AFRL-RH-AZ-TR-2011-0005

Real Time Cockpit Resource Management (CRM) Training

David Kaiser
Jeffery Eberhart
Chris Butler
Gregg Montijo
Michael Vanderford
Crew Training International, Inc.

Alan Spiker
Wayne Walls
Anacapa Scicnces

October, 2010
Final Report

THIS IS A SMALL BUSINESS INNOVATION RESEARCH (SBIR) PHASE II REPORT.

Distribution A: Approved for public release; distribution unlimited. (Approval given by 88 ABW/PA, 88ABW-2011-2393, 2 May 11.)

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

SBIR Rights Legend

Contract Number: FA8650-08-C-6848
Contractor Name: Crew Training International, Inc.
Contractor Address: 9198 Crestwyn Hills Dr Memphis, TN 38125
Location of SBIR Rights Data: None

The Government's rights to use, modify, reproduce, release, perform, display, or disclose technical data contained in this report are restricted by paragraph (b)(2) of the Rights in Noncommercial Technical Data and Computer Software –Small Business Innovation Research (SBIR) Program clause (DFARS 252.227- 7018 (June 1995)) contained in the above identified contract. Any reproduction of technical data or portions thereof marked with this legend must also reproduce the markings. Any person, other than the Government, who has been provided access to such data must promptly notify the above named contractor.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC).

AFRL-RH-AZ-TR-2011-0005 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

ROBERT NULLMEYER

HERBERT H. BELL

JOEL BOSWELL, Lt Col, USAF

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YY)* 1 October 2010	**2. REPORT TYPE** Final	**3. DATES COVERED *(From - To)*** 18 Apr 2008 – 30 June 2010

4. TITLE AND SUBTITLE Real Time Cockpit Resource Management (CRM) Training	**5a. CONTRACT NUMBER** FA8650-08-C-6848 **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Kaiser, David Spiker, Alan; Walls, Wayne Eberhart, Jeffery Butler, Chris; Montijo, Gregg; Vanderford, Michael	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER** 3005HAH9
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Crew Training International, Inc. 9198 Crestwyn Hills Dr Memphis, TN 38125-8538	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate RHA Division Mesa, AZ 85212	**10. SPONSORING/MONITORING AGENCY ACRONYM(S)** AFRL/RHA **11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)** AFRL-RH-AZ-TR-2011-0005

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution unlimited. (Approval given by 88 ABW/PA, 88ABW-2011-2393, 2 May 11.)

13. SUPPLEMENTARY NOTES

This is a Small Business Innovation Research (SBIR) Phase II report developed under a SBIR contract for topic AF071. Report contains color.

14. ABSTRACT Working with the MQ-1 community, four specific training interventions were developed for pilots and sensor operators in initial qualification training: (1) enhanced academics; (2) web-based interactive mishap case histories; (3) a game-based multi-task skills trainer; and (4) GemaSim, a laptop-based team trainer. The four training interventions were introduced cumulatively over the course of 18 months for 27 different MQ-1 training classes (540 aircrew) using baseline, control and experimental classes. Training intervention effectiveness was measured through all four of Kirkpatrick's levels of learning.

15. SUBJECT TERMS

SBIR;Training interventions, Enhanced Academics, web-based interactive mishap case histories, game-based multi-task skills trainer, GemaSim laptop-based team trainer, Kirkpatrick's Levels of Learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Robert Nullmeyer
Unclassified	Unclassified	Unclassified	SAR	163	**19b. TELEPHONE NUMBER** *(Include Area Code)* (480) 988-6561, ext. 283

Table of Contents

ABSTRACT

As the USAF dramatically expands Remotely Piloted Aircraft (RPA) usage, the explosive growth in flying hours is accompanied by an equally rapid growth in the need to train aircrews and by initiatives to use less experienced personnel as operators. Earlier research revealed most MQ-1 accidents involved causal human factors, often in four specific aircrew behavioral areas: inadequate crew coordination, channelized attention, task misprioritization, and wrong course of action.

Working with the MQ-1 community, 4 specific training interventions were developed for pilots and sensor operators in initial qualification training: 1) enhanced academics; 2) web-based interactive mishap case histories; 3) a game-based multi-task skills trainer; and 4) GemaSim, a laptop-based team trainer. The 4 training interventions were introduced cumulatively over the course of 18 months for 27 different MQ-1 training classes (540 aircrew) using baseline, control and experimental classes. Training effectiveness assessment was structured around Kirkpatrick's Level I (student reaction), Level II (evidence of learning), and Level III (transfer of learning). Level III impacts were evaluated during 2 specific sorties in training.

Student reactions (Level I) regarding all four treatments were positive, but highest for the GemaSim team trainer and enhanced academics. Student reactions for interactive case studies and multi-task skills training were higher for sensor operators than experienced pilots. Evidence of learning (Level II) was present for all interventions and was statistically significant with the GemaSim intervention and enhanced academics. Positive transfer of learning (Level III) was observed for enhanced academics and the full complement of interventions. Organizational impact (Level IV) was attempted, but was inconclusive due to syllabus changes, student experience levels, increased operational tempo, and a short evaluation period. As the Air Force moves to more inexperienced RPA aircrews, these interventions may be increasingly useful to not only reduce mishap rates, but also to increase mission effectiveness.

ABOUT THE RESEARCHERS

Alan Spiker is a principal scientist at Anacapa Sciences, which was subcontracted by Crew Training International to assist with the design and analysis of this project. He received his PhD in experimental psychology from the University of New Mexico in 1978. At Anacapa, he is responsible for managing human factors research projects in advanced technology, and has specialized in aviation training and human performance for all branches of service and the airline industry. During the past 11 years, he has conducted research on the cognitive underpinnings of effective, safe aircraft operation, including crew resource management, multi-tasking, planning, and decision-making.

Jeff Eberhart is an Instructor/Subject Matter Expert for fighter, bomber and remotely piloted aircraft for Crew Training International and the on-site manager for the Real Time Cockpit Resource Management Research project at Creech Air Force Base. He is a combat-experienced veteran and military aviator with over 4,300 flying hours, holding various command positions culminating as a numbered USAF vice commander. He earned a Master's degree from Auburn University and has been with Crew Training International since 2007.

Michael Vanderford is the Combat Air Forces Courseware Manager. He is a former F-4, F-16 Instructor Pilot and Weapons School Graduate. He has been the primary Fighter CRM courseware developer for CTI since 1999. He is also a lead fighter I/SME. He earned a Bachelor of Science in Electrical Engineering from Memphis State University and attained his Master's in Business Administration from Golden Gate University. In his USAF career he served as 80 FS Commander, Chief of Range Management, Chief of Wing Training Chief, International Affairs, Weapons Secretary of the Air Force, Chief, Plans And Programs, ACC as well as multiple other roles.

Christopher Butler is an A-10/Fighter I/SME. He is a USAF Academy graduate. He had a distinguished career as an F-16 and A-10 pilot. His organizational skills and attention to detail garnished him praise as one of the best mission planners in Desert Storm. As a result, he was assigned to the Stan/Eval office at HQ ACC. He brought these skills and enthusiasm to the CAF CRM program in which the students consistently identify him as one of the best CRM instructors.

Gregg A. Montijo is the Combat Air Forces Crew/Cockpit Resource Management Program Manager for Crew Training International, Inc. He is a retired USAF command pilot with over 4,000 hours in the A-10, F-16, A-320 aircraft and various gliders. He earned a Bachelor's degree from the USAF Academy in 1981, specializing in Human Factors Engineering. He also earned a Master's degree in Procurement and Acquisition Management from Webster University in 1995. He holds an Airline Transport Pilot's license with a type rating in the B-737 and has over 11 years program management experience at the Pentagon and in the civilian sector. He has been with Crew Training International since 2001.

Wayne Walls serves as the lead developer and principal scientist at Anacapa Sciences, Inc. He has published over 50 reports and a dozen software or web-based applications that serve in a variety of commercial and government venues. He received his PhD in cognitive psychology from Wayne State University in 1986. Since that time he was worked at General Motors, Hughes Aircraft Company, and OACIS Research before going to Anacapa, where he works on issues related to human-computer interaction, human performance evaluation, and design through rapid prototyping. His UI design and software development experience ranges from automotive GPS navigation interfaces to Naval aviation post-flight reporting systems to web-based distance learning applications. Recently, this includes Google Maps-based applications for locating, tracking, and obtaining information about organizational personnel and assets in international locations.

David Kaiser is Vice President and Chief Learning Officer for Crew Training International and is responsible for all courseware development for both Crew Resource Management and Contract Aircrew Training. He earned his Master's degree from Embry Riddle Aeronautical University with a dual specialization in Aviation Education and Aviation Safety. He is a Navy aviator with combat experience and is currently active in the Navy Reserve. He has managed various experimental projects and has been personally awarded two U.S. patents. He has been with Crew Training International since 2006.

INTRODUCTION

When Orville and Wilbur Wright made their series of initial flights in December 1903, they introduced the modern aviation world to human error as a mishap cause. Early aviators did not have an appreciation of the complexity of flight or the impact of many factors to include human performance (Anderson, 2004). A tremendous amount of progress has occurred since then; aircraft are considerably larger, faster, are more reliable and easier to operate and now, fly without humans on board. The development and operational use of unmanned or Remotely Piloted Aircraft (RPA) has expanded exponentially within the last two decades. The 2006 Quadrennial Defense Review set in motion the acquisition of large numbers of these systems (Office of the Secretary of Defense, 2006). Future aircraft design is following current trends towards unmanned aircraft for a variety of missions, both military and civilian.

The U.S. Air Force (USAF) has dramatically increased RPA use over the last 10 years. The use of the General Atomics MQ-1 Predator began in mid-1995 and currently sees widespread operations throughout Iraq and Afghanistan. Surveillance demands in Iraq and Afghanistan have dramatically increased the requirement for RPAs and have grown from several orbits to over sixty 24-hour-a-day orbits.

The explosion in the aircrew requirements needed to operate the expanding fleet of Predator aircraft has been met by a rapidly expanding training system and the quadrupling of aircrews trained over the past five years (Crew Training International, 2010). According to Air Force Safety Center data (2009), the corresponding annual flight hours increased dramatically between fiscal year 2001 and 2009, from 5,751 to 187,393.

The Predator is operated by a crew consisting of a pilot and sensor operator with assistance from an intelligence specialist and often operated via satellite communication links by crews located halfway around the world. Pilots come from a variety of backgrounds within the USAF. Recently, pilot training graduates with no prior operational flying experience and officers with no prior flying experience whatsoever have entered the Predator pilot force. Sensor operators come from various operational Air Force backgrounds, including many who just recently entered the service with no prior military experience. The intelligence specialists are not officially considered part of the flight crew and their participation is not addressed as part of this study.

Despite advances in technology, equipment, training, and the maturation of the Predator weapons system, the absence of a human-on-board has not translated into the elimination of human error. RPA mishaps have been well documented by a variety of sources and studies, where 4 recent military mishap studies are particularly relevant. Tvaryanas, Thompson, and Constable (2005) reviewed UAV mishaps across the U.S. military services. They reported that since the inception of the systems through the end of FY 2003, 32 mishaps per 100,000 flying hours occurred with the USAF Predator system. Typically, USAF Class A mishap rates ($1 million damage or fatality) were in the low, single-digit (1 to 2) range per 100,000 flying hours during that period (O'Toole, Hughes & Musselman, 2006). Nullmeyer, Herz, Montijo, and Leonik (2007) specifically analyzed Predator mishaps through 1996 and found major changes over time. Class A mishap rates dropped to under 10 per 100,000 flying hours by 2006. Mishap causal factors also changed, shifting from equipment failures to human factors, particularly crew skill and knowledge. While early studies theorized that RPAs experience a higher than normal mishap rate, updated USAF Class A mishap information showed mishap rates for the Predator were improving and following a trend line similar to that seen following the introduction of the F-16, a single-engine manned aircraft two decades earlier.

Missing in almost all of the studies to date are concrete recommendations regarding how to best fix deficiencies documented in mishap trends. While many recommendations have been made regarding hardware fixes, few have addressed the human performance part of the effectiveness equation. The effectiveness of human factors skills training has been often debated in the past with few definitive, data-based transfer of training studies published to date (Helmreich, Merritt & Wilhem, 1999; Salas, Wilson, Burke & Wightman, 2006).

The approach was to begin with as complete a picture of Predator aircrew performance as possible. To determine which human factors skills needed improvement, multiple sources of information were used, starting with detailed analyses of USAF Predator mishap reports. While many studies relied solely on mishap reports, these data only tell part of the story – there were thousands of Predator sorties where human factors skills were equally contributory to a successful mission and these data were also used, obtained from a panel of Predator experts. Additionally, the intent was to capture typical student human factor errors that occur daily in training.

