility of online training, <i>from 1 to 5</i>	--	4.53 (.55)

Note: Means are shown with standard deviations in parentheses.

* indicates a significant difference ($p < .05$) from before training score.

** indicates a significant difference ($p < .001$) from control score.

Learning Outcomes

Fifty-five participants (25 trained and 30 control) completed the post-training 10-item multiple-choice test. Tests were then scored from 0 to 100%. On average, trained participants scored higher ($M = 92.8\%$) on the exam than the control group ($M = 47.7\%$). Test scores ranged from 60% to 100% in the trained group and from 20% to 70% in the control group. All but four of the trained participants scored either a 90 or 100% on the test. Both the distributions of the test scores for trained participants, $D(25) = 0.29$, $p < .05$, and the control participants, $D(30) = 0.18$, $p < .05$, were significantly non-normal. Therefore, a Mann-Whitney test was used to determine if the difference in test scores was significant. This difference was indeed significant, $U = 6.5$, $z = -6.31$, $p < .001$, $r = -.85$ (see Table 10).

Table 10

Post-training Test Results

	Condition	
	Control (n = 30)	Experimental (n = 25)
Post-training test score (0 to 100 percent)	47.7 (13.3)	92.8* (9.80)

Note: Means are shown with standard deviations in parentheses.

* indicates a significant difference ($p < .001$) from control scores.

Behavior Outcomes

As discussed in the methods section, the 44 participants were paired into teams of two and evaluated as 22 teams for the simulator portion of the study. Once assigned to their training conditions, there were 11 trained teams and 11 control teams which the two raters (researchers) rated on their performance in the flight simulator. Before analyzing the potential significance of the individual dependent measures, a one-way multivariate analysis of variance (MANCOVA, between-subjects factor: group [trained, control]; covariate: experience [junior pilot aboard, no junior pilot aboard]) was first performed to determine whether there was a significant effect of training condition across four dependent variables: *5-2-1 compliance*, *callout ratio*, *task success*, and *altitude overshoot*. According to Hotelling's trace statistic, there was a significant effect of training group on these four outcomes, $T = 2.09$, $F(4, 16) = 8.34$, $p < .01$. We also confirmed that the covariate, experience, was not significantly related to the same four outcomes, $T = .23$, $F(4, 16) = .92$, $p = .98$.

5-2-1 rule compliance. The next step we took was to follow up the significant MANCOVA with univariate analyses. Looking first at the *5-2-1 rule* dependent variable, teams in the training group made more 5-2-1 callouts ($M = 43\%$, $SD = 12.9\%$) during the last 1,000 feet of climbs and descents than teams in the control group ($M = 17.5\%$, $SD = 4.6\%$). The range for teams' 5-2-1 callout usage in the training group was from 27% to 65% and the range for 5-2-1

callout usage in the control groups was from 11% to 26%. The researchers applied a one-way analysis of variance (ANOVA) on the 5-2-1 variable. The ANOVA revealed a significant effect of training on 5-2-1 compliance, $F(1, 19) = 37.8$, $MSE = .37$, $p < .001$, $\eta_p^2 = .67$, which supported our hypothesis. The 5-2-1 usage scores did violate Levene's test for equality of variance, however the t -test is considered robust to violations of homogeneity of variance when sample sizes are equal (Levene, 1960).

Callout ratio. The ratio of callouts the PNFs voluntarily made in the simulator ranged from 0.55 to 1.40 callouts per minute for the control group and 0.85 to 1.93 per minute for the trained group. On average, PNFs in the trained group made more callouts per minute ($M = 1.38$) than PNFs in the control group ($M = 0.86$). As a follow-up to the significant MANCOVA, the callout ratio measure was also examined using an ANOVA. This analysis revealed a significant effect of training condition on callout ratio, $F(1, 19) = 17.14$, $p = .001$, $MSE = 1.52$, $\eta_p^2 = .47$, and the hypothesis was supported (see Table 12).

A Pearson correlation analysis across the four in-flight variables was also conducted. With regard to the two "behavior" measures, it was found that *5-2-1 rule use* and *callout ratio* were significantly correlated, Pearson's $r(22) = .86$, $p < .001$. This indicated that teams who made more 5-2-1 callouts would be more likely to have higher callout ratios, an expected finding as there were common elements among these two variables. A list of the correlation coefficients is presented in Table 11.

Table 11

Correlations between Four In-flight Variables (N = 22)

	<i>Behavior Outcomes</i>		<i>Task Outcomes</i>	
	<i>5-2-1 Rule Use</i>	<i>Callout Ratio</i>	<i>Task Success rate</i>	<i>Altitude Overshoot</i>
5-2-1 Rule Use	-	.857**	.243	.026
Callout Ratio	-	-	.503*	-.222
Task Success Rate	-	-	-	-.560**

Note: Pearson's r values are listed.

* $p < .05$. ** $p < .01$

Task Outcomes

The 11 trained teams completed between 80% and 100% of tasks successfully, with a mean success rate of 92.4%. The 11 control (untrained) teams completed between 75% and 100% of tasks successfully with a mean success rate of 89.7%. An ANOVA was performed which revealed a non-significant effect of training condition on task success, $F(1, 19) = .54$, $p = .47$, and the hypothesis was not supported. To analyze an alternative measure of level-off task success, the researchers examined the effect of the training condition on the magnitude of the teams' altitude overshoots at level-off tasks. Although no overshoot may be ideal, in the real world, any overshoot less than 100 feet on any non-approach related vertical maneuver would not generally be cause for concern. The control group exceeded the target altitude on 66.7% of the tasks and averaged 17.7 feet of overshoot on the tasks overall. The trained group also exceeded the target altitude on 66.7% of the tasks but overshoot by a slightly greater margin of 19.3 feet across all tasks. An ANOVA was performed and revealed a non-significant effect of training condition on magnitude of altitude overshoot, $F(1, 19) = .26$, $p = .62$, and the hypothesis was not supported. There were only 16 altitude level-off exceedences greater than 50 feet, which equated to 6.7% of all 241 tasks requiring a level-off. There were four altitude

exceedences greater than 100 feet (1.7% of tasks). The largest overshoot on a single task was 155 feet in the trained group and 130 feet in the control group.

With respect to the “tasks” measures, *altitude overshoot* and *task success rate* were found to be significantly correlated, $r(22) = -.56, p < .01$. This indicated that two-pilot teams who committed larger altitude deviations during level-offs would be less likely to succeed on the level-off tasks, an expected finding as there were common elements among these two variables. A significant correlation was also found between a “behavior” and a “task” category variable – *callout ratio* and *task success rate* were significantly correlated, $r(22) = .50, p < .05$. This indicated that teams who made more callouts-per-minute would be more likely to succeed on the level-off task. None of the other simple linear correlations were significant (see Table 11).

Table 12

Monitoring Outcomes in Flight

	<i>Condition</i>	
	<i>Control (n = 11)</i>	<i>Experimental (n = 11)</i>
5-2-1 Rule Use (percentage)	17.6 (4.61)	43.3* (13.0)
Callout Ratio (per minute)	0.86 (0.26)	1.38* (0.32)
Task Success Rate (percentage)	89.7 (8.62)	92.4 (7.62)
Altitude Overshoot (feet)	17.7 (6.84)	19.3 (8.14)

Note: Means are shown with standard deviations in parentheses.

* indicates a significant difference ($p < .05$) from control scores.

Pilot Experience

A team containing a pilot designated either a copilot (CP) or first pilot (FP) was considered a less experienced team and a team not containing a CP or FP was considered a more experienced team. In this study, there were 11 teams which contained a junior pilot and 11 teams which did not. The teams with a junior pilot made 5-2-1 callouts at a nearly identical rate

($M = 30.4\%$) than teams without a junior pilot ($M = 30.6\%$). Examining the callout ratio measure, teams with junior pilots ($M = 1.11$) were virtually equivalent to teams without junior pilots ($M = 1.12$). Similar results were found in the task success rate and overshoot measures (Table 12). Using a MANCOVA with training group as the independent variable confirmed that the covariate, experience, was not significantly related to simulator behaviors, $T = .23$, $F(4, 16) = .92$, $p > .05$. Because this omnibus test failed to reach significance, this test was not followed up with separate ANCOVAs or t -tests.

Table 13

Callouts by Pilot Experience

	<i>Condition</i>	
	<i>Junior pilot on team (n = 11)</i>	<i>No junior pilot on team (n = 11)</i>
5-2-1 Rule Use	30.4 (14.3)	30.6 (18.7)
Callout Ratio (per minute)	1.12 (0.29)	1.11 (0.49)
Task Success (percentage)	91.3 (7.69)	90.7 (8.77)
Altitude Overshoot (feet)	18.3 (7.17)	18.7 (7.94)

Note: Means are shown with standard deviations in parentheses.

Attitudes Questionnaire

Table 12 shows the participants' responses to each of the 23 questions included in the last section of the questionnaire. Significant differences between groups was rare, which indicated the two groups were generally similar in terms of (a) how they prioritized monitoring among other non-flying pilot responsibilities, (b) what indications they felt were most important to monitor, and (c) when during the flight it was most important to actively monitor. A final question asked the participants to evaluate the need for mandating a critical altitude at which time both pilots should be exclusively monitoring the instruments during a climb or descent. There was a significant difference between the groups as 43% of control participants felt there

was no need for such a guideline while only 8% of trained participants felt the same way ($\chi^2(3, N = 55) = 8.78, p < .05$). The majority of this difference was accounted for in one of the other three responses that stated setting an ‘[exact] number of feet would be appropriate.’ Forty percent of the trained participants chose this answer while only 23% of the control participants did (see Figure 4).