To develop a more complete picture of Predator crew performance and which human factors skills needed improvement, three sources of data were analyzed: Class A mishaps, a Delphi panel of warfighter experts, and Predator training records. A detailed explanation of the mishap analysis process was previously described by Nullmeyer, Stella, Montijo & Harden (2005) and Nullmeyer et al. (2007). The top ten Predator human factors mishaps causes are:

1. Channelized Attention
2. No Training for Task Attempted
3. Crew Coordination
4. Selected Wrong Course of Action (COA)
5. Task Misprioritization
6. Checklist Error
7. Inattention
8. Inadvertent Operation
9. Risk Assessment
10. Confusion

Also used were findings from a Delphi Panel of RPA experts with operational combat and instructional experience. The methodology was previously described in Nullmeyer, Spiker, Montijo and Kaiser (2008). The top 10 operational challenges identified by the Delphi Panel are summarized as:

1. Inadvertent Operation
2. Task Misprioritization
3. Crew Coordination Breakdown
4. Channelized Attention
5. Distraction
6. Select Wrong COA
7. Inadequate Inflight Analysis
8. Misperception of Speed, Distance, Altitude
9. Complacency
10. Cognitive Task Oversaturation

The third source of data was 305 student gradesheets from 70 pilots and 75 sensor operators formally enrolled in Predator training. The methodology and results were also reported in detail in Nullmeyer et al. (2008). The top human factors identified in the gradesheet analysis are:

1. Channelized Attention
2. Inadequate Flight Planning Analysis
3. Crew Coordination
4. Course of Action Selection
5. Lack of Task Training

6. Checklist Error
7. Task Misprioritization
8. Limited Total Experience
9. Risk Assessment
10. Inattention

A subset of these human factors skills was selected to be addressed through training interventions; criteria for inclusion were: 1) they were among the leaders in each source of data; 2) they represented a skill that could be most appropriately addressed through training; and 3) they posed a particular problem for the Predator community. Applying these criteria, the most prominent four human factors skills are listed in Table 1.

Table 1

Identified Predator Human Factors Skills

1	Task Prioritization
2	Channelized Attention
3	Selecting an Appropriate Course of Action
4	Crew Coordination

The goal of this project was to determine if the deficient Predator human factor skills identified in Table 1 could be improved through training, and more specifically, to explore the effects of interactive immersive classroom training, eLearning, Web-based gaming and the hands-on team trainer. With the pertinent human factors skills identified, four training interventions described in detail in Nullmeyer et al. (2008) were refined. The training interventions were:

1. Enhanced Academics (EA)
2. Web-Based Interactive Case Histories (ICH)
3. Game-Based Multi-task Trainer (MTT)
4. Computer-Based Team Trainer (GTT)

From an analytic standpoint, a measurement plan was needed that would provide a comprehensive picture of the effectiveness of these training interventions introduced to the RPA squadron. Kirkpatrick's (1976) four-level framework offers such an approach, in which training impact is assessed in terms of student reaction to the training (Level I), amount of learning exhibited (Level II), degree of transfer to the operational environment (Level III), and change within the organization (Level IV). Salas et al. (2006) reported in their meta-analysis of the crew resource management (CRM) literature that very few studies achieve measurement at either the third or fourth level, with the vast majority focused on student reaction and learning. A primary goal of the present effort was to push the measurement boundaries further, by including Level III and attempting to measure Level IV in the study design.

METHODS

Participants

The participants of this study were students (Pilots and Sensor Operators) going through initial qualification training at the MQ-1 Predator Formal Training Unit (FTU), Creech AFB, NV. The study comprised of classes spanning from class number 08-13 to 10-06 which amounted to 287 Pilots and 253 Sensors Operators for a total of 540 participants.

There were two classes of "non-rated" pilots, a test group of pilots that did not complete undergraduate pilot training (UPT), that were omitted from the study. The typical pilot demographic in this study ranged from a First Lieutenant that recently completed UPT to a Colonel that may have thousands of operational hours in a tactical airframe. The Sensor Operator demographics comprised of approximately 50% being junior enlisted that just completed basic training, with the other half having at least one operational tour in a non-flying position to a senior enlisted with operational flying experience.

Treatment

Various training interventions were researched and four were chosen to evaluate. The critieria used to chose these interventions were based upon their perceived ability to impact the deficient HF skills, their ability to work within the constraints of the FTU environment, and if deemed effective, could be operationally implemented. The following are the details of each training intervention and the "flow" in which they were implemented in the study.

Enhanced Academics

Every Predator aircrew member receives four hours of CRM classroom instruction as part of their formal syllabus training. At best, this academic class can be described as fourth generation CRM training. Enhanced Academics (EA) training exposed Predator crewmembers to sixth generation CRM principles of Threat and Error Management. The two-hour facilitated course immersed the student in interactive Predator mishap case studies focusing on task prioritization, situational awareness, crew coordination and decision making, plus mission planning and communication. Unknown to the researchers was if additional immersive multimedia CRM classroom training potentially affected aircrew performance.

Web-Based Interactive Case Histories

With many of the Predator crews being "Generation Y" and therefore familiar with computer-based training (CBT), an interactive CBT module was designed to further enhance human factors skills. This approach was similar to one developed by Spiker, Hunt, and Walls (2005). Written case histories with interactive hyperlinks included detailed explanations, graphics and where applicable, video or computer re-creations. A visual checklist kept the student on task to ensure learning was standardized and complete. Each case study concluded with a set of fairly difficult questions, written so the student had to understand the lesson's main points. Answers were electronically tracked, collected and analyzed for training effectiveness. Would the effectiveness of Web-based Interactive Case Histories (ICH), using essentially the same information as EA, be more effective than the present methods of training?

Game-Based Multi-Task Trainer

The two previous training interventions focused on human factors knowledge, while this third intervention focused on practicing individual skills in an interactive and competitive environment. This intervention allowed each student to improve their task management skills, actively see and learn to avoid channelized attention, and practice task prioritization and decision making. The four-quadrant multitasking screen with audio, visual, memory, and calculation tasks, was an adaptation of SYNWIN (Elsmore, 1994). It replicated similar tasks needed to operate the Predator weapons system. Students not only received individual scores, but also competed within their student class for top scores. Scores were provided at the end of each session, as well as collected for later analysis. Figure 1 shows a sample Multi-Task Trainer (MTT) screen. Would exposing students to game-based skill exercises improve the identified individual human factors skills performance later in training?

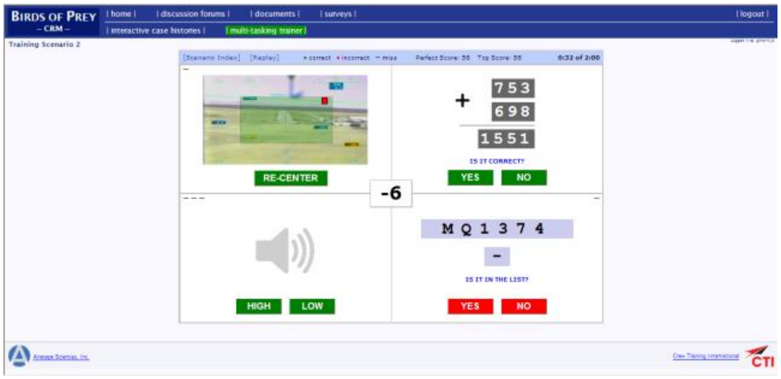


Figure 1. Typical Game-Based Multi-Task Trainer Screen

Computer-Based Team Trainer

This training intervention exercised previously taught and practiced individual human factors skills in a crew environment (crew coordination) under stressful conditions. GemaSim, a proprietary commercial software package, was modified to practice personal and team behavior in a stressful crew flying-type environment. Would practicing individual and crew human factors skills under stress in a non-Predator environment improve their performance on Predator simulator and flying missions? Figure 2 shows Predator crews using the GemaSim Team Trainer (GTT).

Figure 2. GemaSim Team Trainer

Training Intervention Flow

Table 2 shows the class flow, their treatment and the number of students. To establish a typical student performance baseline, student grades were gathered from three classes before introducing the first training intervention. Training interventions were introduced in four spirals. Spiral 1 (classes 09-1, 09-3, 09-5, 09-7, and 09-09) added EA to the initial qualification course syllabus. Spiral 2 (classes 09-11 and 09-13) added EA and ICH. Spiral 3 (classes 09-15 and 10-01) added EA, ICH, and MTT instruction to the syllabus. Finally, Spiral 4 (classes 10-3, 10-5, and 10-6) added all four interventions (EA, ICH, MTT, and GTT). Due to late approval to proceed beyond the first intervention, three extra classes received EA before ICH was introduced. The delay also compressed collecting data for the last class; hence class 10-06 received Spiral 4 instead of class 10-07.

Table 2

Class Treatment Assignment

Class #	Treatment	Size	Class #	Treatment	Size
08-13	Baseline	19	09-10	Control	21
08-14	Baseline	21	09-11	EA + ICH (Spiral 2)	21
08-15	Baseline	19	09-12	Control	19
08-16	EA (Spiral 1)	20	09-13	EA + ICH (Spiral 2)	20
08-17	Control	18	09-14	Control	18
09-01	EA (Spiral 1)	20	09-15	EA+ICH+MTT (Spiral 3)	22
09-02	Control	22	09-16	Control	24
09-03	EA (Spiral 1)	20	10-01	EA+ICH+MTT (Spiral 3)	19
09-04	Control	19	10-02	Control	22
09-05	EA (Spiral 1)	21	10-03	EA+ICH+MTT+TT (Spiral 4)	21
09-06	Control	21	10-04	Control	16
09-07	EA (Spiral 1)	20	10-05	EA+ICH+MTT+TT (Spiral 4)	18
09-08	Control	22	10-06	EA+ICH+MTT+TT (Spiral 4)	18
09-09	EA (Spiral 1)	19		**TOTAL SUBJECTS**	**540**

To prevent student classes from interacting with one another, treatments were given to every other student class with the non-participating class acting as a control group (no treatments, performance evaluated). The interventions were introduced as progressively building spirals rather than as individual treatments since the ultimate focus was on assessing the impact of an integrated curriculum of enhanced knowledge, hands-on practice, and team-building training. Only a spiral methodology would allow this determination to be made with a between-class experimental design.

Instructors/evaluators did not know which classes received treatment(s). Logistically for the USAF, it proved easier to alternate student classes rather than measure select individuals within a class. Dividing classes would have created “compensatory spillover” effects in which members of the control group seek and receive some or all aspects of the treatment once word gets out (Cook & Campbell, 1979). Since Predator students in a given class train together for three months, there would have been no secrets among them once training interventions were introduced. By alternating student classes, all classes were separated by approximately six weeks and rarely, if ever, interacted with one another.

Data Collection and Measurement

Training effectiveness was addressed using Kirkpatrick’s (1976) levels of learning, in terms of student reaction to the training (Level I), amount of learning exhibited (Level II), degree of transfer to the operational environment (Level III), and change within the organization (Level IV).

The nature of the experimental design allows for intervention-specific assessments at Levels I and II, whereas the Level III transfer effects are examined at the spiral level. Level I involved completing student critiques customized for each specific training intervention. Level II looked for evidence of learning within each training intervention. Level III addressed transfer of training to subsequent environments through evaluation at a final simulation check ride (EPE) and the final flight mission CO-3. Baseline student data (Level I from their current CRM class and Level III) were collected to see how an average student performed without any of the training interventions.

During EA, Level I data were collected from a course critique (Appendix A) and Level II data from a seven-question pre- and post-test quiz (Appendix B). ICH collected Level I data from an online survey (Appendix C) and Level II data from quiz results. MTT collected Level I data through an online survey (Appendix D), and the gaming scores for the various exercises provided Level II data. For all students (control and experimental groups), Level III student performance data were collected in the four targeted skill areas during two sorties: the simulator emergency procedures evaluation (EPE) and the final flying mission before course graduation (CO-3). Data collection instruments included the standard Air Combat Command Form 206 gradesheet (Appendix F) and a supplemental HF form (Appendix E). The HF form was designed specifically for this study where instructors used a traditional five-point scale (0-4) in the EPE session and the CO-3 mission. Strengths and weaknesses were identified on this HF form for 25 specific behaviors distributed across the four targeted skill areas. Additionally, all student gradesheets (approximately 21 gradesheets for all 540 students; 11,340 gradesheets) were collected for later data analysis.

Level IV data were collected through the use of an electronic survey given to the gaining units' leadership and the analysis of Hazardous Air Traffic Reports (HATRs), and by measuring the time it took IQT graduates to become mission-ready.

DATA ANALYSIS

SPIRAL 1 (Enhanced Academics [EA])

Level I (Student Critiques of EA)

The classes covered in this analysis are 08-16, 09-01, 09-03, 09-05, and 09-07, which correspond to the five classes that received the EA course (i.e., Spiral 1). Students completed the critiques immediately at the end of the course. By design, the critiques were anonymous, so there was no way to link students' critiques or opinions of the course with either their learning/comprehension data (Level II) or their performance in the two training sessions (CO-3, EPE). Nevertheless, analysis of the critique data yielded immediate, uncensored student opinions of the course they just received. As such, this analysis provided a "first look" at what the students thought about the instruction they just received. While it is possible that students could learn effectively (Level II) from instruction they disliked and transfer that knowledge and skill to the simulator/mission training environment (Level III), the odds of having successful learning and transfer when the training is disliked are not high. Hence, it is desirable to develop and deliver training that is met with favorable responses by the students. The analysis revealed that was certainly the case with the instruction being received. In addition, though, the analysis highlighted some areas where the instruction could be improved.

Critique Sheet.
For the present SBIR project, the critique sheet CTI has been routinely using for its baseline CRM course was modified by adding a second page that asks questions specific to the four human factors (HF) skills being targeted: attention management, task prioritization, course of action (COA) selection, and crew coordination. As before, the first page consisted of 15 Likert-scale type questions that ask students to express their degree of agreement with various positively-worded statements concerning course relevance, content, usefulness of CRM skills, and instructor proficiency. A five-point scale was used in which the student could either circle "strongly agree" (i.e., the highest rating), "agree," "neutral," "disagree," or "strongly disagree." There was a sixth option, "not applicable," which corresponded to a

no-rating. To support quantitative analysis, these responses were converted into a numeric scale by assigning “strongly agree” to 5, “agree” to 4, “neutral” to 3, “disagree” to 2, and “strongly disagree” to 1. The data were entered into an Excel worksheet where statistical calculations (means, variance) were performed. Responses corresponding to “not applicable” were disregarded in any calculation of average ratings.

With one exception, the remaining questions on the front page and all the questions on the back side asked for comments concerning course likes, dislikes, and ways the course did, or should have, improved the student’s knowledge of the four targeted HF skills. The other question asked students to indicate in which of the seven primary CRM areas (mission planning, debriefing, flight integrity/crew coordination, communication, situation awareness, task management, risk management/decision making) they thought they needed more training. The comments for each student were also entered into an Excel worksheet in which qualitative content analysis was performed.

A separate worksheet tab was created for each class, where quantitative and qualitative analyses were performed on each class separately. Interpretive analyses were then performed that compared the five classes on an item-by-item basis to examine trends. The results of these analyses are described in the following subsections.

Before revealing the results, a brief note is needed about the quantity and quality of data collected. One student in class 08-16 failed to respond to any of the Likert-scale items, yet did supply comments to many of the qualitative questions. With a class size of 20, this omission resulted in a total of only 19 respondents for analysis of the rating data. For class 09-05, it was necessary to delete one student from the analysis since this individual provided very minimal, non-informative data concerning his/her dislike of the class. In particular, the student rated all items a ‘1’ and made general negative comments with no specifics. Since the attitude and non-responsiveness was so inconsistent with all of the other respondents in all the classes, it was decided to delete the entire student’s data since it would unduly influence the class average data, rendering comparisons with other classes meaningless. The loss of this student resulted in a total of 19 students for analysis. Other than these two students, complete critique sheet data were received from all the students who took these five academic classes; hence this demonstrates a fairly complete picture of what students thought of the class. In all, there are rating data from 100 students.

With regard to comments many students opt not to respond either due to lack of time, difficulty in free-form writing, or lack of interest. Nevertheless, there was a positive degree of responsiveness observed from the students in the five classes. In particular, 75% of students (15/20) in class 08-16 provided at least one comment, and in most cases more. The corresponding percentages for the other classes were even higher, corresponding to 86% (18/21) for 09-01, 86% (18/21) for 09-03, 89% (17/19) for 09-05, and 85% (17/20) for 09-07. These higher percentages are an indication of significant interest in providing course feedback to the instructors and, as the qualitative analysis will show, most students took the time to write very explicit, helpful comments expressing their likes and dislikes of the course content. This information will be extremely valuable in making improvements to the course in the future.

On a slightly less positive note, responses to questions 19-22, corresponding to specific feedback about what was learned regarding the four HF skills, reveal the percentage of students responding dropped considerably. While 30% of students in class 08-16 failed to provide comments to the HF skill-oriented items, the percentage of non-responders was higher in the other classes. Specifically, the percentage of non-responders was 76% (16/21), 38% (8/21), 42% (8/19), and 35% (7/20) for classes 09-1, 09-03, 09-05, and 09-07, respectively. The lack of comment data from class 09-01 is particularly troubling since these questions are a major source of data concerning the impact of the course instruction on the targeted HF skills. In some cases, it may be that students simply failed to turn the critique sheet over to answer questions 19-22, or possibly, there may have been insufficient time to provide comments. In any event, it is strongly recommended future classes be given sufficient time and encouragement to complete the comment-oriented questions on the back side of the critique sheet. Despite the absence of comments from all students, the quality of the comments from the students who did respond was quite outstanding and, in many instances, extremely informative.

The other aspect of data quality examined concerns the Likert-scale ratings on items 1-15. As in other analyses, a rough gauge on respondent "effort" in completing the ratings can be obtained by determining how much variability there is in the ratings within a given respondent's data set. That is, if respondents have dismissed ("pencil-whipped") the exercise, then respondents often assign the same rating, be it high or low, to all items. This is, of course, an indication that they are not actually reading the questions since the odds of all aspects of the course generating the same reaction to a truly interested student are fairly low. Hence, a key index considered for rating data quality is the percentage of respondents who have at least "some" variability in their ratings. The analysis showed, across the four classes, only 16-20% of students in any class assigned the same Likert-rating to all 15 questions. Notably, of the 18 students who did this, 14 assigned the maximum rating, 5 (or "strongly agree"). For the other 4 students, 3 assigned all items a "4" and the other student assigned the items a "3." Thus, not only was student interest in assigning quality ratings quite high, with over 80% of them doing so, even the majority of students who "pencil-whipped" their ratings were very favorable about the course. These are all excellent signs of having an effective instructional process in place.

Rating Data.

Using the numeric conversion scheme described above, the average Likert-rating for each critique item over each of the five classes was computed. The critique sheet selections were given numerical values of 1-strongly disagree, 2-disagree, 3-neutral, 4-agree, and 5-strongly agree. Table 3 presents the average of these ratings, where a paraphrased version of the question appears in the left-most column. The overall average rating for each class appears (in bold font) in the bottom row of the table; the corresponding overall averages for each critique item appear in the right-most column.

Table 3

Average Likert-Ratings for the First 15 Items on the CRM Course Critique Sheet

	Class					Item
Critique Item	**08-16**	**09-01**	**09-03**	**09-05**	**09-07**	**Average**
1. Course relevant to job	4.4	4.4	4.7	4.4	4.6	4.5
2. Course covered enough material	4.4	4.0	4.6	4.4	4.5	4.4
3. I learned about CRM	3.8	4.0	4.5	4.3	4.4	4.2
4. I used CRM since last course	4.6	3.6	4.1	4.1	4.3	4.1
5. I will use at least 1 skill	4.5	4.4	4.8	4.5	4.7	4.6
6. Videos of good quality	4.4	3.7	4.4	4.3	4.6	4.3
7. Videos engaging & relevant	4.4	4.1	4.5	4.4	4.5	4.4
8. I learned from others' experiences	4.3	4.3	4.3	3.9	4.0	4.2
9. Instructor organized & well-prepared	4.7	4.6	4.7	4.5	4.8	4.7
10. Instructor had strong teaching skills	4.6	4.6	4.7	4.4	4.8	4.6
11. Instructor was encouraging	4.6	4.6	4.6	4.4	4.8	4.6
12. Instructor shared experiences	4.6	4.4	4.4	4.3	4.7	4.5
13. Instructor translated case studies	4.5	4.4	4.6	4.4	4.7	4.5
14. CRM toolkit well organized	4.3	4.0	4.5	4.1	4.3	4.2
15. I will use CRM toolkit	4.0	3.5	4.2	3.6	4.3	3.9
OVERALL CLASS AVERAGE	**4.4**	**4.2**	**4.5**	**4.3**	**4.5**	4.4

Even a cursory look at the table reveals the ratings in general are quite high. Indeed, the overall average rating, across classes and items, is 4.4, which is roughly halfway between "Agree" and "Strongly Agree" on the critique sheet's Likert scale. Clearly, the students' reactions to the enhanced CRM course were quite positive. This view is further seen in the class averages in the bottom row, where the means ranged from a "low" of 4.2 in Class 09-01 to a high of 4.5 in Classes 09-03 and 09-07. To provide a statistical yardstick by which to view average differences, the average standard error about the mean for each class was computed; then that standard error was averaged. In essence, a difference between averages of .3, or almost a third of a unit difference, can be statistically meaningful (p<.05). By this reckoning, it could be concluded, on average, the students in classes 09-03 and 09-05 (average = 4.5) had a higher overall impression of the course relative to the students in class 09-01 (average = 4.2). The averages for the other two classes fall in between these extremes, and are not statistically different. Nevertheless, it is clear student reactions from *all* the classes are positive. To see how or why one class had a higher or lower average rating, it is necessary to examine how students responded to the individual critique items.

This is seen by examining the far-right column which provides the averages, across the five classes, for each of the 15 Likert-rating items. The items form a logical grouping, in which items 1-5 concern the class as a whole and the utility of CRM skills, items 6-7 pertain to the use of videos, items 9-13 cover the instructor, and items 14-15 concern the CRM toolkit, a paper-based job aid covering the six CRM core areas, passed out to the students during the class.

As before, .3 is used as an index of notable difference to gauge the relative ranking of the critique items and their logical groupings. Thus, the CRM instructor received the highest marks, as his average is 4.5 to 4.7 across the four items in this group. Just slightly lower are the videos, which with average ratings of 4.3 to 4.4 are fairly well-received. On a less positive note, the two items that should be examined further, either through the comments (see below) or with follow-up queries of students themselves, are item #3 (using the CRM skills since the last course) and item #15 (I will use the CRM toolkit). These items, with averages of 4.2 and 3.9, respectively, are lower and suggest future utilization of CRM skills and associated products should be a focal concern for the CTI/Air Force Research Laboratory research team. Of course, examination of the CRM behaviors in the designated training missions will give some additional evidence (Level III), concerning the degree to which the knowledge and skills acquired in the EA (and, eventually, the other spirals) will transfer to the hands-on environment.

Comment Data.

Following the Likert-items, the critique sheet asked for student comments indicating one or two good things about the course, one or two things that need improving, which CRM skills need more training, and how the course helped/or could help each of the four targeted HF skills: attention management, task prioritization, COA selection, and crew coordination. Following are the most frequent (modal) comments for each question, as well any other comments that provide unique and critical insights concerning the CRM instruction delivered. Because qualitative/content analysis is far from an exact science, considerable interpretation is required. However, the concrete nature of the course content and the questions asked make most of the interpretation fairly straightforward. Nevertheless, reviewing the comments in their entirety on the original Excel worksheets would provide for better interpretation and judgment of these comments.

One or two good things about the course.

Comments on this item were placed into the following seven categories: 1) instructor, including his experience, presentation style, and expertise; 2) case examples, including their Predator-specific nature, real-world mishap, and covering a range of possible causes; 3) videos; 4) student interaction, including discussion and benefiting from others' experience; 5) in-class exercises, including dissection of a mishap and/or stimulating student's thinking; 6) presentation of academic concepts, including the core content; and 7) the CRM toolkit.

The most frequent positive comments concerned the use of the scenarios and Predator-specific examples, with a tally of 22. The next most popular topic was the instructor, with 17 positive comments. The other comment categories had the following tallies, indicated in parentheses: videos (14), student in-class discussion (11), core content (6), in-class exercises and dissection of mishaps (5), and the toolkit (1).

One or two things about the course that need improvement.

Given the high positive marks for the course, it is not surprising there were far fewer responses to this item than the previous one. Hence, it was more difficult to categorize the critical comments and construct tallies. Nevertheless, there were several comments that were similar and could be combined. The following is a breakdown of most of the comments concerning course improvement.

The most frequent response to this item was "none" or "NA," which is indicative of the positive view most students had of the course overall. Beyond this response, the next most frequent comment was that the course could have even more examples, as indicated by seven students. Also, four students recommended the course length be shortened, though a time limit was not specified. Two students commented that mission planning/debriefing should receive more attention during the class; and two students requested the scenarios/examples be of an interactive-branching sort, such that the class' choice of action would dictate the outcome of the scenario (this is analogous to adaptive training where the instruction provided is heavily response-driven). Also, two students suggested a Predator-specific course instructor be provided.

The other negative comments or suggestions to this item were indicated by only one student in each case. These suggestions are listed in the following bullets:

- Include a Predator CAS case example
- Provide better HUD footage
- Include a Sensor Operator-oriented scenario
- Provide the course earlier in FTU
- More in-class discussion among experienced and inexperienced aircrew
- Encourage more class participation
- Include positive CRM behaviors in the scenarios
- Include scenarios from other weapons systems
- Improve the CRM toolkit

CRM skills needing more training.

Responses to this item identified which of the seven primary CRM areas (including debriefing) students felt needed more training; they were allowed to check as many areas as they wanted. The purpose of this item was to determine where future efforts in CRM training and instruction might be focused rather than to indicate any particular shortcoming with the present course. However, many students failed to provide any response, most probably an indication of their overall satisfaction with the present course.

The frequency of students' selecting the seven CRM areas across the four classes was tallied. The results are as follows, with the number of students selecting that category in parentheses:

- Situation Awareness (17)
- Flight Integrity/Crew Coordination (14)
- Task Management (12)
- Risk Management/Decision Making (10)
- Debriefing (6)
- Communications (4)
- Mission Planning (4)

Course impact on attention management.

For this and the remaining items, the analysis focused mainly on extracting comments that offer useful insights regarding students' expression of what they learned about the targeted skill. These expressions provide the best and most direct indication of what they are taking out of the course and hopefully applying later in the training curriculum. The focus is not on the quality of word-smithing or clarity of expression, but rather, on what they indicate they are learning.