Table 14

Attitudes Questionnaire Results

	Overall (N = 55)	Trained (n = 25)	Control (n = 30)
As PNF, rate the importance of...			
Completing checklists?	6.1	6.0	6.1
Radio communications?	5.5	5.3	5.7
Operating the flight director (autopilot)?	5.2	5.0	5.3
Monitoring the actions of the other pilot?	6.6	6.7	6.5
Monitoring the status of aircraft systems?	6.1	6.1	6.1
Monitoring the aircraft's flight path?	4.9	4.8	4.9
Conducting briefings?	6.5	6.5	6.5
As PNF, rate importance of monitoring each instrument in actual IMC...			
BarAlt (Barometric Altimeter)	6.4	6.2	6.6
RadAlt (Radio Altimeter)	4.8	5.0	4.6
Airspeed Indicator	6.0	5.8	6.2
Heading	6.2	5.9*	6.4
FMS Modes	5.1	4.8*	5.4
ADI (Attitude Direction Indicator)	6.4	6.5	6.4
VSI (Vertical Speed Indicator)	6.2	6.3	6.0
As PNF, rate importance of monitoring during each flight regime...			
Pre-takeoff/taxi:	5.1	5.0	5.1
Takeoff:	6.6	6.6	6.7
Enroute to destination/operational area:	5.4	4.9**	5.7
Approach to landing:	6.6	6.6	6.5
Approach to overwater hover:	6.9	7.0	6.9
Landing:	6.4	6.3	6.4
Hover operations:	6.5	6.3	6.6
As PNF, how vigilant are you when autopilot is engaged?			
	3.9	4.1	3.8

Note: Mean scores are shown. Items were rated on a scale from *no importance* (1) to *extremely important* (7).

* indicates a significant difference ($p < .05$) from control scores.

** indicates a significant difference ($p < .01$) from control scores.

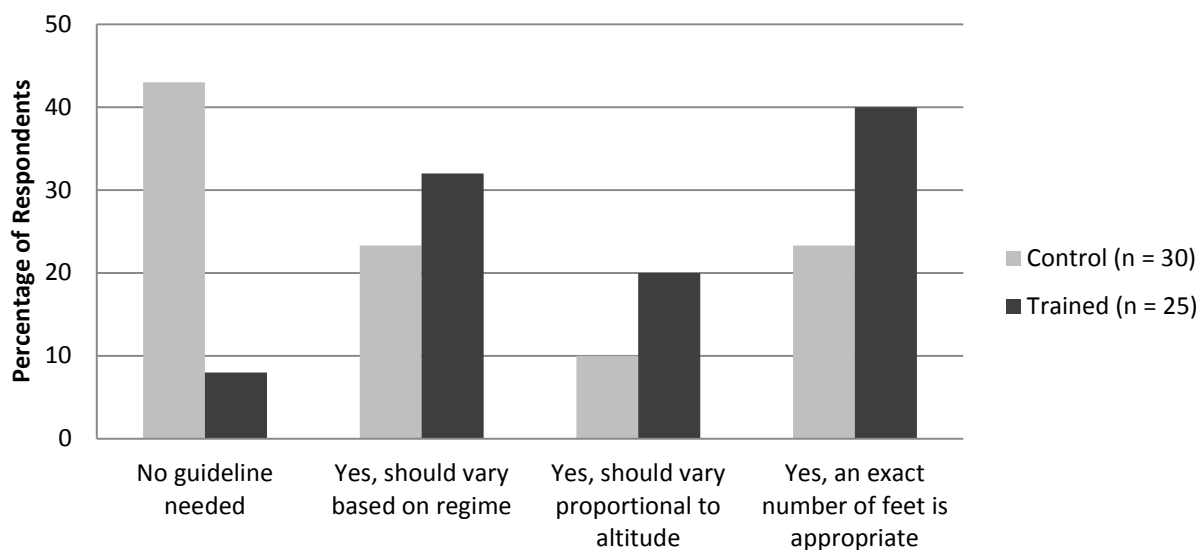


Figure 4. Participant Attitudes towards Mandatory Monitoring Altitudes

Training Latency Analysis

The time interval between participants completing the training and flying the simulator varied from an hour to three full days (1.0 to 72.5 hours), and the average time from training to simulator was about a day and a half ($M = 39.4$; $SD = 27.1$). The interval between the training and the first simulator flight was averaged for each of the two-pilot teams (in the trained group only) and a simple linear regression was run for each of the behavior variables to see if performance decayed as training latency increased. Training latency was used as the predictor variable and 5-2-1 rule usage as the outcome variable in the first analysis. Training latency was not a significant predictor of 5-2-1 compliance, $F(1, 9) = 4.58$, $p = .061$, although this result did approach significance. Similar results were found for the second simple regression. Training latency was used as the predictor variable and callout ratio as the outcome variable, and training latency was also not found to be a significant predictor of callout ratio, $F(1, 9) = 4.83$, $p = .056$.

Discussion

Overview

Our modern understanding of aircraft mishaps has grown to appreciate that shortcomings in pilots' technical flying skills are rarely the cause of accidents – errors of communication and team coordination are the more likely culprits (Gladwell, 2008, p.186). The two-pilot flight deck environment is one in which the flying responsibilities must be strategically shared to enable the crew to perform well as a team. One issue that researchers believe deserves more attention is the pilots' role to cross-check each other's actions as a defense against preventable errors (Sumwalt et al., 1997, 2002; Dismukes & Berman, 2010). The current study examined the effectiveness of an online training program specifically designed to strengthen pilots' voluntary monitoring behaviors during climbs and descents. The participants were instrument-rated helicopter pilots and represented a cross-section of experience levels and geographic locations. The researchers measured changes in participants' knowledge and flight behaviors across all four dimensions of Kirkpatrick's (1976) training evaluation framework as they pertained to pilot monitoring behaviors. The results indicate that a brief online training tool can teach pilots to be more proactive monitors of aircraft altitude during critical flight regimes.

Hypotheses 1-2

Participants in the training group provided highly favorable ratings of the training. In fact, only three participants rated any aspect of the training below a four on a one (1) to five (5) scale, with five, the highest score, being the most frequent response to all questions. Ratings of the online training by the trained group were significantly greater than those rating the control activity, suggesting that the participants thought the training was worthwhile and applicable to their jobs. A second measure of participants' reactions – a comparison of the participants' self-

assessed improvement in knowledge of the subject matter – also rose appreciably in the trained group.

When the researchers tested the participants' knowledge of the training program's learning objectives, the participants in the training group scored extremely well and significantly better than the control group. There are three reasons having these quantifiable results are important. First, the assumption that a training need did exist was validated: the control group's significantly lower test scores were a clear indicator that pilots did not have pre-existing knowledge of the training material. Second, the trained groups' results validated that the test questions were closely correlated to the training content. The consistently high scores (and lack of outliers) among the trained participants indicate this group paid attention to the training and were able to recall information from the training to answer the test questions. Third, these results support the principles behind the cognitive load theory. Given the learners' consistently successful retention of the training material, the efforts to reduce extraneous workload and to increase germane factors on the learners' cognitive processes worked to promote learning of the course objectives. The remaining hypotheses built on these results by testing to see whether the pilots would apply what they had learned in the flight environment.

Hypotheses 3-4

The behavior measures taken from flight simulator recordings also indicate the online training was effective. The scripted flight simulator events allowed for a standardized and informative performance comparison – flight conditions in the simulator (weather and types of emergencies encountered, approaches flown, airports used) were controlled and kept relatively equivalent. The first measure examined the teams' use of the 5-2-1 rule. The measure was developed to see if pilots who were taught a simple memory tool to promote altitude monitoring

would apply it in the flight environment. The 5-2-1 rule was the focal point of the training – it was first explained in the online training session, then participants were given an opportunity to practice it with the assistance of computer-based feedback, if needed. Never did the training program, researchers, or instructors explicitly state that these callouts were required. The rule was presented only as a suggestion to help the pilots fly more safely. Still, performance differences indicated that trained pilots used the 5-2-1 callouts at a significantly higher rate than the control group. This is significant for two reasons: (a) at a conscious level, pilots felt the rule was beneficial to apply in flight, and (b) that pilots considered 500, 200, and 100 feet to be appropriate points during these maneuvers to apply the rule. If either assumption was not true, it is unlikely the pilots would have applied the rule at all in the in-flight (simulator) setting. The statistical predicament in the pilots' use of the 5-2-1 rule as a variable was that the same callouts factored into the crews' overall callout ratios (i.e., second in-flight dependent measure), making the measures almost too highly correlated to warrant the distinction. The data showing a strong positive correlation between 5-2-1 compliance and callout ratio ($r^2 = .73$) was hardly surprising in light of the fact that the 5-2-1 callouts counted towards callout ratios. The two measures were both included from the perspective that (a) these performance measures were highly heuristic given the exploratory nature of the training program, and (b) if the measures did not correlate, this would have had significant implications on the fundamental training design we used. For example, if there were significant differences in 5-2-1 rule usage but non-significant differences in callout ratios, this might have indicated that pilots were already being proactive with altitude callouts, just not in accordance with our specifically-trained 5-2-1 criteria. The researcher's intent, from the flying behavior perspective, was that the rule would both provide structure to the non-flying pilots' duties and increase altitude monitoring during vertical flight regimes.