A number of students offered fairly insightful responses to the question of how the course helped their attention management skills. These included the following, some of which were paraphrased for readability and/or conciseness:

> *It helped me understand how fast fixation can develop into a much bigger problem and how to recognize it in someone else.*

> *It taught me how to ask questions to get my SA.*
>
> *It helped to know the basics and keep my cross-checking going.*
>
> *It reinforced crew coordination and cross-check.*
>
> *It taught me not to fixate on any one thing too long.*
>
> *It taught me to be more aware of what's going on.*
>
> *It showed me how to prioritize my workload.*
>
> *It taught me task management skills to free up brain cells to maintain SA.*
>
> *It taught me to focus attention on the most important tasks.*
>
> *It made me aware that each crewmember should be more aware of their surroundings.*
>
> *It taught me to focus on what's important for the current situation.*

In keeping with the positive reaction of the students to the course, there were far fewer indications of what else could be done in the class to improve attention management (or any of the other skills for that matter). Only two actionable suggestions were provided, both of which apply to all four skills: 1) make the graphics easier to read; and 2) make the course shorter.

Course impact on task prioritization.
There were a comparable number of student responses to the question concerning course impact on task prioritization. As expected, most of the comments were positive, and included the following:

> *It taught me that the aircraft isn't going to fall out of the sky; focus on aircraft control, then prioritize the problems.*
>
> *Make sure we know the aircraft presets as much as possible.*
>
> *It makes you look at the overall picture and realize if you are becoming over-tasked.*
>
> *Refocused on basic airmanship (storms don't care how many hours you have); pay attention to airspeed, alt., etc.*
>
> *Take care of the higher priorities while keeping your crosscheck going.*
>
> *Understanding of fly the aircraft first, checklists second.*
>
> *Focus on aviate, communicate, navigate.*
>
> *It taught me details on task management.*
>
> *It made me realize that tasks change and you must focus on most important task at hand.*
>
> *Make sure not to leave out important decisions.*
>
> *It helped me realize that kinds of tasks should take priority.*
>
> *It gave me a reminder of all the resources available in the UAV.*

There were only several critical comments concerning ways to improve task prioritization training. One student suggested that the scenarios have alternative branching points depending on how the class prioritized the tasks. Another student expressed a preference for learning with hands-on equipment to help learn to prioritize tasks; this desire should be satisfied with the other spirals under development.

Course impact on COA selection.

The student comments to this question were just as insightful as with the two previous HF skills. The comments included the following (some of which are condensed and/or paraphrased):

> *Understanding the BIDE principle and specifically buying time*. (BIDE is a CRM acronym that stands for **B**uy time, **I**dentify problem, **D**ata collection, and **E**xecute)
>
> *Helped remind me to get all the info before jumping into a decision too quickly*.
>
> *When in doubt ask questions and know the basics, collect information from every source*.
>
> *Think about the big picture first for your decisions*.
>
> *Solicit as much help and feedback from the entire crew*.
>
> *Ask for suggestions*.
>
> *It taught me to consider more options and not act preemptively*.
>
> *It showed that it happens early in the plan*.
>
> *It taught me to make a decision and make it timely*.
>
> *It made me realize I should do more earlier, before situations arise*.
>
> *It taught me to use my SO*.
>
> *It gave me examples of who should do what in situations and that you should question if in doubt*.
>
> *It taught me to notice problems before something happens and fix them*.
>
> *Make sure to discuss decisions with the pilot*.
>
> *It gave me a technique of asking, "if we were in the plane would we do this?"*

Again, there were very few negative critiques to this item. One student commented that a student-decided branching scenario would help train this skill more effectively. Two other students felt more emphasis should be given to the importance of sticking with a decision and monitoring to ensure desired results are obtained.

Course impact on crew coordination.

A number of positive comments were obtained regarding student's learning from the course about crew coordination. These comments included the following:

> *Solicit information from everyone, including MIC, SO, SOF, CAOC, etc.*
>
> *The pilot may not always know or understand what is going on in the plane at every single moment – SO input is needed.*
>
> *Make sure we are all on the same sheet of music given all of our various backgrounds*.
>
> *Rank doesn't exist on the airplane; both pilot and SO are responsible for the aircraft and must ask questions if things don't look right and when in doubt call in external resources*.
>
> *Better comm. is needed.*
>
> *It helped with my communications skills.*
>
> *It made me realize that if I speak up problems can be avoided.*
>
> *Learning to work as a crew and understanding that seniority does not mean safe flight.*

> *It's all about communication.*
>
> *Be directive and clear.*
>
> *Solicit inputs from anyone who has information.*
>
> *Need to rely on using my SO better.*
>
> *It taught me the ways to have effective crew coordination starting with effective communication.*
>
> *Crew needs to communicate to work together.*
>
> *I now realize that I have to know least a little about other crewmembers' responsibilities in order to back them up.*
>
> *It taught me not to be afraid of speaking up when I notice something.*

Finally, one student commented that training in this skill could be improved if the course emphasized the SO's potential for helping the crew see the picture.

Level II (Evidence of Learning for EA)

The classes analyzed were 08-16, 09-01, 09-03, 09-05, and 09-07. Recall there were five classes in the Spiral 1 condition, rather than the planned-for two classes, due to the unexpected delay in receiving IRB approval for the project research.

From a logical standpoint, student learning (i.e., Kirkpatrick's Level II) occupies a middle ground between student reaction (Level I – Do they like the instruction?) and transfer of training (Level III – Do they transfer the knowledge and skills they have learned into the operational environment?). In particular, it would seem reasonable to expect students who fail to learn the requisite knowledge-skills-attitude (KSAs) during the instructional intervention will not be in a position to transfer those unlearned KSAs into the operational environment. Put another way, empirical evidence for student learning (Level II) is a necessary but not sufficient condition for transfer of KSAs to the operational environment (Level III). The purpose of this analysis is to make the case that a significant amount of student learning *did* occur over the course of Spiral 1 instruction.

Learning was measured by comparing a given student's performance on 2 tests, 1 given prior to the course (pretest) and a second given after the course (post-test). The pretest and post-test each contained 7 questions, where the items (and their associated foils) were jumbled on the post-test so students were not be able to just use memory to answer them a second time. The 2 tests did indeed "look different" even though they contained the same content. Since the tests were taken during class, there wa s a 100% response rate. Of the 102 students (50 Pilots, 52 Sensors) from the 5 classes, data from 1 student, a Pilot, (in Class 09-05) were discarded. This individual did not take the exercise seriously, as evidenced by creating new options [e.g., a foil e) All of the Above] to some items as well as circling all 4 foils, a-d, to indicate his response. Other than this person, everyone else clearly took the testing seriously, leaving 101 total students for analysis. This clearly constitutes an excellent sample size on which to make assessments of learning magnitude stemming from Spiral 1.

The analysis is divided into two parts. In the first part, pretest and post-test scores are compared to determine how much learning occurred, both by crew position and across classes. The amount of learning (as indicated by the average improvement from pretest to post-test) varied considerably across both classes and crew positions, but was consistently positive indicating a notable learning effect. In the second part, students' performance on each of the seven items is examined, comparing pretest and post-test. During the first two classes, several "weak" items were identified, where content and phrasing needed modification. Data from all five classes collected indicated which of the seven items were primarily responsible for the observed learning effects. A good item is one that is neither too hard nor too easy, since not everyone should choose the correct answer (particularly during pretest) nor the incorrect answer (particularly during post-test).

Assessment of Learning.
Recall there were seven items on each test, so a maximum score was seven for a given student. Post-test scores should be higher than pretest scores, where the difference (or "gain" score) indicates the amount of learning. Note there are several limitations to this approach. The first is, if students came into the course knowing a lot about CRM already, or the test items were really easy, the pretest scores would be high, leaving little room for seeing a gain in the post-test. Second, if students did not think the tests were important (there is no a priori reason why they would think this, but this is always a possibility), one would see pretty much a random distribution of scores with little evidence of movement. Fortunately, there was no evidence of either problem in the data set.

Table 4 shows the average pretest and post-test scores for each class, broken out further by crew position. Looking at the table, there are clearly substantial differences in initial level of knowledge between Pilots and Sensors, with Pilots having almost a full-point advantage (5.5 vs 4.6) over the Sensors on the pretest. This was hardly surprising given the greater experience levels of most Pilots relative to their Sensor counterparts Importantly, demonstrated in the last row of the table, there was a notable learning effect for both crew positions, as Pilots and Sensors each averaged a gain of .6 (just over a half a point) from pretest to post-test.

Table 4

Learning Scores for the Five Spiral 1 Classes

Spiral 1 Class	**Pilots**			**Sensors**		
	Pretest	Posttest	Difference	Pretest	Posttest	Difference
08-16	4.2	5.3	+1.1	4.2	4.3	+.1
09-01	5.2	5.7	+.5	4.5	5.1	+.6
09-03	5.7	6.2	+.5	4.5	6.1	+1.6
09-05	6.2	6.9	+.7	4.7	5.0	+.3
09-07	6.1	6.4	+.3	5.2	5.7	+.5
Overall Average	***5.5***	***6.1***	***+.6***	***4.6***	***5.2***	***+.6***

The interior cells of the table, show that the size of the learning effect (i.e., the size of the difference scores) varied considerably across classes. While all classes "learned," the amount of learning was vastly different. For Pilots, the size of the average difference score ranged from .3 (Class 09-07) to 1.1 (Class 08-16). For Sensors, there was an even wider range, with Class 08-16 showing almost no learning (.1) whereas Class 09-03 exhibited an increase of 1.6 from pretest to post-test. These differences across classes were consistent with the "cohort effect" described in the Level III data analysis summaries, in which, for whatever reason, some classes had a disproportionately higher number of stronger students than did others. Despite these class differences, students, at least on average, demonstrated an increase in their CRM knowledge.

Although the average gain score was sizeable for both crew positions, it was necessary to look at the scores for individual students to ensure the mean differences were not simply due to one or two students exhibiting strong learning effects. That is, it is preferable that the majority of students in each class are contributing to the average gain score differences displayed in Table 4. To make this determination, the gain score for each student was computed individually, and then the number of students having positive, negative (i.e., their scores actually declined from pretest to post-test), or zero (i.e., no change from pretest to post-test) scores was tallied.

These individual gain score tallies are displayed in Table 5. Consistent with the mean scores in Table 4, a majority of students, both Pilots and Sensors, had positive gain scores. Specifically, 53 of the students (26 Pilots, 27 Sensors) had higher scores on the post-test, whereas only 17 students (7 Pilots, 10 Sensors) had negative scores. The remainder, 31 students (16 Pilots, 15 Sensors), showed no change in score from pretest to post-test. The numbers within the classes were consistent with the mean score differences, where again, there was a wide variation across classes in the number of students who actually experienced a learning effect. Indeed, this number ranged from a high of 9 students (Sensors, Class 09-03) to a low of 3 students (Sensors, Class 08-16). Overall, it appears the learning effect was notable and of an equal magnitude for Pilots and Sensors.

Table 5

Frequency of Students Exhibiting Positive, Negative, or Zero Gain Scores in Spiral 1 Classes

Spiral 1 Class	**Pilots**			**Sensors**		
	Positive	Negative	Zero	Positive	Negative	Zero
08-16	6	2	2	3	3	4
09-01	4	2	4	5	1	5
09-03	6	1	3	9	0	2
09-05	6	1	2	4	3	3
09-07	4	1	5	6	3	1
Overall Frequency	***26***	***7***	***16***	***27***	***10***	***15***

Analysis of Individual Test Items.

One of the objectives of the preliminary analysis of Level II data from the Spiral 1 conditions was to analyze the individual test items to determine if any needed to be changed. Test item analysis is like tuning a frequency radio channel – the goal is to maximize the signal coming out (in this case, the item's measure of actual learning) and minimize the noise component (for items, it is making sure the foils are not too hard or too easy). Developing a good test is not easy and can be expensive. For example, the Educational Testing Service (ETS) spends roughly $1000 to develop and validate a single test item for the ACT college entrance exam. Since test item development is part art and part science, there is an inevitable iterative quality to constructing items that challenge a student's knowledge of the course material while ensuring students who actually know the information can answer the item correctly.

In this regard, a good test item is one that is neither too hard nor too easy, and shows substantial improvement when there has been some type of training intervention (in this case, Spiral 1 Focused Academics) that occurs between the first (pretest) and second (post-test) administration of the test. For these purposes, an ideal item is one that about half the students get wrong on the pretest and most get right on the post-test. Table 6 shows the extent to which each of the test items contributed to the learning effects observed in Tables 4 and 5. Specifically, the number of students who answered the item correctly on the post-test relative to the pretest was computed. With this calculation, the more positive the number, the greater value the item provided to the assessment of the learning effect with Spiral 1. The right-hand column gives a synopsis of the item's contribution to the learning effect. Following the table, there is a brief explanation of each item's estimated learning value based on the test results from the five Spiral 1 classes.

Table 6

Pretest/Posttest Gain Scores Associated with Each Learning Test Item

Test Item	**Class**					**Item Assessment**
	08-16	**09-01**	**09-03**	**09-05**	**09-07**	
1. Defense against threats	+7	+4	+6	+2	+3	This was one of the best items on the test.
2. Crewmembers involved in checklist	+3	+1	+1	+1	0	This was an acceptable though not exceptional item.
3. Pilot put down gear	+2	0	+2	+3	+2	This was a moderately useful test item.
4. Which best describes a threat	+5	+4	+5	+2	-2	Other than Class 09-07, this had been a good test item.
5. Flying mission above FL230	0	+4	+3	+3	+5	This was one of the best items on the test.
6. A/C heading toward a mountain	0	-2	0	0	+1	This was a weak test item.
7. Which statement about error is true	-1	0	+3	-1	-1	This was a weak test item.

Item #1: Defense against threats.
This appeared as Item #1 on the pretest and Item #2 on post-test. It consistently exhibited positive gain scores across classes, making it one of the best indicators of learning on the test.

Item #2: Crewmembers intently involved in checklist procedure.
This appeared as Item #2 on the pretest and Item #7 on the post-test. The gain scores on this item were consistently low, either +1 or even 0, as in the last Spiral 1 class (09-07). It contributed very little to the learning effect, making it a weak test item.

Item #3: Pilot put gear down.
This appeared as Item #3 on the pretest and Item # 4 on the post-test. This item was typically missed by a half-dozen or more students, where its gain scores (+2-3) made it a moderately useful test item.

Item #4: Which best describes a threat.
This appeared as Item #4 on the pretest and Item #1 on the post-test. Other than the last Spiral 1 class, this item was consistently an excellent contributor to the learning effect. The drop to a negative gain score in Class 09-07 was surprising, and hopefully, was only a fluke occurrence.

Item #5: Flying mission above FL 320.
This appeared as Item #5 on the pretest and Item #5 on the post-test. Like the first item, this test item consistently showed significant gain scores, making it one of the best items on the test.

Item #6: Aircraft heading toward a mountain.
This appeared as Item #6 on the pretest and Item #6 on the post-test. This was a difficult item, as it was often missed by about half of the students. It did not yield any discernible gain score, making it a weak test item. Part of the problem may have been the length of the question stem, such that reading comprehension entered into the student performance along with content knowledge.

Item #7: Which statement about error is true.
This appeared as Item #7 on the pretest and Item #3 on the post-test. Other than Class 09-03, which exhibited the most learning of any of the Spiral 1 classes, this item performed fairly poorly. It was a fairly easy item and was rarely missed by any of the students. It was the weakest item on the test.

Level III (Transfer of Training for Spiral 1)

Analysis of targeted skills from HF form.
The analysis reported here is based *solely* on the ratings provided by Predator instructors. As such, it does not include comments or the checked behaviors underneath each of the four skills. Of concern with these data is the fact that there was no control over which instructor was doing the rating. Nevertheless, there was evidence for some beneficial effects of Spiral 1 (EA) that rose to the level of statistical significance in a number of cases.

Below is a table of mean ratings organized by a number of factors, including rating dimension, crew position, session, and condition. First, ratings on six dimensions on the HF sheets were collected. These included a rating (from 0 to 4) of attention management, task prioritization, course of action (COA) selection, crew coordination, instructor intervention (i.e., how much did instructors have to intervene to help the student through the session), and crew CRM performance. Next, the means were computed separately for Pilots and Sensors because these are VERY DIFFERENT populations of students and they are being evaluated, often, on different criteria. Their data should not be combined. The mean rating data for the EPE and CO-3 sessions were calculated separately since they were at different points in the curriculum, involved different training environments (Sim vs aircraft), and were scored by different populations of instructors (Stan/Eval vs. mix of contractor and AF instructors). So, like the other factors, data from the two different sessions were not combined.

Finally, a variety of conditions was defined, all of which bear on Spiral 1 and which, collectively, make up the "Spiral 1 comparisons." Looking at Table 7, the top row in each subtable corresponds to the "baseline" condition, consisting of Classes 08-13, 08-14, and 08-15. Because the return rate was pretty weak, there was a need to aggregate the numbers over all three classes to have enough to calculate a meaningful average. Next is the "Control 1" condition. This includes data from Classes 08-17 and 09-02. These were combined to compensate for reduced return rates. Next is Spiral 1, which includes data from three classes: 08-16, 09-01, and 09-03. The return rates were still low at this point, so the data from three classes were combined to achieve a reasonable sample size. Next is Spiral 1-1, which consists of data from Class 09-05. Control 1-1 has data from Class 09-04. Finally, Spiral 1-2 is based on data from Class 09-07, while Control 1-2 has data from Class 09-06. As demonstrated, the later classes had return rates sufficiently high to analyze each one individually. ALL spiral-labeled groups received the EA as their intervention. Five classes received this intervention (08-16, 09-01, 09-03, 09-05, and 09-07) because of the delay of IRB approval before moving forward with the other interventions.

For statistical comparisons, the “spiral” means was compared to the corresponding prior control condition. The Spiral and Control groups were labeled in terms of this comparison. Specifically, the Spiral 1 means were compared to two control conditions: Baseline and Control 1. Spiral 1-1 was compared to Control 1-1, whereas Spiral 1-2 was compared statistically to Control 1-2. To make the statistical comparison, a confidence interval was computed around the Spiral mean, based on the respective variances of the spiral condition and its control counterpart. A confidence interval was computed that would cover a 95% probability, by chance, of having another mean within its range. If the control condition mean was outside this confidence interval, then they were considered statistically different. This test turned out to be conservative since it was not adjusted for the inter-correlations between the six rating dimensions, which could have been done to reduce overall error variance. What this means is, the fact that instructors tended to give similar ratings to the same student on the six dimensions (though they were often not identical but were related) was not taken into account. It was decided not to do this because it would be prohibitively difficult to compute with this study due to incomplete data sets within a session for students (i.e., sometimes there were two sessions for a given student, sometimes just one). While it would have helped, perhaps, to identify some additional cases of significance, the calculation would be very complex. Consequently, there was a conservative testing method, which means the differences reported were likely to be real and not the result of chance variation.

Table 7 reports only means. The means that are statistically different from their counterparts are color-coded in the table, either green, meaning the spiral mean is higher than the control mean, or red, meaning the reverse. Obviously it is preferable to have more green and hopefully no red. There are seven means color-coded in green and one in red. The latter is a case where the control group scored higher than their spiral counterpart. This is not desireable, but it only happened in one case.

Table 7

Mean Ratings for the various Classes Comprising the “Spiral 1 Comparisons” (0-4 scale)

Sensor CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Baseline	2.00	2.00	2.14	2.17	2.33	2.00
Control 1	2.56	2.44	2.67	2.67	2.78	2.56
Spiral 1	2.42	2.50	2.42	2.63	2.71	2.50
Control 1-1	3.13	3.00	3.00	3.31	2.50	2.81
Spiral 1-1	2.65	2.80	2.70	2.80	2.95	2.80
Control 1-2	2.38	2.63	2.25	2.38	2.38	2.38
Spiral 1-2	**3.00**	2.86	**3.00**	3.00	**3.00**	2.71

Sensor EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Baseline	2.56	2.56	2.75	2.78	2.72	2.67
Control 1	3.00	2.88	2.88	3.13	2.88	2.88
Spiral 1	2.70	2.70	2.33	**2.56**	2.38	2.56
Control 1-1	2.60	2.60	2.70	2.67	2.90	2.60
Spiral 1-1	2.56	2.44	2.33	2.61	2.78	2.44
Control 1-2	2.40	2.10	1.90	2.30	2.40	2.10
Spiral 1-2	2.56	2.56	2.44	2.44	2.56	2.22

Pilot CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Baseline	2.50	2.50	2.25	2.50	2.50	2.38
Control 1	2.58	2.92	2.88	3.00	2.83	2.67
Spiral 1	2.73	2.64	2.82	2.68	2.95	2.18
Control 1-1	2.59	2.73	2.73	2.64	2.73	2.36
Spiral 1-1	2.64	2.73	2.55	2.75	2.40	2.50
Control 1-2	2.55	2.55	2.36	2.30	2.50	2.30
Spiral 1-2	3.17	3.00	2.83	**3.00**	3.00	2.33

Pilot EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Baseline	2.81	2.69	2.79	2.90	3.19	2.69
Control 1	2.33	2.33	2.79	2.33	2.67	2.17
Spiral 1	2.88	**3.06**	3.00	**2.93**	3.13	**2.93**
Control 1-1	2.67	2.72	2.78	2.88	2.88	2.38
Spiral 1-1	2.80	2.80	2.80	2.90	3.20	2.80
Control 1-2	2.45	2.55	2.45	2.55	2.73	2.55
Spiral 1-2	2.60	2.50	2.50	2.60	3.10	2.20

Comparing the means revealed some instances where some mean differences, on the order of .60-.70 (i.e., six to seven tenths of a rating point difference), did not reach significance. In these cases, the variability within a condition was so high it was not possible to achieve significance. More often than not throughout the table, the spiral mean was higher than its control counterpart (although there are certainly instances of the reverse), but the difference was not large enough to overcome the high variance to reach significance. Examining the complete table, which includes the sample size and variance, reveals why some of these seemingly large differences did not reach the level of statistical significance.

So where are the significant differences? This appeared for three of the ratings in Spiral 1for Pilots in the EPE session: task prioritization, crew coordination, and CRM performance. There was also significance for Spiral 1-2 for Sensors in the CO-3 session for three dimensions: attention management, COA selection, and instructor intervention. Note these are three different dimensions than above, suggesting the effects are widespread. There was also a Spiral 1-2 effect for the crew coordination dimension for Pilots in the CO-3 session. The lone opposite finding was for Spiral 1 with Sensors in the EPE session for crew coordination.

So what does this all mean? It is difficult to define completely until the comments, behaviors underneath each HF skill/category, and the Creech training gradesheets are included in the data. There appeared to be a general positive advantage of the Spiral classes (i.e., the ones receiving EA) when the within-group (error) variability was low. However, the variability in ratings was quite high, both within a class and particularly between classes. This was not surprising since a large number of instructors provided ratings. Even though the rating sheet had some anchors to guide the ratings, the instructors were not really calibrated in providing the ratings. However, the quantity of data collected revealed evidence of learning in the interventions.

Analysis of negative behaviors from HF form.

This second analysis examined the percentage of negative behaviors (measurement defined below) scored by instructors during the CO-3 and EPE sessions. Recall the first look examined the instructor rating data obtained from the HF measurement sheets. The family of analysis, “Spiral 1 comparisons,” was so named because it combined the low-return rate classes to have sufficient sample sizes to perform valid statistical tests. In particular, the following comparisons were formed, which were repeated in this second-look analysis:

Baseline:	Classes 08-13, 08-14, 08-15
Spiral 1:	Classes 08-16, 09-01, 09-03
Control 1:	Classes 08-17, 09-02
Spiral 1-1:	Class 09-05
Control 1-1:	Class 09-04
Spiral 1-2:	Class 09-07
Control 1-2:	Class 09-06

The logic of the statistical analyses is that Spiral 1 was compared to both Baseline and Control 1 combined whereas Spiral 1-1 and Spiral 1-2 were compared, respectively, to Control 1-1 and Control 1-2. In this way, the spiral-treatment students’ performance was always compared to the immediately-preceding control (non-treatment) class.

This second analysis examined the percentage of students, within a condition (Baseline, Spiral 1, Control 1, etc.), who received a minus on the six to seven behaviors associated with each of the four HF skills. Examples of these behaviors included *effective cross-check*, *cross-check doesn’t stagnate*, and *switches attention* under the Avoids Channelized Attention skill. Note not every instructor used the minus designation, where some simply just scored behaviors as zero (i.e., they left it blank) or a +. Others scored ALL behaviors a +, with the assumed meaning that it occurred. However, the assessment of the HF forms was that the instructors who took the scoring process most seriously, scored some behaviors negative, others positive, and the rest neutral. Consequently, the performance was best revealed with this measure by tallying the number of negative behaviors within a class and then converting that to a percentage so all classes were on a common scale.

For this analysis, the data were combined across Pilots and Sensors to have a sufficient base of possible observations. That is, for a given session and behavior, there might have been anywhere from 13-21 students contributing to the tally. If this number was halved, by doing separate analyses for Pilots and Sensors, there was not enough data to achieve stable (and statistically reliable) percentages. Admittedly, the precision of being able to separately analyze Pilots and Sensors was lost, but it was still possible to determine which behaviors the spiral was affecting. This was a more precise determination than the first analysis, which was only at the skill level (which was the level the ratings were assigned).

The analytic strategy with these data is to start with the most aggregated look and then "unpack" the data by considering more subgroups in subsequent analyses. For the first look, numbers were aggregated across control and treatment (spiral) conditions to get a sense of which behaviors were stronger or weaker than the others. This aggregated tally is shown in Table 8.

Table 8.

Percentage of Negative Behaviors across Session and Condition

Skill/Behavior	Frequency	Percentage
Avoids Channelized Attention		
effective cross-check	23	0.081
cross-check doesn't stagnate	25	0.088
switches attention	19	0.067
adjusts to different cockpits	6	0.021
not distracted by radios	19	0.067
able to shift attn w/o cues	22	0.077
Task Prioritization		
knows high priority task	14	0.049
handle interruptions	17	0.060
returns to interrupted task	9	0.032
can suspend lower priority task	25	0.088
do tasks concurrently	26	0.091
aviate-navigate-communicate	14	0.049
Select COA		
considers all options	25	0.088
facts vs assumptions	20	0.070
avoids hasty decisions	19	0.067
doesn't take too long	24	0.084
ID potential risks	17	0.060
follow-on decisions	20	0.070

Skill/Behavior	Frequency	Percentage
Crew Coordination		
divide tasks	12	0.042
perform team tasks	10	0.035
anticipate info needs	27	0.095
provides timely data	18	0.063
cross-checks others	25	0.088
maintain SMM	21	0.074
convey SMM	12	0.042
N	**285**	

The middle column of the table indicates the frequency or number of students in both control and spiral conditions who received a negatively scored behavior, considering both the CO-3 and EPE sessions together. That frequency was converted to a percentage by dividing each by 285, which was the total number of observations in the sample. Though no statistics were performed, there was a fairly even distribution of percentages across the behaviors, allowing identification of the strongest and weakest behaviors within the Spiral 1 comparison data set. The lowest percentages of negative behaviors (i.e., those for which the aggregate percentage was below .04 [4%]), are left-justified. These were the "strongest" behaviors, which corresponded to *adjusts to different cockpits*, *return to interrupted tasks*, and *perform team tasks*. Note these stronger behaviors came from 3 different HF skills. The right-justified percentages indicate the behaviors which received the highest percentage of negative scores. The table reveals 6 such behaviors: *cross-check doesn't stagnate*, *can suspend lower priority task*, *do tasks concurrently*, *consider all options*, *anticipate information needs*, and *cross-check others*. Again, these represented all the HF skill areas, so no clear trend was apparent.