Broadening the scope of observations, PNF-initiated altitude callouts during the last 1,000 feet of climbs and descents were examined. The online training taught pilots that commercial aviation SOPs require that all non-essential tasks to be suspended within 1,000 feet of any level-off as a counter to the risk of making a level-off error while multi-tasking – in other words, to maintain a “sterile cockpit.” This meant two training “forces” were potentially at work in the trained participants’ minds: (a) the sterile cockpit mentality below 1,000 feet, and (b) the 5-2-1 rule. Neither the USCG’s existing SOPs nor flight operations manuals require that pilots make intermediate callouts at specific altitudes during climbs and descents. The altitude “callout ratio” measure was important to test whether the participants gave verbal indications that they were actively watching the aircraft’s altitude during climbs and descents. For example, an untrained pilot making callouts at 1000, 600, 400, and 200 feet prior to level off would score higher than a trained pilot who ignores callouts until 500 feet and then completes all three 5-2-1 calls. In other words, if the callout ratios had not differed between experimental groups, it may have indicated the strategy of using randomly arranged callouts was an equal or more effective method of altitude monitoring than 5-2-1 rule, rendering the 5-2-1 rule unnecessary. The number of voluntary callouts participants made per minute, however, did increase in the trained group. This confirms the hypothesis regarding the effect the online training had on aircrews’ callout ratios and strengthens the overall effect the training seemed to have. Without conducting a separate experiment to tally the locations where (number of feet prior to level off) pilots’ made altitude callouts during climbs and descents, it is impossible to say if the significant differences in callout ratio were only an effect of the 5-2-1 rule training or indicative of something that had transferred from the general training emphasis on altitude awareness, i.e., one which had an effect on pilots’ behaviors even before they reached the 500-feet-prior point. As one participant

reported, “I had never thought of using specific altitudes... I normally just make random callouts.”

Responses to one of the survey questions shed some light on reasons for the group differences. When the participants were asked how many callouts they would make during the last 500 feet of a climb or descent, participants in the control group, on average, suggested they would make 2.5 callouts ($n = 30$). By comparison, the trained participants exclusively reported they would make three callouts during the final 500 feet (note: this was consistent with the 5-2-1 rule contained within the training and was unlikely a pre-existing difference). This further confirms that after receiving training, the non-flying pilots were mentally prepared to make more frequent altitude callouts during the final few hundred feet prior to level off.

There were other indications that pilots' altitude monitoring did extend beyond tacit acceptance of the first 500 foot callout recommended by the 5-2-1 rule. Were trained crews actually putting other tasks on hold during the final stages of vertical phases of flight, freeing mental capacity for focusing their attention on monitoring altitude and other instruments during the level-off task? Observations indicated a small percentage of crews (two of the trained teams) explicitly state they were putting all other cockpit tasks on hold during the entire 1,000 foot window, as the airlines already do. From an aviation safety perspective, it was a positive sign to hear aircrews who had received the training explicitly state they were putting the checklists on hold at the 1,000 feet-to-go point of a climb until the level-off task was accomplished. A comment like this suggests that the behavioral transfer of the training material, at least in some cases, extended beyond the simple application of the 5-2-1 rule.

Hypothesis 5

Contrary to this hypothesis, no statistically significant support was found to indicate the online training affected teams' bottom-line flight task outcomes. Task success was looked at as a measure to see if the users of this online training would change their cross-checking behaviors to such an extent they achieved a higher level of proficiency at vertical flight tasks. This was not the case, as teams in both groups demonstrated similar proficiency at following procedures and backing up the flying pilots as the aircraft leveled off. There was one major hindrance to the online training having an effect on this outcome, and it came from using a particularly motivated group of participants for this study. Going into each IFR simulator event, the pilots knew they were receiving their annual graded instrument flight proficiency check ride. These check flights held a high level of personal and professional significance to the pilots, and as a result, the participants would have been highly motivated. Pilots typically study and prepare for several days (or longer) for this flight. If a participant was to fail this flight, there would be negative and professionally embarrassing consequences. Further, the approach tasks-of-interest themselves were of significant importance to passing these flights – it is a basic aviation tenet not to descend below a minimum descent altitude unless the airport is in sight. The “results” measures that were used, unlike the previous “behavior” measures (Kirkpatrick, 1976), were ones which the USCG instructors were also evaluating the participants on, and the participants knew this.

Results from the altitude overshoot measure were also non-significant, and did not support the hypothesis. Although this was not what was hypothesized, the measure of overshoot was an interesting one as some techniques were discovered which pilots used to level off. In particular, it was found that pilots preferred using automation to level off the aircraft whenever possible. The overshoot measure became more a function of which one of three different

techniques the pilots used to level-off the aircraft than a measure of the focus the pilots were directing at the altimeter readings just prior to leveling off. Using the first level-off technique, a.k.a. the 'hand fly' method, pilots would manually fly the aircraft during the climb or descent, slow the rate of climb or descent as they approached the level off altitude, and zero out the vertical speed as they arrived at the target altitude. This technique was mostly observed in simulator events in instances when the autopilot was made inoperative (for training) by the instructor. Survey data shows participants were hesitant to manually fly the simulator and more reliant on the autopilot than they would have been in the actual aircraft. The second and third techniques involved using the autopilot (automation) during the entire climb or descent and level-off. The difference between the second and third techniques was a measure of the dependence the PF placed in automation to do its job. Using one technique (one more often used in climbs), pilots would keep the rate-of-climb or rate-of-descent established for the entire vertical maneuver and then call for the PNF to engage the altitude hold mode as they passed the target altitude. During a climb for example, this technique would mean the autopilot would capture while the aircraft was still climbing. Since the computer captures the target altitude at the moment the button is depressed, the autopilot computer would need 25 to 50 feet to arrest the aircraft's climb before it began to descend back down to the proper altitude. Using an alternative autopilot method, pilots would use the autopilot to slow the rate of climb or descent (usually about 100 to 200 feet prior to level off) and make a more controlled and gradual approach to the target altitude. Using this method, the PF would have slowed the climb or descent to less than 200 feet-per-minute (fpm) before having the PNF engage the "altitude hold" autopilot mode. This technique was more common to see pilots using during approaches and descents. Because about half of the teams used the former autopilot method (to accept a 25 to 50 foot overshoot as

part of allowing the autopilot to handle the level off), the overshoot measure data tended to be a representation of this technique. In fact, the strong and significant negative correlation between overshoot and the task success rate ($r^2 = .31$) confirmed this – as the teams' average overshoot variable increased, their task success rates decreased. This result was anticipated and attributed largely to the fact that overshoots greater than 50 feet were counted as single-task failures in the teams' overall task success rate score. Ultimately, neither of these “task outcomes” measures was able to identify an effect on the teams' bottom line performance at vertical flight maneuvers in the flight simulator.

The correlation analysis provided some encouraging results. The moderate-to-strong correlation between callout ratio and task success rate ($r^2 = .25$) meant that the teams (regardless of group membership) making more callouts during in-flight tasks were also more successful at the level-off tasks. This is an important finding in light of the prior discovery of a significant main effect of the online training on teams' callout ratios. This is a subtle indication that there may be a link between online training (independent variable) and task success (outcome variable), and the only tangible evidence of a training effect at the organizational (i.e., Kirkpatrick's “results”) level. This is an important finding from the perspective of developing content validity for these types of measures as good, valid measures of “monitoring skills”. Had the tasks not been completed by considerably experienced and proficient pilots flying a scripted check-flight, the result might have increased our chance of detecting a difference between the training groups with regard to their task success rates.

Hypothesis 6. It was expected that teams with junior pilots would make fewer callouts than teams with senior pilots, however, this hypothesis was not upheld. Previous research indicates that junior pilots tend not to be as assertive as senior pilots (Jentsch et al., 1997) and

that crew pairings containing senior pilots paired with very junior pilots are involved in more mishaps (NTSB, 1994). These were indirect indications that monitoring skills may improve with pilot experience. In this study, crew performance was measured in pairs of pilots and, therefore were unable to distinguish between the callouts made by each pilot on an individual basis. In other words, within a team's total callout ratio, it is impossible to know which pilot, if either, made more callouts. These performance indications were relatively identical across the four dependent measures regardless of pilot experience level, and none even approached significance. Because individual data was not gathered on junior pilots, the only conclusion that can be drawn from this measure is that the data did not give any preliminary indications that there is an effect of experience on monitoring performance. From a training design perspective, it is a positive sign that junior and senior pilots have similar baselines monitoring performance levels and seem to be similarly receptive to the training. Still, the scope of this comparison in terms of overall pilot monitoring was very narrow (altitude callouts only) and did not relate to monitoring for deviations, which is the primary area of concern regarding the monitoring performance of junior pilots. Future studies would benefit from finding better ways to control for pilot experience in order to draw more statistically useful comparisons.

Helicopter “Areas of Vulnerability”

In the background section, the author described that industry analysts had developed an “areas of vulnerability” (AOV) diagram used to depict to pilots the highest workload flight regimes to help pilots' awareness of error vulnerability (Sumwalt et al., 2002; see Figure 1). The diagram stresses the importance of both pilots maintaining their awareness during vertical phases of flight. The AOV diagram was written with long-haul, commercial aviation aircrews in mind. Since this study instead used military helicopter pilots, it is worth discussing some of the

differences in the flight regimes and the implications for aircrew errors. The first major difference is that most helicopters stay below 10,000 feet MSL (i.e., low-altitude flight). In commercial aviation, by comparison, all attempts are made to minimize flight time below 10,000 feet MSL, mostly because there is a much greater amount of civil aviation traffic flying at these altitudes. Along these lines, the current AOV chart discusses “transition altitudes,” an altimeter change procedure which only applies to aircraft flying over 18,000 feet MSL. Any mention of flight over 10,000 feet and transition altitudes would be easy to remove in a discussion of AOV for helicopters.