The second pass through the behavior data peeled away one layer of aggregation, by looking at the percentage of negative behaviors received by students in all the control classes (combined) and all the Spiral classes (combined), where separate tallies were provided for the CO-3 and EPE sessions. This methodology was chosen to compensate for the slightly different class sizes to enable meaningful comparisons. This breakout is depicted in Table 9.

Table 9

Percentage of Negative Behaviors by Spiral and Session

	Control – CO3		Control - EPE		Spiral – CO3		Spiral - EPE	
Skill/Behavior	Freq	Pct	Freq	Pct	Freq	Pct	Freq	Pct
Avoids Channelized Attention								
effective cross-check	6	0.09	9	0.09	2	0.04	6	0.10
cross-check doesn't stagnate	10	0.14	8	0.08	4	0.07	3	0.05
switches attention	3	0.04	8	0.08	2	0.04	6	0.10
adjusts to different cockpits	4	0.06	1	0.01	1	0.02	0	0.00
not distracted by radios	7	0.10	7	0.07	2	0.04	3	0.05
able to shift attn w/o cues	7	0.10	6	0.06	3	0.05	6	0.10
Task Prioritization								
knows high priority task	2	0.03	6	0.06	3	0.05	3	0.05
handle interruptions	3	0.04	6	0.06	5	0.09	3	0.05
returns to interrupted task	3	0.04	3	0.03	2	0.04	1	0.02
can suspend lower priority task	5	0.07	13	0.14	3	0.05	4	0.06
do tasks concurrently	7	0.10	12	0.13	5	0.09	2	0.03
aviate-navigate-communicate	4	0.06	5	0.05	2	0.04	3	0.05
Select COA								
considers all options	6	0.09	11	0.11	2	0.04	6	0.10
facts vs assumptions	8	0.11	8	0.08	2	0.04	2	0.03
avoids hasty decisions	5	0.07	7	0.07	5	0.09	2	0.03
doesn't take too long	7	0.10	10	0.10	3	0.05	4	0.06
ID potential risks	3	0.04	8	0.08	4	0.07	2	0.03
follow-on decisions	5	0.07	9	0.09	3	0.05	3	0.05

	Control – CO3		Control - EPE		Spiral – CO3		Spiral - EPE	
Skill/Behavior	Freq	Pct	Freq	Pct	Freq	Pct	Freq	Pct
Crew Coordination								
divide tasks	4	0.06	5	0.05	2	0.04	1	0.02
perform team tasks	4	0.06	5	0.05	0	0.00	1	0.02
anticipate info needs	8	0.11	8	0.08	8	0.14	3	0.05
provides timely data	5	0.07	8	0.08	2	0.04	3	0.05
cross-checks others	8	0.11	8	0.08	6	0.11	3	0.05
maintain SMM	5	0.07	9	0.09	4	0.07	3	0.05
convey SMM	5	0.07	5	0.05	1	0.02	1	0.02
N	70		96		57		62	

There were 2 statistical tests performed on these data. The first was a simple sign test in which the number of behaviors (out of a total of 25 behaviors evaluated under the four HF skills) that were higher for one condition over the other were compared. This was done for the CO-3 and EPE sessions separately. This type of test is a non-parametric test since it does not involve any assumptions about the underlying distribution of data. It seemed a reasonable attempt since the percentage of a negatively scored behavior was consistently higher for the Control subjects than the Spiral subjects.

In particular, looking at the CO-3 session, 16 of the Spiral cells had lower percentages than the Control; 4 were lower; with 5 ties (the ties were discarded). The probability this breakout arose by chance was computed by comparison to a binomial distribution in which the hypothesized probability was .50 (Miller & Freund, 1965). The resulting p-value was .006, which was statistically significant. For the EPE session, there were 21 Spiral cells with lower percentages than their Control counterparts; 3 were higher; with 1 tie (ties were again dropped). Since the binomial table only goes up to N=20, the normal approximation to the binomial distribution was used to calculate a z-statistic of 3.674, which was significant at the p < .001 level. In both comparisons, there was a statistical advantage for the Spiral students over their Control condition counterparts.

The second test compared each of the Spiral vs. Control percentages on each behavior under the two sessions (CO-3 and EPE). This involved a lot of tests, so a fairly stringent alpha-level was required to avoid inflation of Type I error rate (Harris, 1994). In all cases, the percentages were compared using a z-statistic, corresponding to the equation: z = (p1 – p2)/SQRT[(1-p)*(1/n1)+(1/n2)], where p1 and p2 were the percentages of the two comparison cells, p was the average of the two percentages, and n1 and n2 were the number of scores contributing to the percentages in each case. Applying this statistic to all the cell percentages in Table 9, two cases were found that met or exceeded significance. These are denoted by yellow highlighting, and correspond to the *perform team tasks* in the CO-3 session (z=-1.972) and *do tasks concurrently* (z = -2.262) in the EPE session. In both instances, the Spiral percentage was lower, indicating better performance.

Figure 3 is a combined graphical representation of Table 9 in which the percent of negative comments of each targets' skill were aggregated. The mean of each targeted skill was then calculated and graphed. A Wilcoxon sign test across these combined data revealed this Spiral I reduction was statistically significant (p<.035) for both CO3 and EPE as compared to the control groups.

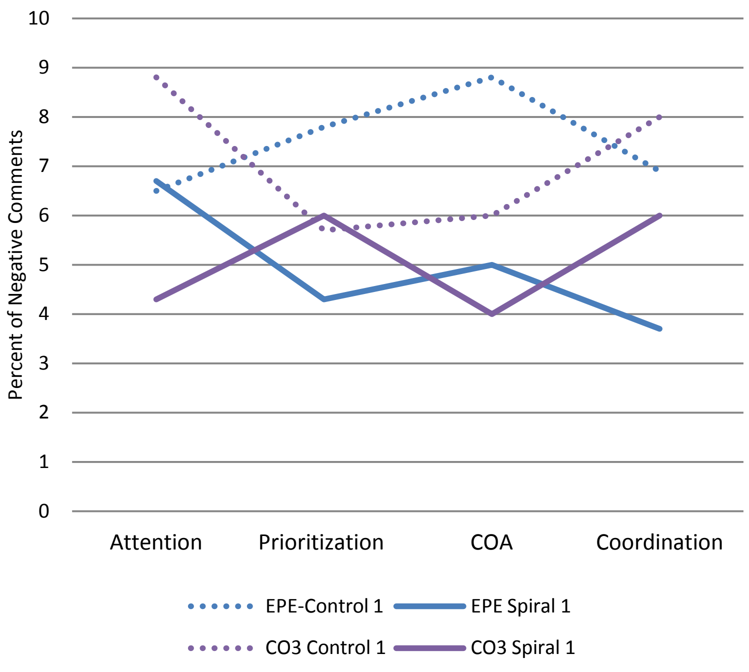


Figure 3: Percent of Negative Comments for Spiral 1

For the final set of comparisons, the Spiral and Control groups were broken out into their subconditions (i.e., Spiral 1, 1-1, 1-2; Baseline, Control 1, 1-1, 1-2) to more precisely pinpoint where the differences in the data resided. These breakouts are depicted in Table 10. The cells having unusually high negative percentages of .15 or higher are color-coded. They are green-shaded when they correspond to the Spiral 1 condition (i.e., the preferred direction) and red-shaded when they correspond to a Control condition (going in the opposite direction of a positive effect of enhanced CRM training). Not surprisingly, there were more green- than red-shaded cells since far more cell comparisons were in favor of the Spiral relative to the Control condition.

Table 10

Percentage of Negative Behaviors by Session and Condition

Skill/Behavior	Baseline		Spiral 1		Control 1		Spiral 1-1		Control 1-1		Spiral 1-2		Control 1-2	
	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE
Avoids Channelized Attention														
effective cross-check		0.11	0.04	0.12	0.14	0.15	0.05	0.11	0.05	0.05		0.06	0.11	0.05
cross-check doesn't stagnate	0.09	0.08	0.13	0.04	**0.24**	**0.20**	0.05	0.05	0.05			0.06	**0.16**	0.05
switches attention	**0.18**	0.08	0.04	0.08	0.05	0.15	0.05	0.11		0.11		0.11		
adjusts to different cockpits	0.09	0.03	0.04		0.05				0.05				0.05	
not distracted by radios	**0.18**	0.08	0.09	0.04	0.05	0.10			0.11	0.11		0.11	0.11	
able to shift attn w/o cues	0.09		0.09	0.04	**0.24**	**0.25**		0.11	0.05	0.05	0.08	**0.17**		
Task Prioritization														
knows high priority task	0.09	0.08	0.13	0.04		0.15		0.05				0.06	0.05	
handle interruptions		0.03	0.09	0.00	0.05	**0.20**	0.10		0.05	0.05	0.08	**0.17**	0.05	
returns to interrupted task	0.09	0.03			0.05	0.10	0.05	0.05	0.05		0.08			
can suspend lower priority task	0.09	**0.17**	0.09	0.12	0.05	0.25				0.11	0.08	0.06	**0.16**	
do tasks concurrently	0.09	**0.17**	**0.17**	0.04	**0.24**	**0.20**	0.05			0.05		0.06	0.05	0.05
aviate-navigate-communicate	**0.18**	0.06	0.09	0.08	0.05	0.05		0.05	0.05	0.11				
Select COA														
considers all options	**0.18**	0.08	0.04	0.12	0.14	**0.25**		0.11		0.11	0.08	0.06	0.05	0.05
facts vs assumptions	0.09	0.08	0.04	0.00	0.14	**0.20**		0.05	0.11	0.05	0.08	0.06	0.11	
avoids hasty decisions	0.09	0.06	0.13	0.04		**0.20**	0.10	0.05	0.05				**0.16**	0.05
doesn't take too long	**0.27**	0.11	0.13	0.04	0.10	**0.20**		0.11		0.11		0.06	0.11	
ID potential risks	0.09	0.06	0.13	0.08		**0.25**			0.05	0.05	0.08		0.05	
follow-on decisions	**0.18**	0.06	0.04	0.12		**0.20**	0.05			0.11	0.08		**0.16**	0.05

Skill/Behavior	Baseline		Spiral 1		Control 1		Spiral 1-1		Control 1-1		Spiral 1-2		Control 1-2	
	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE	CO-3	EPE
Crew Coordination														
divide tasks		0.06	0.09		0.05	0.15		0.05	0.05			0.06	0.11	
perform team tasks		0.06				0.15		0.05	0.05				**0.16**	
anticipate info needs	0.09	0.06	**0.26**	0.08	0.10	**0.25**	0.05	0.05	**0.16**		0.08		0.11	0.05
provides timely data	0.09	0.11	0.09	0.08		**0.25**		0.05	0.05				**0.16**	
cross-checks others	**0.18**	0.14	**0.17**	0.08	0.14	0.15		0.05	0.05		0.15	0.06	0.11	
maintain SMM	0.09	0.06	0.13	0.08	0.05	**0.25**		0.05	0.05	0.05	0.08		0.11	0.05
convey SMM	**0.18**			0.00	0.14	**0.20**		0.05		0.05	0.08			
N	11	36	23	25	21	20	21	19	19	19	13	18	19	21

For statistical effects testing, the z-statistic noted above was again used. The yellow-shaded cells indicate comparisons that approached or exceeded statistical significance. The table indicates they were all in favor of the Spiral, and were all localized in the Spiral 1 cells. It was apparent they cut across HF skills, where the behaviors showing an advantage of the Spiral treatment corresponded to *able to shift attention without cues*, *handle interruption*s, *facts vs assumptions*, and *convey Shared Mental Model*. Though not shown with color coding, there were a number of other comparisons that approached significance, but were excluded to avoid the inflation of Type I error that comes with conducting multiple tests.

Analysis of student gradesheets.

The dependent measures examined with this third analysis were selected training item ratings from the Creech training gradesheets during the CO-3 and the session that immediately precedes EPE, SIM-14.

In laying the groundwork for this analysis, the project team reviewed the training items from the two sessions, for both Pilots and Sensors, to identify those items that "best capture" or represent elemental behaviors that would be facilitated by CRM training. While admittedly subjective, this assessment utilized the subject matter expertise of some half-dozen warfighters and researchers to pinpoint those training items that were the most likely "beneficiaries" of CRM training. These items were certainly not as well-specified (for these purposes) as the human factors skills and behaviors that were delineated for the project's sheets, but they have the advantage of being a normal part of Predator operations training and hence, receive the undivided attention of instructors in assigning grades.