A more interesting discussion may be whether “1,000 feet prior to any level off” is an appropriate area of vulnerability in helicopter IFR operations. After all, helicopters typically operate at lower altitudes, make much smaller magnitude altitude adjustments, and use smaller vertical speeds ($\pm 1,000$ fpm) to maneuver vertically. To adhere to a typical 3 degree approach glideslope, a helicopter would not typically exceed a 600 fpm rate-of-descent. A commercial jet, by comparison, routinely exceeds 3,000 fpm for enroute climbs and descents, a number that decreases to around 1,000 to 1,500 fpm during approach descents. So the task of monitoring a commercial jet’s altitude may take as little as 15 to 20 seconds during the last 1,000 feet of a climb and as much as 60 seconds during an approach descent. According to the helicopter simulator data from this study, 1,000 foot climbs in the helicopter simulator took just over one minute to complete on average and the same 1,000 foot approach descent typically took crews over two minutes. The questionnaire in this study asked participants if there was a specific altitude prior to level off that would be an ideal time for both pilots to focus on the level off task. Among those who agreed there should be a standard, 500 feet was the most frequent response, and 82% of the responses were equal to or less than 500 feet ($n = 17$). Five hundred feet would

equate to 30 seconds of a climb and 60 seconds of an approach, much closer to the commercial aviation standard used in the AOV chart. Accordingly, this research suggests that a more appropriate area of vulnerability for helicopter pilots during vertical maneuvers would be within 500 feet of the assigned level-off altitude. This number is more aligned with both the survey data from pilots and the reduced severity of risk (inherent to the slower airspeeds and lower rates of descent) associated with helicopter IFR operations. A more thorough analysis of accident or operational flight data would be necessary to draw any more conclusions regarding how “areas of vulnerability” affect helicopter operations differently than commercial ones. From an instructional perspective, it is highly worthwhile to continue to explore these differences to improve training programs and develop aircraft-specific AOV charts.

Limitations

Although this study utilized a fairly simple experimental design, the execution and nature of the experiment was susceptible to several limitations. Many of these limitations arose from using real world observational data which would not have existed had we used a tightly controlled laboratory setting. The limits of the field study can be categorized into the following categories: selection bias, efficacy of training, missing data, and experimenter bias.

Selection bias. As in most field studies, many experimental factors were outside the control of the experiment. For example, a preferred experimental design is to use a pre- and post-test experimental design, not a post-test only design as was performed here. This was, as the case with some of the other limitations, primarily a function of the access the researchers were granted in this field study. Without having pre-test scores to compare our results to, the central threat to the validity of this study is that a selection bias existed. In other words, what assurance did the researchers have that our sample of trained participants were equivalent or

comparable to the control group before the study in terms of the behaviors of interest to this study? Selection bias states that when non-random assignment exists (e.g., pre-assigned groups were used for this study), it is difficult for the researchers to ensure the groups are identical. Pilot experience is considered a characteristic that may correlate significantly with the dependent variables at play (Bordens & Abbott, 2008), which suggests that for studies using pilots as participants is that having similar levels of flight experience within each group may actually be preferable to a lopsided, yet random distribution of flight experience. In this case, we gathered several measures of flight experience information from the participants to investigate these potential types of critiques. As researchers, a decision was also made not to randomly assign pairs to training groups, as it made more sense in the bigger picture to apply an alternating, week-on, week-off systematic schedule to assign participants into their groups. It was felt that alternating the groups' treatments week-to-week reduced the chance of diffusion bias between the groups, as would have been the case had the participants discussed the training during their social time together. This non-random assignment method also was accepted by ensuring the two groups did not significantly differ across several demographic variables. In fact, the groups were very similar across demographic categories and pilot experience levels (see Tables 1 and 2).

Efficacy of online training. A major critique of any training research is that the researchers cannot control for learners' level of motivation. With online, asynchronous training, it is even more difficult to ensure learners pay attention to the training program. For this study, testing the efficacy of the training was accounted for from the earliest stages of its development. Beyond the use of SMEs during the development phases, the content itself was designed to be engaging by providing the trainees with feedback. It is assumed that this feedback played a role

in the effectiveness of the training observed in the reactions, learning, and behavior outcomes. During this study, there were no reports of software failures or confusing directions within the program itself which would have discouraged participants from completing the training. To further address this limitation, this study's results show both overwhelmingly positive participant reactions and post-training test results as quantifiable measures of the trainings' efficacy.

Training latency. Another limitation in terms of the training's efficacy was the variation in the amount of elapsed time between the participant taking the training and the start of the flight simulator events. The time interval between participants completing the training and flying the simulator varied from an hour to three full days. The reason for the large variation in training latency was because the USCG requested the researchers to use voluntary participants and operate on a "low-interference" basis. In order to respect the need of students to prepare for the normal events (tests, lectures, and check flights) during the week, the researchers made the training available online starting the day the students arrived at the proficiency course (Sunday) with the explanation that the training could be taken any time prior to their first scheduled simulator event (Tuesday, and in a few cases, Wednesday). This method created a limitation, but does suggest that the results would have more consistent and reliable if the researchers had required participants take the training within a more restricted time period prior to the simulator events. Had the researchers arranged for a longer period of time to accumulate data, this method might have been feasible. The other option would have been to treat latency as another independent variable or covariate, but the logistics of this study and small sample size here did not allow for this. We do not assume these results tell suggest whether long-term monitoring behaviors are impacted by this training – obviously training effects are susceptible to decay. Future research should be expanded into longer-term studies of monitoring skill (or any other

non-technical skill) retention. This study merely provides a starting point for studies into the effectiveness of online, non-technical skill training programs that test the duration of the training effect both immediately and then three, six, or 12 months after the initial training treatment.

Incomplete data. This research was conducted in a real-world military training environment and used a high-fidelity simulator. This applied setting was a rare research opportunity and served as a valuable setting to conduct this research – it enhanced the overall external validity of our hypothesis which stated that online training can be an effective training method. The flight simulator was conducive to testing representative, intact pairs of pilots in an appropriately realistic, yet safe environment.

Unfortunately, not every group in the study could be observed across all 16 tasks-of-interest. In one case, the recording of an event began a few minutes late (when the instructor remembered to turn on the recording device). In another case, the recording ended early because the maximum recording length for an event was exceeded. This happened because an instructor did not stop and reset the CAD recording system between two separate events. The other form of missing data would occur if either an aircrew aborted an approach for some legitimate reason (e.g., an emergency situation) or an instructor (to increase workload and test pilot decision making for training purposes) changed an altitude assignment task in the middle of a climb. In these cases, the researchers changed the appropriate denominators in the 5-2-1 callouts, task success, and overshoot measures to reflect the missing data. For callout ratio, the total number of seconds providing callout “opportunities” would have been slightly less for the teams over the two events if an event was missed. In cases where the crew was, within a few hundred feet of leveling off when the task was changed or aborted, the data up to that point was used. For example, if the in-flight task changed from a climb to a descent but the pilots were already within

150 feet of leveling off from the climb, the clock would stop, and callouts would be averaged over the time it took the aircraft to climb from 1,000 feet prior to 150 feet prior to level off. Also, only the opportunities for 500 and 200 foot prior callout would be used and the team would get a 5-2-1 score for that task out of two instead of three. One team included in this study completed 13 tasks, one team completed 14 tasks, and five teams completed 15 tasks. None of the teams used in the final results completed less than 13 total tasks.

Instruction differences. Another limit of the study is that each simulator event is an actual training event and there were inconsistencies in the workload imposed by instructor that were outside the control of the experiment. Instructors do follow a script, but are empowered to insert simulated emergencies or other distractions at random points throughout the flight to test the pilots. This would add to the workload at certain points for the crew and may confound results of the crews' monitoring performance. For example, if an instructor activated an engine fire warning light as the aircraft approached a level-off point, the crew would be less engaged (justifiably) in the monitoring task. Another method of instructor inference affecting pilot workload would be for the instructor pilot to discuss the participants' performance on a previous task during the subsequent tasks, potentially affecting the participants' performance ratings on the next task. Still, one could argue that this lack of experimental control actually promoted the accrual of more realistic monitoring performance data within a wide range of environments.

Experimenter/measurement bias. Despite having planned for and developed a straightforward measurement procedure, the researchers knew two raters would be required to observe the tasks-of-interest and ensure the callout tallies were consistent. Also, to ensure they were kept impartial, the raters were kept blind to the condition they were watching. The use of two raters was beneficial and allowed the SME raters to identify ambiguous moments and come

to agreement. When watching and listening to the recordings, however, there were several instances where radio calls would be “stepped on,” other instances where it would be very difficult to tell which pilot had made the callout, or what exactly they said with a high level of certainty. An example would be a pilot saying, “*one hundred*,” and the rater having to listen to context at that point in the recording to determine if the pilot meant 100 feet, 100 degrees of heading, or 100 knots. A final issue was that some pilots would specifically mention the 5-2-1 calls leading up to a level-off, which clearly identified that team as belonging to the trained group and opening up the researchers to potential for experimenter bias.

Conclusion

In recent years, researchers have suggested there is a need for non-technical aviation training that is “effective, standardized, widely distributable, and affordable” (Kearns, 2010). The challenge for training pilot monitoring and cross-checking skills is two-fold in that (a) these programs still lack attention within the industry, and (b) researchers have not explored which delivery methods – online, classroom, or blended; asynchronous or synchronous – would be most suitable for pilot training. The current study tested the efficacy of a relatively brief, inexpensive training program on one aspect of monitoring (a non-technical flight skill): the 5-2-1 rule. The evaluation included trainee reactions, knowledge, behavior, and task outcomes. Training effects were evident in the first three levels and not the fourth. We can therefore infer that just because we did not see changes at the results level does not necessarily mean that the training did not affect the “results” level, i.e., the organization’s bottom line. Future research into monitoring and other non-technical skill training programs should continue to seek unique and improved ways to measure results.

Despite the lack of results indicating a change in the task outcome measures, the other results were very promising. This study successfully showed a positive transfer of pilot behaviors from an online training program to an operational environment. According to D. L. and J. D. Kirkpatrick (2006), for change to happen a learner must: (a) desire to change, (b) know what to do, (c) be in the right climate to facilitate the change, and (d) be rewarded for changing. Since this study observed measureable changes in pilot behaviors, it can be assumed the online training program influenced the learners at least to some degree along these criteria. Further, in light of the results, these methods deserve consideration from aviation organizations – particularly the paradigm of using observational data from recorded simulator events in making data-driven decisions regarding on the utility of their training programs.

Recommendations

This research supports the use of low-cost, online training to advance the training of non-technical skills such as pilot monitoring. The elements of this online training considered to be the most effective include the self-paced nature of the training, allowing participants an opportunity to practice what they learned, and the use of SME input in its development. Along these lines, future studies are warranted to dissect which aspects of online training was most beneficial to the learners. For example, was it the flexibility to take the training at a convenient time that made the 5-2-1 rule so agreeable to the learners or was it the opportunity to practice the rule in some actual scenarios? From the perspective of the need to train pilots to be more active monitors, more research is needed into the usefulness for a memory aid like the 5-2-1 rule. This rule was devised for training, but did not generate any feedback whether pilots or organizations would accept a rule like this in the long-term. A survey study could be very useful here, but the more important finding would be to conduct more field research to determine if applying a 5-2-1

does produce some bottom-line performance improvements at vertical flight tasks, i.e., “results,” per Kirkpatrick’s typology – the types of results we did not see here but assumed some level of a ceiling effect from the significance of the training event to the participant.

Getting back to the impetus for this research, the results here do suggest we should apply some of the general monitoring concepts and training design criteria used for this study back to fixed-wing aircraft and the realm of commercial aviation. The resources within commercial aviation for exploratory studies using flight simulators would advance our understanding of what monitoring strategies to train and how to train them. As monitoring research continues to mature, there would hopefully be discussions of the critical role of the monitoring pilot into the CRM training as well. Given the fact that many aviation multi-crew fatal accidents occur in part due to the pilots’ shortcomings in monitoring responsibilities, this is certainly a training domain that is increasingly relevant to safer aviation operations.

References

- Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Walseth, P. A., Fiset, M., & Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research, 74*(3), 379-439.
- Besco, R. O. (1995). Releasing the hook on the copilot's catch 22. In *Proceedings of the 39th Annual Meeting of the Human Factors and Ergonomics Society*, pp. 20-24, San Diego, CA.
- Bordens, K. S. & Abbott, B. B. (2008). *Research design and methods* (7th ed.). New York: McGraw-Hill.
- Brandimonte, M. A., Ferrante, D., Feresin, C., & Delbello, R. (2001). Dissociating prospective memory from vigilance processes. *Psicologica, 22*, 97-113.
- Brandimonte, M.A. & Passolunghi, M.C. (1994). The effect of cue-familiarity, cue-distinctiveness, and retention interval on prospective remembering. *The Quarterly Journal of Experimental Psychology, 47*(3), 565-587.
- Brannick, M. T., Prince, C., & Salas, E. (2005). Can PC-based systems enhance teamwork in the cockpit? *International Journal of Aviation Psychology, 15*(2), 173-87.
- Clark, R. C., & Mayer, R. E. (2006). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning* (2nd ed.) San Francisco, CA: Pfeiffer.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). New York: Academic Press.
- Compliance with ATC Clearances and Instructions, 14 C. F.R. § 91.123 (1995).

- Curzio, J. C. G. & Arroyo, C. (2010, April). *Pilot monitoring training*. Paper presented at First Pan American Aviation Safety Summit, Sao Paulo, Brazil. Retrieved from <http://www.mexico.icao.int/summit/Day%204%20-%20Thursday/>
- Dismukes, R. K. (2007). Prospective memory in aviation and everyday settings. In M. Kliegel, M. A. McDonald, & G. O. Einstein (Eds.) *Prospective memory: cognitive, neuroscience, developmental, and applied perspectives* (pp. 411-431). New York: Taylor & Francis.
- Dismukes, R. K., & Berman, B. A. (2007, March). *Checklists and Monitoring: Why Two Crucial Defenses Against Threats and Errors Sometimes Fail*. Paper presented at the 2007 Human Factors in Aviation Conference, San Antonio, TX. Retrieved from humanfactors.arc.nasa.gov/flightcognition/Publications/HFAvnConfSlides_FINAL%201.ppt
- Dismukes, R. K., & Berman, B. A. (2010). Checklists and monitoring in the cockpit: Why crucial defenses sometimes fail, *NASA Technical Memorandum (NASA TM-2010-216396)*. Moffett Field, CA: NASA Ames Research Center.
- Dismukes, R. K., Berman, B. A., & Loukopoulos, L. D. (2007) *The limits of expertise: Rethinking pilot error and the causes of airline accidents*. Aldershot, UK: Ashgate.
- Dismukes, R. K., Loukopoulos, L. D., & Jobe, K. (2001). The challenges of managing concurrent and deferred tasks. In R. S. Jensen (Ed.), *Proceedings of the 11th International Symposium on Aviation Psychology*, Columbus, OH: Ohio State University.
- Dismukes, R. K., Young, G. E., & Sumwalt, R. L. (1998). Cockpit interruptions and distractions: Effective management requires a careful balancing act. *ASRS Directline, 10*, 4-9.
- Ellis, J. A. (1996). Prospective memory or the realization of delayed intentions: A conceptual framework for research. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.),

- Prospective memory: Theory and application* (pp. 1-22). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97-101). Santa Monica, CA: Human Factors Society.
- Endsley, M. R. (1995). Toward a theory of situation awareness. *Human Factors*, 37(1), 32-64.
- Federal Aviation Administration (FAA). (2003). *Advisory Circular No 120-71A: Standard Operating Procedures for Flight Deck Crewmembers*. Washington, DC: Author.
- Federal Aviation Administration. (2004a). *Advisory Circular No 120-51E: Crew resource management training*. Washington, DC: Author.
- Federal Aviation Administration. (2004b). *Advisory Circular No 120-35C: Line Operational Simulations: Line Oriented Flight Training, Special Purpose Operational Training, Line Operational Evaluation*. Washington, DC: Author.
- Federal Aviation Administration. (2004c). *Instrument rating practical test standards* (FAA-S-8081-4D). Washington, DC: Government Printing Office.
- Federal Aviation Administration. (2006). *Advisory Circular No 120-90: Line Operations Safety Audits*. Washington, DC: Author.
- Fioratou, E., Flin, R., Glavin, R., & Patey, R. (2010). Beyond monitoring: distributed situation awareness in anaesthesia. *British Journal of Anaesthesia*, 105(1), pp. 83-90.
- Fischer, U., & Orasanu, J. M. (2000). Error-challenging strategies: Their role in preventing and correcting errors, In *Proceedings of the 41st Meeting of the Human Factors and Ergonomics Society*. San Diego, CA: Human Factors Society.

- Flight Safety Foundation (2005, February). Line operations safety audit (LOSA) provides data on threats and errors. *Flight Safety Digest*, 24(2), 1-18.
- Flight Safety Foundation (2010). ALAR Toolkit. Retrieved from <http://flightsafety.org/current-safety-initiatives/approach-and-landing-accident-reduction-alar/alar-tool-kit-and-resources>.
- Flin, R., Martin, L., Goeters, K., Hoermann, J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety*, 3, 95-117.
- Foushee, H. C., & Helmreich, R. L. (1988). Group interaction and flight crew performance. In E. L. Weiner & D.C. Nagel (Eds.), *Human factors in aviation* (pp. 189-227). San Diego, CA: Academic Press.
- Gagne, R. M. (1984). Learning outcomes and their effects: Useful categories of human performance. *American Psychologist*, 39, 377-385.
- Gladwell, M. (2008). *Outliers: The story of success*. New York: Little, Brown and Company.
- Guynn, M. J. (2008). Theory of monitoring in prospective memory. In M. Kliegel, M. A. McDonald, & G. O. Einstein (Eds.) *Prospective memory: cognitive, neuroscience, developmental, and applied perspectives* (pp. 53-76). New York: Taylor & Francis.
- Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 3-45). San Diego, CA: Academic Press.

- Helmreich, R.L., Klinec, J.R., & Wilhelm, J.A. (1999). Models of threat, error, and CRM in flight operations. In *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 677-682). Columbus, OH: The Ohio State University.
- International Civil Aviation Organization (ICAO). (1994). *Safety Analysis: Human Factors and Organizational Issues in Controlled Flight Into Terrain (CFIT) Accidents, 1984–1994*. Montreal, Quebec: ICAO.
- Jensen, R. S. (1995). *Pilot judgment and crew resource management*. Ashgate Publishing, Aldershot.
- Jentsch, F., Martin, L., & Bowers, C. (1997). *Identifying critical training needs for junior first officers*. (Tech. Report for the Naval Air Warfare Center Training Systems Division and the Office of the Chief Advisor for Human Factors at the Federal Aviation Administration). Orlando, FL: University of Central Florida.
- Jentsch, F., Barnett, J., Bowers, C. A., & Salas, E. (1999). Who is flying this plane anyway? What mishaps tell us about crew member role assignment and air crew situation awareness. *Human Factors*, 41(1), 1-14.
- Jones, R. A., & Endsley, M. R. (1995). Investigation of situation awareness error. In *Proceedings of the 8th International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University.
- Kearns, S. K. (2009). e-CRM: The advantages and challenges of computer-based pilot safety training, In *Proceedings of the 53rd Meeting of the Human Factors and Ergonomics Society*. San Antonio, TX. Human Factors and Ergonomics Society. 1569-1573. doi: 10.1518/107118109X12524444080233
- Kearns, S. K. (2010). *E-learning in aviation*. Aldershot, UK: Ashgate.