Table 11 provides the list of training items examined for this analysis. The right-hand column indicates the corresponding human factors skill area based on the SME's assessments. Each item, as well as the overall session grade, was graded on a 0 to 4 five-point scale. Virtually all grades in the gradesheets were either a '2' or '3', although there was an occasional '1'. Typically when this happened, the student was required to repeat the session as an extra ("X") ride.

Table 11

Training Gradesheet Items Evaluated as Level III Data

Training Item	Applicable CRM/HF Skill Area
Sensor CO-3	
1. Mission planning/preparation	Mission Planning
2. Admin checks/checklist procedures	Task Management
10. Airmanship	Decision Making & Situation Awareness
14. CRM/crew coordination	Crew Coordination
15. ORM/safety	Decision-making
16. Flight Discipline	Multiple CRM skills
19. Emergency Procedures & Knowledge	Task prioritization, COA, crew coordination
20. Mission checks/checklist procedures	Task Management
Overall Grade	All
Sensor SIM-14 & S-EP-2	
1. Mission planning/preparation	Mission Planning
2. Admin checks/checklist procedures	Task Management
10. Airmanship	Decision Making & Situation Awareness
14. CRM/crew coordination	Crew Coordination
15. ORM/safety	Decision-making
16. Flight Discipline	Multiple CRM skills
18. Emergency Procedures & Knowledge	Task prioritization, COA, crew coordination
Overall Grade	All
Pilot SO-3	
1. Mission planning/preparation	Mission planning
2. Admin checks/checklist procedures	Task Management
8. Airmanship/aircraft control	Decision-making & Situation Awareness
12. ATC Comm/coordination	Communication
17. CRM/crew coordination	Crew Coordination
18. ORM/safety	Decision-making
19. Flight Discipline	Multiple CRM skills
21. Emergency Procedures & Knowledge	Task prioritization, COA, crew coordination
23. Mission checks/checklist procedures	Task Management
43. Tactical communication/coordination	Communication
Overall Grade	All

Training Item	Applicable CRM/HF Skill Area
Pilot SIM-14	
1. Mission planning/preparation	Mission planning
2. Admin checks/checklist procedures	Task Management
8. Airmanship/aircraft control	Decision-making & Situation Awareness
10. ATC Comm/coordination	Communication
15. CRM/crew coordination	Crew Coordination
16. ORM/safety	Decision-making
17. Flight Discipline	Multiple CRM skills
19. Emergency Procedures & Knowledge	Task prioritization, COA, crew coordination
Pilot S-EP-2	
1. Mission planning/preparation	Mission planning
2. Admin checks/checklist procedures	Task Management
8. Airmanship/aircraft control	Decision-making & Situation Awareness
9. ATC Comm/coordination	Communication
14. CRM/crew coordination	Crew Coordination
19. ORM/safety	Decision-making
21. Flight Discipline	Multiple CRM skills
23. Emergency Procedures & Knowledge	Task prioritization, COA, crew coordination
Overall Grade	All

Note. SIM-14 was the designated session through Class 09-04. After that, SIM-14 was split into S-EP-1 and S-EP-2, the latter class was demonstration of skills so that session was used. Training items were the same for the Sensor in SIM-14 and S-EP-2. The items are numbered differently (though labeled the same) for the Pilot. Both lists are included for completeness.

Recall the first and second analysis examined the instructor rating data and negatively-scored behaviors, respectively, from the HF measurement sheets designed specifically for this project. The family of analysis was termed "Spiral 1 comparisons" because it combined the low-return rate classes to have sufficient sample sizes to perform valid statistical tests. In particular, the following comparisons were formed, which was repeated in this final analysis:

Baseline:	Classes 08-13, 08-14, 08-15
Spiral 1:	Classes 08-16, 09-01, 09-03
Control 1:	Classes 08-17, 09-02
Spiral 1-1:	Class 09-05
Control 1-1:	Class 09-04
Spiral 1-2:	Class 09-07
Control 1-2:	Class 09-06

As in the previous analyses, the logic of the statistical analysis was that Spiral 1 was compared to both Baseline and Control 1, whereas Spiral 1-1 and Spiral 1-2 were compared, respectively, to Control 1-1 and Control 1-2. In this way, the spiral-treatment students' performance was always compared to the immediately-preceding control (non-treatment) class.

To set the stage for the analysis, Anacapa project staff reviewed the Creech training folders (in hardcopy form for classes 08-13 through 09-04 and in electronic form thereafter) and extracted the CO-3 and SIM-14/EPE session gradesheets for each student. Then the grades were entered for the training items identified in Table 11 into an Excel spreadsheet. Separate worksheet tabs were created for each of the Spiral 1 comparisons specified at the beginning of this analysis. The statistics resident within Excel were used to calculate the means and variances necessary to perform the required analyses. To make the statistical comparisons, the same method was used as in the first look report: 1) the within-group variances were pooled from the conditions being compared in order to formulate a t-test (Hayes, 1973); and 2) a conservative Bonferroni criterion was used to control Type I error inflation due to making multiple tests (Harris, 1994).

There were three general observations about these data. First, there was considerable variability in the "grades" (or ratings) across both training items and across sessions. The high within-group variances made it hard to achieve statistical significance, even in cases where the mean differences were substantial. Nothing could be done about this. It was for this reason multiple dependent measures were necessary, so the location of the likely effects in the data could be "triangulated."

Second, and an added complication, is there were probably some notable "cohort" effects in the data. That is, it is very possible certain classes had, on average, better students while other classes had weaker students. Since the treatment conditions were assigned by class, it was necessary to be wary of cohort effects (as they could either mask or unfairly enhance the impact of the treatments) and, for this reason, it was necessary to give each spiral at least two distinct classes to spread these cohort effects out across conditions. Of course, there was no way to be sure this balance was acheived, nor do was there any clear index of the size of these cohort effects. However, if a given class consistently received lower (or higher) than average grades on all of the training items, this was certainly suggestive of a cohort effect. As will be seen in the tabled data, there is some indication this might have occurred.

Third, it should be noted Creech AFB made some subtle, yet notable, changes in their gradesheets during the course of the study. Specifically, in the CO-3 session, the Mission Checks/Checklist Procedures items for both the Pilot and Sensor were removed, as was the Tactical Communication/ Coordination item for Pilots. This change occurred prior to Class 09-05. Consequently, there were incomplete data reported for these two training items.

Tables 12-15 present the gradesheet data for the CO-3 sessions (Pilots), SIM-14/EPE sessions (Pilots), CO-3 sessions (Sensors), and SIM-14/EPE sessions (Sensors), respectively. For each table, the columns correspond to the various training items whereas the rows provide the means, variances, and sample sizes (N) for the different subgroups that made up the Spiral 1 comparisons. To permit the mean comparisons to stand out more, the cells contain the variance and N statistics in light-gray font. Using the same convention adopted in the other reports, any statistically significant differences in favor of the treatment (Spiral) are highlighted in green whereas the opposite finding (Control superior to Spiral) is highlighted in red. Recall the Spiral 1 mean was compared to both the Baseline mean and Control 1 mean; the mean for Spiral 1-1 was compared to Control 1-1, whereas the Spiral 1-2 mean was contrasted with Control 1-2.

Looking first at Table 12, there was only 1 significant difference in the CO-3 session for Pilots, which unfortunately was in the opposite direction of that desired. Specifically, the Control 1-1 mean for the ORM training item (2.55) was significantly higher than the Spiral 1-1 mean (2.09). This achieves significance because the within-group variability for Spiral 1-1 (.09) was quite low, allowing the 2 means to have a non-overlapping confidence interval. The means across the training items for Control 1-1, were generally quite high, suggestive of the "cohort effect" mentioned above. Interestingly, the mean grade for the Tactical Communication item, 2.00, was quite low. Unfortunately, this item was no longer included in the gradesheet for the Spiral 1-1 (09-05) class, so it is impossible to determine if an advantage might have materialized there. However, the mean grades for the Tactical Communications items that do exist were uniformly quite low (all below 2.10). The lack of solid performance on this item might have been a reason for the gradesheet change.

Table 12

Gradesheet Data and Analysis Results for CO-3, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	12. ATC Comm	17. CRM	18. ORM	19. Flight Discipline	21. EPs	23. Msn Checks	43. TAC Comm	Overall Grade
Baseline (mean)	2.18	2.18	2.55	2.27	2.55	2.36	2.36	2.27	2.18	2.09	2.18
Baseline (VAR)	0.16	0.16	0.27	0.22	0.27	0.25	0.25	0.22	0.16	0.09	0.16
Baseline (N)	11	11	11	11	11	11	11	11	11	11	11
Spiral 1 (mean)	2.39	2.21	2.36	2.29	2.36	2.43	2.39	2.25	2.29	2.07	2.21
Spiral 1 (VAR)	0.25	0.17	0.24	0.21	0.24	0.25	0.25	0.27	0.21	0.07	0.25
Spiral 1 (N)	28	28	28	28	28	28	28	28	28	28	28
Control 1 (mean)	2.54	2.23	2.62	2.38	2.54	2.62	2.62	2.23	2.23	2.08	2.31
Control 1 (VAR)	0.44	0.19	0.26	0.26	0.27	0.26	0.26	0.19	0.19	0.08	0.23
Control 1 (N)	13	13	13	13	13	13	13	13	13	13	13
Spiral 1-1 (mean)	2.45	2.36	2.18	2.09	2.18	**2.09**	2.18	2.09			2.27
Spiral 1-1 (VAR)	0.27	0.25	0.16	0.09	0.25	0.09	0.16	0.09			0.22
Spiral 1-1 (N)	11	11	11	11	11	11	11	11			11
Control 1-1 (mean)	2.55	2.45	2.55	2.18	2.45	**2.55**	2.64	2.27	2.45	2.00	2.50
Control 1-1 (VAR)	0.27	0.27	0.27	0.16	0.27	0.27	0.25	0.22	0.27	0.00	0.29
Control 1-1 (N)	11	11	11	11	11	11	11	11	11	11	11
Spiral 1-2 (mean)	2.63	2.50	2.50	2.50	2.63	2.38	2.13	2.25			2.36
Spiral 1-2 (VAR)	0.27	0.29	0.29	0.29	0.27	0.27	0.13	0.21			0.25
Spiral 1-2 (N)	8	8	8	8	8	8	8	8			8
Control 1-2 (mean)	2.45	2.45	2.27	2.27	2.45	2.45	2.45	2.27			2.27
Control 1-2 (VAR)	0.27	0.27	0.42	0.22	0.47	0.47	0.47	0.22			0.42
Control 1-2 (N)	11	11	11	11	11	11	11	11			11

Table 13 presents the Pilots' data for the other session of interest, SIM-14/EPE. Once again, there was 1 significant result in the opposite direction desired. As can be seen, it resided in the Spiral 1-1 vs. Control

1-1 comparison again, this time for the Flight Discipline training item. This difference seemed to be less the result of a "cohort effect," since the means for Control 1-1 were not uniformly high. The difference in this case stemmed from an unusually high mean on that particular training item (2.60), coupled with relatively low variance (0.16) from the Spiral 1-1 group. No other difference in the table reached significance, where a typical difference to achieve significance would be on the order of .40 (i.e., four-tenths of a grade).

Table 13

Gradesheet Data and Analysis Results for SIM-14/EPE, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	10. ATC Comm	15. CRM	16. ORM	17. Flight Discipline	19. EPs	Overall Grade
Baseline (mean)	2.64	2.45	2.27	2.27	2.55	2.40	2.36	2.00	2.45
Baseline (VAR)	0.25	0.27	0.42	0.22	0.27	0.27	0.45	0.00	0.27
Baseline (N)	11	11	11	11	11	11	11	11	11
Spiral 1 (mean)	2.50	2.38	2.43	2.32	2.39	2.24	2.46	2.22	2.32
Spiral 1 (VAR)	0.26	0.24	0.25	0.23	0.25	0.19	0.26	0.18	0.23
Spiral 1 (N)	28	24	28	25	28	25	28	27	28
Control 1 (mean)	2.54	2.50	2.33	2.50	2.50	2.25	2.58	2.08	2.17
Control 1 (VAR)	0.27	0.28	0.24	0.27	0.45	0.21	0.27	0.27	0.33
Control 1 (N)	13	10	12	12	12	8	12	12	12
Spiral 1-1 (mean)	2.36	2.18	2.27	2.18	2.18	2.09	**2.18**	2.09	2.18
Spiral 1-1 (VAR)	0.25	0.17	0.22	0.16	0.16	0.09	0.16	0.09	0.16
Spiral 1-1 (N)	11	11	11	11	11	11	11	11	11
Control 1-1 (mean)	2.30	2.33	2.50	2.22	2.30	2.22	**2.60**	2.20	2.40
Control 1-1 (VAR)	0.23	0.25	0.28	0.19	0.23	0.19	0.27	0.18	0.27
Control 1-1 (N)	10	9	10	9	10	9	10	10	10
Spiral 1-2 (mean)	2.40	2.11	2.22	2.33	2.44	2.56	2.56	2.22	2.20
Spiral 1-2 (VAR)	0.27	0.11	0.19	0.25	0.28	0.28	0.28	0.19	0.18
Spiral 1-2 (N)	10	9	9	9	9	9	9	9	10
Control 1-2 (mean)	2.40	2.40	2.40	2.30	2.40	2.30	2.40	2.30	2.30
Control 1-2 (VAR)	0.27	0.27	0.27	0.23	0.27	0.23	0.27	0.23	0.23
Control 1-2 (N)	10	10	10	10	10	10	10	10	10

Table 14 presents the Sensor data for the CO-3 session. There were 2 significant differences in the table, 1 in each direction. For the Mission Planning item, there was, once again, the control mean (2.60), this time from Control 1, higher than Spiral 1 (2.24). This did not look like a "cohort effect" since the training items were not uniformly high for Control 1 nor uniformly low for Spiral 1. The other difference was in the Airmanship item, where the Spiral 1-1 mean (2.50) was significantly higher than Control 1-1 (2.10). This, too, looked like a real effect as there was normal variability across training item means for both groups.