- Khatwa, R. & Helmreich, R. L. (1998). Analysis of critical factors during approach and landing in accidents and normal flight: Data acquisition and analysis working group final report. *Flight Safety Digest, Killers in Aviation: FSF task force presents facts about approach and-landing and controlled-flight-into-terrain accidents*. Nov-Dec 1998, Jan-Feb 1999, 1-77.
- Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resources development* (pp. 18.1–18.27). New York: McGraw-Hill.
- Kirkpatrick, D. L. & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). San Francisco: Berrett-Koehler.
- Klinec, J. R., Wilhelm, J. A., & Helmreich, R. L. (1999). Threat and error management: Data from line operations safety audits. In R.S. Jensen (Ed.), *Proceedings of the 10th International Symposium on Aviation Psychology* (pp. 683-688). Columbus, OH: The Ohio State University.
- Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science*, 47(5), 693-706.
- Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology* 78(2), 311-328.
- Levene, H. (1960). Robust tests for equality of variance. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

- Lindblom, R. A. (2007, April). The two-challenge rule. *Flying Safety Magazine*, 63(4), 26-27.
Retrieved from <http://www.afsc.af.mil/information/magazinearchive/fsm07.asp>
- Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2003). Concurrent task demands in the cockpit: Challenges and vulnerabilities in routine flight operations. *Proceedings of the 12th International Symposium on Aviation Psychology*, (pp. 737-742). Columbus, OH, USA.
- Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009). *The multitasking myth: Handling complexity in real-world operations*. Aldershot, UK: Ashgate.
- Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009, August). The perils of multitasking. *AeroSafety World*, 4(8), 18-23. Retrieved from <http://flightsafety.org/aerosafety-world-magazine/>
- National Transportation Safety Board (NTSB). (1994). *Safety study: A review of flightcrew-involved, major accidents of U.S. air carriers, 1978 through 1990* (NTSB/SS-94/01). Washington, D.C.: National Technical Information Service.
- National Transportation Safety Board. (2007). *Crash During Approach to Landing, Circuit City Stores, Inc., Cessna Citation 560, N500AT, Pueblo, Colorado, February 16, 2005*, Aircraft Accident Report NTSB/AAR-07/02 (Washington, DC: NTSB, 2007).
- National Transportation Safety Board. (2010). *Loss of Control on Approach, Colgan Air, Inc., Operating as Continental Connection Flight 3407, Bombardier DHC-8-400, N200WQ, Clarence Center, New York, February 12, 2009*, Aircraft Accident Report NTSB/AAR-10/01 (Washington, DC: NTSB, 2010).
- Noe, R. A. (2004). *Employee training and development* (3rd ed.). New York: McGraw-Hill.

- O'Connor, P., Hörmann, H.-J., Flin, R., Lodge, M., Goeters, K.M., & The JARTEL Group (2002). Developing a method for evaluating CRM skills: A European perspective. *International Journal of Aviation Psychology*, 12(3), 263-286.
- Orasanu, J. M., Murray, R., Rodvold, M., & Tyzzer, L. K. (1999). Has CRM succeeded too well? Assertiveness on the flight deck, In R. Jensen (Ed.), *Proceedings of the Tenth International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University.
- Orlady, L. M. (2010). Airline pilot training today and tomorrow. In B. Kanki, R. Helmreich and J. Anca (Eds.), *Crew resource management* (2nd ed.) (pp. 469-492). Amsterdam: Elsevier.
- Orlady, H. W., & Orlady, L. M. (1999). *Human factors in multi-crew flight operations*. Burlington, VT: Ashgate.
- Pass, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1), 1—4.
- Prince, C., Salas, E., Brannick, M., & Prince, A. (2010). The influence of experience and organizational goals on leadership in the military cockpit. *The International Journal of Aviation Psychology*, 20(4), 375 – 389.
- Raisinghani, M. S., Chowdhury, M., Colquitt, C., & Reyes, P. M. (2005). Distance education in the business aviation industry: Issues and opportunities. *International Journal of Distance Education Technologies*, 3(1), 20-43.
- Rosenberg, M. J., (2001). *e-Learning: Strategies for delivering knowledge in the digital age*. New York: McGraw Hill.

- Salas, E., Burke, C. S., Bowers, C. A., & Wilson, K. A. (2001). Team training in the skies: Does crew resource management (CRM) training work? *Human Factors*, 43(4), 641–674.
- Salas, E., & Cannon-Bowers, J. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 52, 471–499.
- Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors*, 48(2), 392-412.
- 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, 189–206.
- Sarter, N., Mumaw, R., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on highly automated flight decks: An empirical study combining behavioral and eye tracking data. *Human Factors*, 49, 347–357.
- Sumwalt, R. L. (1999). Enhancing flight crew monitoring skills can increase flight safety. *Flight Safety Digest*. March, 1999, pp. 1-8.
- Sumwalt, R.L (2009, May). *Training for Enhancing Flight Crew Monitoring and Cross-checking Skills*. Retrieved from http://www3.nts.gov/speeches/sumwalt/RAA_Training_Committee_052009.pdf
- Sumwalt, R. L., & Lemos, K. A. (2010) The accident investigator's perspective. In B. Kanki, R. Helmreich and J. Anca (Eds.), *Crew resource management* (2nd ed.) (pp. 469-492). Amsterdam: Elsevier.
- Sumwalt, R. L., Morrison, R., Watson, A., & Taube, E. (1997). What ASRS data tell about inadequate flight crew monitoring. In *Proceedings of the Ninth International Symposium*

- on Aviation Psychology*, R. S. Jensen, & L. Rakovan, (Eds). Columbus, OH, United States: Ohio State University.
- Sumwalt, R. L. III, Thomas, R. J., & Dismukes, R.K. (2002). Enhancing flight crew monitoring skills can increase flight safety. In *Proceedings of the 55th International Air Safety Seminar*; Flight Safety Foundation (pp. 175-206), Dublin, Ireland, November 4-7.
- Thomas, M. J. W. (2004). Predictors of threat and error management: Identification of core non-technical skills and implications for training systems design. *International Journal of Aviation Psychology*, 14(2), 207-231.
- Thomas, M. J. W. (2005). *Error Management Training: An Investigation of Expert Pilots' Error Management Strategies During Normal Line Operations and Training: Final Report*. (No. 2004/0050 – ATSB Aviation Safety Research Grant Scheme Project). Adelaide, SA: University of South Australia.
- Thomas, M. J. W. & Petrilli, R. M. (2006). Crew familiarity: Operational experience, non-technical performance, and error management. *Aviation, Space, & Environmental Medicine*, 77(1), 41-45.
- Tullo, F. J. (2001, May). Responses to mistakes reveal more than perfect rides. *Aviation Week & Space Technology*, 154(21), 106.
- Tullo, F. J. (2010). Teamwork and organizational factors. In B. Kanki, R. Helmreich, and J. Anca (Eds.), *Crew resource management* (2nd ed.). (pp. 59-78). Amsterdam: Elsevier.
- Wickens, C. D., Gordon, S. E., & Liu, Y. (2004). *An introduction to human factors engineering*. (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Appendix A: Informed Consent

Informed Consent Form

Purpose of Research

The purpose of this project is to determine effects of training on aircrew monitoring and cross-checking performance.

Specific Procedures

There are two possible training conditions, each of which will expose you to a training activity. You will be assigned to only one of the conditions through a random process. In either condition, you will begin by completing a short training activity. Responses/feedback in this section are not recorded or graded. You will then be asked to answer 10 questions about monitoring and cross-checking. In the final section of the training, you will answer a short set of questions regarding your feelings toward the training and give your opinions on certain flying techniques and cross-checking responsibilities. The total duration of the session will be 20-30 minutes, depending on the condition.

Regardless of the condition you are assigned, you will next participate in the P-course simulator training session. This research will not modify the P-course simulator training curriculum in any way. If you consent, ATC will allow us access to your simulator data to analyze for effects of the two training conditions.

Benefits to the Individual

Following the data collection period, a researcher will contact you and provide you with a debriefing. It is possible for you to gain a deeper understanding of aircrew monitoring and cross-checking behaviors.

Risks to Participants

While the materials in the training conditions may change your behavior during the simulator training session, it is unlikely that the training conditions will impact your in-flight simulator performance.

Confidentiality

Your privacy will be protected. Your name will not be associated with neither the answers you provide in the training nor with simulator data. A code number will be used to identify the data, and this code number will not be associated with your name in any documentation upon completion of the study.

Voluntary Participation

You do not have to participate in this research project. If you do agree to participate, you can withdraw participation at any time without penalty.

Human Subject Statement

If you have any questions about this research project, contact Brian Potter at (386) 402-8762. If there are concerns about the treatment of the participants, contact Dr. Albert Boquet at the Embry-Riddle Aeronautical University Institutional Review Board at (386) 226-7035. The email address is boque007@my.erau.edu.

I HAVE HAD THE OPPORTUNITY TO READ THIS CONSENT FORM AND AM PREPARED TO PARTICIPATE IN THIS PROJECT.

Participant's Signature

Date

Participant's Name

Researcher's Signature

Date

Appendix B: Demographic Questionnaire

Demographics questionnaire*	
1	What is your age?
2	What is your pilot designation?
3	How many total flight hours do you have (all aircraft)?
4	How many flight hours do you have in the HH-65?
5	How many CG flight hours do you have in the past 60 days?
6	What is your gender?

*Demographics collected using www.surveygizmo.com

Appendix C: Monitoring and Cross-checking Online Training

Crew Monitoring and Cross-Checking Skill Training

Instructions: This is a USCG training program designed to get you to advance your overall monitoring/cross-checking skills. You will get some interactive tasks and questions along the way – these are designed to reinforce concepts and are not graded. At the end of the training, you will receive a ten-question quiz at which point your answers will be recorded, but kept confidential. Enjoy!

[Click here to Start Training!](#)

Monitoring Training

The following 5 topics will be covered in order...

- Importance of Monitoring
- What to Monitor
- How to Monitor “5-2-1” Rule
- Monitoring Practice
- End of Course Quiz

Northwest Airlink Flight 5719

- One cold night in December 1993, a Jetstream 31 (twin turboprop) began its descent on a localizer approach into Hibbing, MN. The captain had intentionally delayed some discretionary step-down altitudes given by the controller to avoid entering a cloud layer (potential icing conditions) as long as possible and how the crew would have to fly a steeper-than-normal approach.
- Another factor adding to the complexity of the situation was that the captain had insisted the first officer delay the landing checks until they reached the final approach fix (FAF).



Northwest Airlink 5719

- A few minutes later, the first officer made an altitude callout as the aircraft arrived at the FAF. The plane was over 1,000 feet too high and in a steep 2,250 from descent. At this point, the captain added to the first officer’s workload by giving the him two additional tasks – to call ahead for fuel and to activate the runway lights using the CTAF frequency.
- As the descent continued, the first officer seemed preoccupied with these tasks and failed to give the captain any additional altitude callouts or reminders of an upcoming intermediate step-down altitude prior to the MDA.
- The captain never slowed down the 2,000+ from descent rate. The diving aircraft crashed 2 miles short of the runway killing all 18 people on board.

Northwest Airlink 5719

- Obviously, at some point inside the final approach fix, both pilots lost altitude awareness during this night approach.
- The first officer was actively clicking his radio as the aircraft passed through the step-down altitude, oblivious to the incomplete landing checks and the fact that the captain had essentially given him a distraction which would prove to be fatal. The runway lights activated just as the airplane’s wings began to crash through some trees.



Airlink 5719 – Could it happen again?

- A few years after this incident, the same scenario was presented to pilots during a line audit scenario. Scarily, the result was nearly the same, although the aircraft did successfully land (barely, though) – touching down halfway down the runway.
- This time, the crew even had the added safety benefit of a ground proximity warning system (GPWS) that the mishap aircraft did not which arguably helped prevent a worse outcome.

The primary objective of this training is to develop a skill for safety pilots (SPs) to use to catch flying errors by the flying pilot. The learner will practice a monitoring/cross-checking technique which promotes a "monitor-first" mentality.

"Pilot training programs should be modified to contain modules which train and emphasize monitoring skills."

-NTSB recommendation to the FAA in a post-incident analysis in 1994. Since the NTSB has reissued the same recommendation following each of three multiple-fatality accidents in the past 5 years.



Importance of monitoring/cross-checking

- Monitoring serves three important functions:
 - Keeps crew apprised of status of aircraft, automation & flight path
 - Helps crew catch errors
 - Helps detect evolving aspects of flight situation
- Empirical data from line observations found pilots make an average of 3.7 monitoring errors on each flight.

Line audit data shows pilots of all experience levels make the same number of monitoring errors – indicating that pilots do not naturally become better monitors as their experience increases!

NTSB data

In commercial aviation, inadequate crew monitoring was found to be a factor in 84% of all accidents attributed to pilot error. (NTSB, 2004)



Examples of monitoring errors

An example of a monitoring error would be during a IFR descent to 1,000 feet, you hear the RADAT bug chime which cues you to look up and see you're at 900 feet and you've busted your target altitude.

- Other examples:*
- During an instrument approach, a SP fails to make one of the "standard calls" at decision height (DH).
 - During level-off and landing checks, monitoring is built into steps like "center instrument panel – check" and "landing gear – down." Omitting either of these steps would be both a checklist error and a monitoring error.

Problem

- Among the duties of the safety pilot (SP) – a list that includes radios, checklists, briefings, and navigation duties – is monitoring. While we assume the PAC is monitoring the flight path, *monitoring needs to be a shared task by both pilots.*
- This is sentiment is shared by the vice chairman of the NTSB, CRM training experts, and the FAA (see AC 120-71A and AC 120-51E).

Which of the following is a function of SP monitoring?

- A. Keep crew apprised of the status of aircraft, automation, and flight path.
- B. Help crew catch errors.
- C. Helps detect evolving aspects of flight situation.
- D. All of the above are correct.

Which of the following is considered a monitoring error?

- A. During an approach, the SP forgets to make one of the standard calls at the missed approach point.
- B. During level off, you fail to check and report the center instrument panel as "checked."
- C. Before landing, the SP fails to report the landing gear "down and locked."
- D. All of the above are correct.

Summary of "Importance of monitoring" section


- *Monitoring* is a crew duty that can prevent deadly mistakes from occurring. Potentially serving as the last line of defense, the SP plays a critical role in monitoring the aircraft's flight path and system status.
- *The next sections will cover some best practices for SP monitoring while on IFR flight plans.*

Monitoring Training

- Importance of Monitoring
- What to Monitor
- How to Monitor: "5-2-1" Rule
- Monitoring Practice
- End of Course Quiz

Areas of Vulnerability

Objective:
Students shall be able to describe the *areas of vulnerability* chart and apply it to their own workload management strategy.

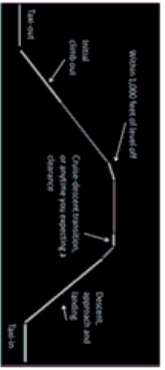


Areas of Vulnerability

- For IFR flights, the flight times when pilots commit the most errors are on takeoffs, landings, and during vertical phases of flight.
- These flight phases tend to carry a higher workload, so it's easier for the overtasked flying pilots to lose track of an important piece of information.
- Some safety experts developed the concept of "areas of vulnerability" to describe these phases.

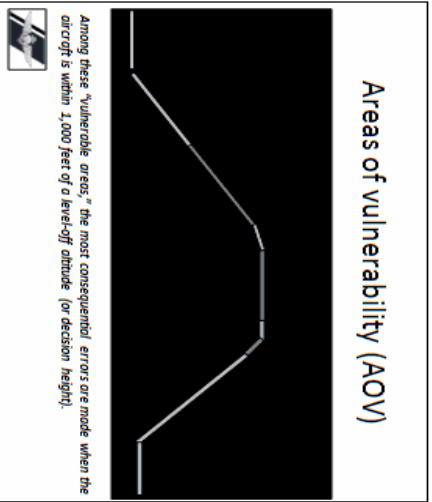


Areas of vulnerability (AOV)



The **green** segments show times when a lower workload prevails. This allows adequate time to conduct tasks such as brief approaches, listen to ATIS, do checklists, stow charts, etc.

The **orange** segments show times when a higher workload prevails and both pilots should be monitoring the aircraft and flight systems. By having the SP actively monitoring, the crew is most prepared to detect and prevent deviations.



Research into AOV & monitoring

- Climbs/descents: In a study looking at all types of aircraft operations, 75% of monitoring errors occurred while the aircraft was in a vertical phase of flight.
- IER flightplans: NASA identified most the 3 most common consequences of monitoring errors are altitude deviations (54%), course correction (1.7%), airspeed departure (6%)
- The same study found 30% of all safety threats arise during the approach-to-landing phase of flight

Based on this data, the focus of the training is specifically on improving the SP's role of monitoring aircraft altitude during vertical flight.

1,000 feet to Level Off

- The key to having success in monitoring during approaches is first having both pilots get into a 'monitoring' mindset.
- One recommendation based on the AOV chart is to take advantage of the instrument approach brief where there's a spot for you to brief "crew duties":

When you brief this section, you should insist the SP back you up during the last 1,000 feet of the descent.

The "areas of vulnerability" chart would be useful for...

A. Selecting pilots to fly on a high risk flight
 B. Choosing times not to conduct briefings, run checklists, etc.
 C. Making a pre-flight risk assessment
 D. Deciding which flight phases carry the highest risk of making an error.

Which of the following flight phases are considered "areas of vulnerability"? (check all that apply)

taxi-out
 within 1,000 feet of level off
 within 30 degrees of rollout on heading
 during an airspeed adjustment
 descent
 approach and landing
 taxi-in

Which of the following flight phases are considered "areas of vulnerability"?

taxi-out
 within 1,000 feet of level off
 within 30 degrees of rollout on heading
 during an airspeed adjustment
 descent
 approach and landing
 taxi-in

Correct for passengers →

The AOV chart (and most airline SOPs) suggest that both pilots should focus on ensuring the aircraft levels at the assigned altitude during the last feet of altitude change.

(click box with the answer)

The chart shows a line representing altitude change from 2000 feet down to 300 feet. The line is divided into segments with labels: 2000, 1500, 1000, 500, and 300. The line is horizontal at each of these levels, indicating a level-off period.

Using the line below, identify the six areas of vulnerability...

(click directly on the line)

The graph shows a line with a question mark in the center. Below the line is a button that says "Click here to continue only once you've selected all six."

Areas of vulnerability

The diagram shows a line representing flight phases: "Initial climb out", "Cruise/Descent transition or arrival you're expecting a clearance", "Descent approach and landing", and "Taxi-in". A label "Within 1,000 feet of level off" points to the top of the cruise/transition phase. A button at the bottom says "Section complete - next section" with a right-pointing arrow.

Monitoring Training

- Importance of Monitoring
- What to Monitor
- How to Monitor "5-2-1" Rule
- Monitoring Practice
- End of Course Quiz

5-2-1 Rule

- Objective:** This section of training will provide you with a behavioral strategy to use as the safety pilot (SP) to avoid making errors during vertical phases of flight.

5-2-1 rule: Context

- We tend to think of bad vs. good flying habits, but for a change let's consider bad vs. good non-flying habits.
- We've seen the data shows a majority of accidents involve a failure by the SP to monitor the actions of the flying pilot.
- We've also discussed that the majority of these accidents occur during instrument approaches, with plenty of additional errors occurring during other vertical phases of flight.

5-2-1 rule

- The 5-2-1 rule states the SP should monitor the flight path and verbally callout altitudes at 500, 200, and 100 feet prior to any level off altitude.
- For example, if the DH for an approach is 200 feet, the SP should make at least 3 altitude callouts: one at 700 feet, one at 400, and at 300 feet.
- As another example, if you were assigned a climb to 1,500 feet by ATC, at 1,400 feet, the SP would callout either "100 feet to go" or "passing through 300 feet for 200," whichever is more appropriate.

5-2-1 rule

- It's critically important that both pilots concentrate on leveling off the aircraft during the final 1,000 feet of any climb or descent, and make at least 3 callouts during the last 500 feet. The rationale behind this concept is that other "distractions" and routine tasks are usually of a lower priority and can wait. If something else can't wait (like a landing clearance) and you're on an approach, you've drastically underestimated your workload and probably a go-around is in order.

Distractions are a huge problem that factor prominently in 72% of all approach and landing accidents. These distractions break up the flow pattern of cockpit activities such as the pilots' monitoring the aircraft systems and each other's actions.

5-2-1 Rule

This callout task engages both pilots in the task at hand and benefits them by putting them in an excellent position to complete the task. The SP is also in a great position to respond to any deviations or take the controls.



5-2-1 rule

This rule reinforces the SP's core responsibility to be vigilant in monitoring the flight path as the flying pilot completes a climb or descent.



Section summary

- The 5-2-1 rule is a process that virtually assures that the crew will get the results they want at the completion of the climb or descent.
- Put in plain English, the rule says, *stop what you're doing and focus your attention on the most likely thing to get you in trouble right now.*
- With practice, you'll become increasingly systematic as you start watching the altitude very closely during the last 500 feet of a vertical maneuver.



Question

A rule-of-thumb for monitoring/cross-checking states ...

- The PIC would make altitude callouts at 500 and 200 feet prior to any level off in altitude.
- The SP would make altitude callouts at 200 and 100 feet prior to any level off in altitude.
- Either crew member would make callouts 5, 2, and 1 seconds prior to completion of any time maneuver (e.g., approach, holding etc.)
- Either crew member would make altitude callouts at 500 and 100 feet prior to any altitude level off.

Question

According to accident data, the prominent cause of the crew committing a monitoring error during an approach is:

- Lack of crew coordination
- Multitasking
- Disorientation
- Distraction



• As safety pilot on an IFR flight, how many times during the last 500 feet of vertical flight should you make altitude callouts to the flying pilot prior to reaching the target altitude?

- 0
- 1
- 2
- 3
- 4
- 5



Correct. According to the 5-2-1 rule, calls would be made at 500, 200, and 100 feet prior.

Go to next module



Monitoring Training

- Importance of Monitoring
- What to Monitor
- How to Monitor: "5-2-1" Rule
- Monitoring Practice
- End of Course Quiz



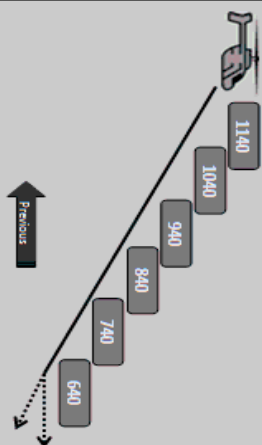
Practice

• You will get five scenarios to apply the 5-2-1 rule. Read the questions carefully.

Continue



• Your MDA is 540 feet. According to the 5-2-1 rule, the first required altitude callout the SP would make during the descent would be at:
(Click the box with the correct answer)



• Your DH is 350 feet. According to the 5-2-1 rule, the *second* required altitude callout during the descent the SP would make would be at:

(click the box with the correct answer)

850
650
550
450

Previous Q

• ATC instructs a climb to 3,000 feet MSL. According to the 5-2-1 rule, the *first* required altitude callout the SP should make would be when passing through:

2,600
2,800
2,700
2,900
3,000

Previous Q

• Your MDA is 420 feet. According to the same rule, the *third* required altitude callout (last before making a standard call at the MDA/DH) would be at:

(click box with the answer)

620
540
520
420

Previous Q

• ATC instructs a climb to 1,500 feet MSL. Again according to the same rule, the *second* required altitude callout the SP should make would be when passing through:

1,100
1,000
1,300
1,200
1,500
1,400

Previous Q

Monitoring Training

- Importance of Monitoring
- When and What to Monitor
- How to Monitor: "5-2-1" Rule
- Monitoring Practice
- End of Course Quiz

End of course quiz and Post-training survey

This completes the training. Please close this window and answer the remaining survey questions.

Appendix E: "Reactions" Questionnaire

-
- 1 **Please indicate how much you either agree or disagree with the following three statements: ***
(use scale from 1 [strongly disagree] to 5 [strongly agree])
- a. The course objectives were clear.
 - b. The course material was organized and presented in an effective manner.
 - c. The course effectively used practical applications to make the subject matter more relatable.

-
- 2 **To what extent do you expect this training will make a difference in the way you do your job?**
() 1 - No difference..... () 5 - Tremendous difference

-
- 3 **How would you rate your overall knowledge of monitoring before the training activity?**
() 1 - Low ... () 5 - High

-
- 4 **How would you rate your overall knowledge of monitoring after the training activity?**
() 1 - Low () 5 - High

-
- 5 **How would you rate this training course overall?**
() 1 - Poor..... () 5 - Excellent
-

Note: Questionnaire was distributed using www.surveygizmo.com

* Question 1 was not presented to the control group.

Appendix F: Attitudes Questionnaire

-
- 1 **When you are in the role of the safety pilot, how would you rate the importance of the following duties?** (score on scale from 1 = *of no importance* to 7 = *extremely important*)

Completing checklists
 Radio communications
 Operating the flight director (autopilot)
 Monitoring the actions of the other pilot
 Monitoring the status of aircraft systems
 Monitoring the aircraft's flight path
 Conducting briefings

-
- 2 **When the flight director (i.e., automation) is helping to "fly" the aircraft and I am not at the controls, I am _____ to monitor the actions of the flying pilot.** (Choose a score from 1-6)

1 = significantly less vigilant ... 7 = significantly more vigilant

-
- 3 **Assume you are out flying in actual instrument conditions. If you are the safety pilot, how would you rate the importance of monitoring each of the following instruments/indications?** (score on scale from 1 = *of no importance* to 7 = *extremely important*)

Barometric Altimeter Radio Altimeter Airspeed Indicator
 Heading FMS Modes ADI (Attitude Direction Indicator)
 VSI (Vertical Speed Indicator)

-
- 4 **When you are the safety pilot, how would you rate the importance of "monitoring" (i.e., observing flying pilot's actions, instrument scanning, & system status) during the following flight regimes:** (score on scale from 1 = *of no importance* to 7 = *extremely important*)

Pre-takeoff/taxi Takeoff Enroute to destination/op area
 Approach to landing Approach to overwater hover Landing
 Hover operations

-
- 5 **Both pilots should be monitoring the aircraft's altitude and not be involved in other tasks (radio calls, checklists, etc.), during the last x feet of any climb or descent to ensure the aircraft levels off at the appropriate altitude. In regards to the variable "X", I feel that...**

X (number) of feet is an appropriate guideline (X should be ____)
 More guidelines are needed. Exact number for "X" should vary proportionally to aircraft altitude (e.g., 100 feet prior during low level maneuvers, 500 feet prior during enroute flying, etc.)
 More guidelines are needed. Exact number for "X" should vary based on flight regime (IFR, VFR, overwater, pattern ops, etc.)
 No guideline needed. The existing rules for altitude callouts that are in place are fine, and leaving decisions how to allocate other safety pilot responsibilities to the discretion of the crew is ok.

Note: Questionnaire was distributed using www.surveygizmo.com.

Appendix G: Participant Debriefing

Participant Debriefing script

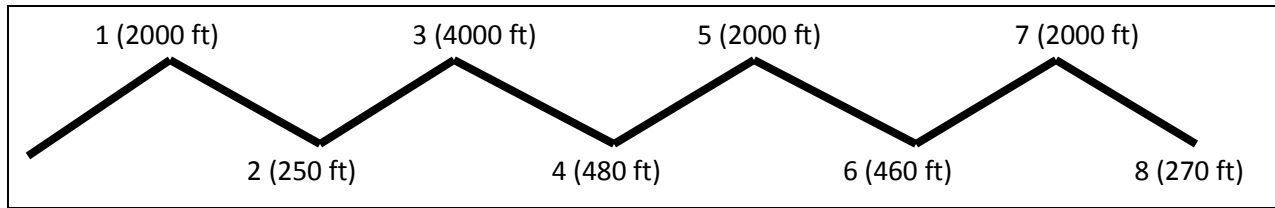
Thank you for participating in this research project. Your participation will assist in understanding the effectiveness of using online and computer-based methods to train non-technical skills such as monitoring and cross-checking to pilots. Instances of non-flying pilots failing to keep track of the flying pilot's actions and aircraft's trajectory are increasingly common in aviation accidents; however, there is little effort at this time to specifically train pilots the skills or strategies to catch these types of errors. In some ways, these types of skills are related to those trained during CRM training, but need to be explored before this relationship and training needs are fully understood.

For this study, the researchers were primarily interested in the effectiveness of an interactive, online monitoring training activity. Half of the participants received this online training and half did not. In order to measure its effectiveness, we looked at several different potential outcomes. Specifically, the researchers had you answer a group of test questions, and give us your reactions and opinions of the training. Then, using recordings of your simulator flights while at ATC Mobile, we also observed and measured aspects of pilot-to-pilot communications, with particular attention to the vertical phases of flight.

It will take up to three months to complete this study and publish the results. It is requested that you not discuss your experiences in this session until after the results are published. Please contact the researcher if you would like to receive a summary of the results; we just ask you be patient.

If you have any questions, please let the researcher (Brian) know at brian.a.potter@uscg.mil or you can contact the Human Factors and Systems department of Embry-Riddle Aeronautical University. Results of this study will be published and placed in the Hunt Library (Daytona Beach Campus) upon completion of this project.

Appendix H: Sample Rater Grade Sheet – IFR scenario #1



IFR#1	Time: _____	Pilot IDs: _____	Trained/Untrained
1.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
2.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
3.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
4.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
5.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
6.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
7.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
8.	Start time _____ End time _____ Ttl Secs _____	500 200 100	Overshoot ____ or N/A Success? Y/N
	SP Callouts ____	Total: ____ (0-3)	
Enter total Callouts here _____		Total 5-2-1 = _____	Ttl Overshoot = _____
Divide by Grand total secs _____		Divide by 24 = _____	Divide by 8 = _____
= _____ *60 = callouts per minute: _____			