Table 14

Gradesheet Data and Analysis Results for CO-3, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	20. Msn Checks	Overall Grade
Baseline (mean)	2.29	2.21	2.33	2.46	2.39	2.38	2.13	2.17	2.25
Baseline (VAR)	0.22	0.17	0.23	0.26	0.25	0.24	0.12	0.14	0.20
Baseline (N)	24	24	24	24	24	24	24	24	24
Spiral 1 (mean)	**2.24**	2.07	2.18	2.54	2.29	2.29	2.07	2.14	2.21
Spiral 1 (VAR)	0.19	0.07	0.15	0.26	0.21	0.21	0.07	0.13	0.17
Spiral 1 (N)	29	27	28	28	28	28	28	28	28
Control 1 (mean)	**2.60**	2.16	2.32	2.68	2.42	2.42	2.17	2.05	2.32
Control 1 (VAR)	0.20	0.14	0.23	0.23	0.26	0.26	0.15	0.05	0.23
Control 1 (N)	19	19	19	19	19	19	18	19	19
Spiral 1-1 (mean)	2.40	2.00	**2.50**	2.50	2.40	2.40	2.10		2.30
Spiral 1-1 (VAR)	0.27	0.00	0.28	0.28	0.27	0.27	0.10		0.23
Spiral 1-1 (N)	10	10	10	10	10	10	10		10
Control 1-1 (mean)	2.30	2.10	**2.10**	2.40	2.40	2.22	2.11	2.10	2.11
Control 1-1 (VAR)	0.23	0.10	0.10	0.27	0.27	0.19	0.11	0.10	0.11
Control 1-1 (N)	10	10	10	10	10	9	9	10	9
Spiral 1-2 (mean)	2.30	2.10	2.20	2.40	2.10	2.20	2.10		2.33
Spiral 1-2 (VAR)	0.23	0.10	0.18	0.27	0.10	0.18	0.10		0.25
Spiral 1-2 (N)	10	10	10	10	10	10	10		10
Control 1-2 (mean)	2.30	2.11	2.20	2.20	2.00	2.00	2.00	2.00	2.10
Control 1-2 (VAR)	0.23	0.11	0.18	0.18	0.00	0.00	0.00	0.00	0.10
Control 1-2 (N)	10	9	10	10	10	10	10	1	10

The final analysis, Sensors in the SIM-14/EPE session, is depicted in Table 15. The 1 significant difference in the table was again in the opposite direction desired. In this case, the mean Administrative Checks item for Control 1-2 (2.20) was significantly higher than the Spiral 1-2 mean (2.00). This appeared to be a cohort effect, in that the means for all Spiral 1-2 training items were uniformly depressed. This wass particularly reflected in the zero variance in the Administrative Check cell (which partly explains the significant difference), which indicated all 11 students had the same score, '2'.

Table 15

Gradesheet Data and Analysis Results for SIM-14/EPE, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Baseline (mean)	2.25	2.17	2.38	2.17	2.25	2.33	2.26	2.21
Baseline (VAR)	0.46	0.49	0.24	0.41	0.20	0.23	0.38	0.35
Baseline (N)	24	24	24	24	24	24	23	24

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Spiral 1 (mean)	2.18	2.14	2.39	2.29	2.25	2.29	2.16	2.18
Spiral 1 (VAR)	0.30	0.35	0.25	0.36	0.19	0.21	0.14	0.30
Spiral 1 (N)	28	28	28	28	28	28	25	28
Control 1 (mean)	2.17	2.33	2.39	2.39	2.33	2.33	2.13	2.22
Control 1 (VAR)	0.14	0.32	0.26	0.34	0.23	0.23	0.20	0.25
Control 1 (N)	19	19	19	19	19	19	16	19
Spiral 1-1 (mean)	2.30	2.40	2.40	2.40	2.40	2.50	2.30	2.20
Spiral 1-1 (VAR)	0.23	0.27	0.49	0.71	0.27	0.50	0.23	0.40
Spiral 1-1 (N)	10	10	10	10	10	10	10	10
Control 1-1 (mean)	2.13	2.13	2.50	2.25	2.50	2.38	2.13	2.25
Control 1-1 (VAR)	0.13	0.13	0.29	0.21	0.29	0.27	0.13	0.21
Control 1-1 (N)	8	8	8	8	8	8	8	8
Spiral 1-2 (mean)	2.18	**2.00**	2.18	2.09	2.18	2.18	2.18	2.00
Spiral 1-2 (VAR)	0.16	0.00	0.16	0.29	0.16	0.16	0.16	0.20
Spiral 1-2 (N)	11	11	11	11	11	11	11	11
Control 1-2 (mean)	2.30	**2.20**	2.20	2.20	2.20	2.10	2.30	2.20
Control 1-2 (VAR)	0.46	0.18	0.18	0.40	0.18	0.10	0.46	0.40
Control 1-2 (N)	10	10	10	10	10	10	10	10

Conclusions and Recommendations.

The extremely positive reactions of students to the EA course were very encouraging and suggest any future revisions should be more “fine-tuning” rather than extensive. Students in all five Spiral 1 classes rated the Likert-scale items positively for the most part, with an average rating mid-way between “Agree” and “Strongly Agree.” The instructor quality was consistently high across the classes and was one of the major selling points of the instruction. As well, the Predator-specific case examples that populate course content were well-received by virtually every student, though several students would have liked to see even more included. The videos were also lauded, though elimination of technical glitches would be appreciated. Other highly rated features include student participation, particularly learning from student interactions. Two points that should be addressed in the future include ensuring students are both instructed to and encouraged to apply the CRM skills they have learned to future training and operational situations. Also, the CRM toolkit received somewhat mixed reviews. It would be advisable to query students directly concerning the toolkits and whether improvements in formatting, organization, and/or content would be desirable.

Regarding future CRM skill instruction, students most often cited SA as the one area they would like additional training in. Also highly desirable for further instruction were task management and flight integrity/crew coordination. Many students were extremely articulate in expressing what they learned from the course in each of the four targeted HF skills. One possible use of this information in the future might be to transform some of these statements into training objectives that could be posed to the students at the outset of the course. In addition, these objectives could be used to stimulate and vet the Predator-specific examples that presently populate the course. Another point to consider for future development is the use of adaptive case histories in which students’ COA selection dictates the “path” the scenario subsequently follows. With video-constructed scenarios, multiple branch points could be developed for each main decision that could be rewound and shown to students so they could see what might have happened if another decision had been made. This response-contingent branching has exciting pedagogical potential and there is certainly ongoing training research in related areas that could be exploited to utilize this technology in future implementations of Spiral 1.

Looking at the pattern of data from Table 6, it was evident four items (1, 3, 4, and 5) were responsible for virtually all of the observed learning effects. The other three items (2, 6, and 7) showed little or no improvement in performance from pretest to post-test. This finding – that only some of a test's items were good indicators of learning – is fairly common in applications such as this. These items were not changed for the other spirals, though modifying these three items should be considered (either their wording, the foils, or perhaps the content itself) for future use. In that way, there is the opportunity to create an even more sensitive test of Level II student comprehension and learning.

There was a fairly robust and statistically reliable effect of the Spiral treatment in the Level III behavioral data, which seemed to cut across the four HF skills. While the effects were not overwhelming, they were persistent and were evaluated with two different measures: the ratings the instructors provided at the HF skill level, and the more in-depth assessment at the behavior level. Clearly, these results were encouraging and identification of a positive impact of CRM training with Level III data has not been typically found in past studies. This was certainly an encouraging development and was the preferred impact.

On balance, it is probably fair to conclude the analysis of Creech gradesheets did not shed very much light on the effects of Spiral 1 training. Only five differences were significant, and two of these (CO-3, Pilots; SIM-14/EPE Sensors) were most likely the results of a cohort effect. The large within-group variability, which for statistical testing purposes counts as error variance, made it difficult to discern the impact of real treatment effects. Of the other three differences, two were in the opposite direction to that desired, while the third was to the advantage of a Spiral condition. Interestingly, all of the significant differences were observed in different training items, making it difficult to spot any trends or patterns. As was noted at the outset, the Creech training items were not specifically designed to highlight the four Human Factors skills of focus in the training. Indeed, this was why the HF Measurement Sheet was created. Thus, it was not surprising the gradesheets served as a less-than-desirable source of data.

SPIRAL 2 Analysis (EA + ICH)

Level I (Student Critiques of ICH)

The Web-based Interactive Case Histories (ICH) were implemented along with EA as Spiral 2. ICH was designed as a self-study training module MQ-1 student Pilots and Sensor Operators were to complete during the two-week period following their EA and prior to being scheduled for CO-3 and EPE training sessions. The critiques were implemented as online surveys, via the SurveyMonkey service, accessed with a link to the main Birds of Prey website. These student critiques, corresponding to a Level I Kirkpatrick analysis, were filled out by each class receiving ICH training. Given the study design for this project, classes receiving ICH training fell under Spiral 2 (EA + ICH), Spiral 3 (EA + ICH + Multi-tasking Trainer [MTT]), and Spiral 4 (EA + ICH + MTT + Gemasim Team Trainer [GTT]). The data described in this analysis pertain to the two classes that received Spiral 2 training and one that received ICH in beta-test form. Since the latter was fully implemented and was successful from a technology standpoint, their data are included in this analysis. Thus, the Level I data described for ICH came from Classes 09-09, 09-11, and 09-13.

Survey Methodology.
A 41-question student critique survey was developed, adapted for a SurveyMonkey format, and put on the Birds of Prey website as a link. Most questions were either yes/no or multiple choice, with several asking for more detail via comments. During the ICH introduction, given by a CTI instructor following the EA course, students were asked to complete the survey once they finished 2 (of 4) ICH modules. Over the 3 classes, a total of 22 students responded to the survey. These included 11 Pilots, 10 Sensors, and 1 student who did not indicate their crew position. While this was not an overwhelming response rate, only about 33%, it was sufficient to perform the Level I analyses reported herein.

To simplify the analysis and presentation, the data from all three classes were combined to create a suitable sample size for interpretation. Where appropriate, the data were broken out for the two crew positions. SurveyMonkey provides a convenient facility for tallying responses by survey item, and contains a viewing capability by which student comments can be examined for each item for which that option was included. The survey items were organized into four categories: technical issues, usability, usefulness, and case history use. The survey data are summarized for each category.

Technical Issues.
Before considering issues of usability and utility of ICH, it was important to establish whether students had any technical difficulties in operating or accessing the system. The first item asked whether students experienced any sluggishness in system response, owing either to poor connectivity or limited bandwidth. Of the 21 students responding to this item, only one indicated having any notable system lag. When asked further whether this lag caused any negative impact on system use, the lone student indicated “no.” A related item asked students whether they encountered any delays when accessing the video clips present in each ICH module. Of 19 respondents, again only one reported encountering a delay. Again, there was no negative impact of lag on user experience indicated by that student.

Students were then asked whether they had any trouble accessing ICH, either because a computer in the learning lab was unavailable or because the server hosting the Birds of Prey website was down. Of 18 respondents to this item, no student expressed problems in gaining access to ICH. Finally, students were asked whether they encountered any error messages when working through the ICH modules. Out of 19 students responding to this item, no student indicated receiving an error message during their learning experience. In sum, it appears there were no notable technical difficulties in using ICH as a Web-based training capability.

Usability Data.
The first survey item in the usability category asked for students' reactions to the overall look and feel of the ICH. The responses are tabulated in Table 16. Of the 20 respondents, 15 (75%) viewed the ICH as positive or very positive. Further analysis of the data indicated the Sensors were, on average, somewhat more positive in their assessment of ICH look and feel. This was evident in their greater frequency of "very positive" responses and fewer "neutral" responses.

Table 16

Frequency of Student Reaction to Overall "Look and Feel" of ICH

Survey Response Item	Pilot	Sensor	Total
Very Negative	---	----	
Negative	----	----	
Neutral	4	1	5
Positive	6	7	13
Very Positive	0	2	2

The next question was how long students took reading the ICH materials, including the tutorial and whichever modules they reviewed. A total of 14 students provided time estimates for their module review. The average amount of time spent reviewing the ICH materials was just under 37 minutes. Breaking this out by crew position, Pilots spent an average of 31 minutes reviewing the materials compared to 39 minutes for Sensors. These estimates were in line with what was recommended by the instructor and indicate students took the assignment seriously.

The next usability item asked students how "overall hard or easy" it was to use ICH. A five-point scale was again used; their frequency of responding is shown in Table 17. Of 18 respondents, 16 reported ICH to either be Easy or Very Easy to use. Half of the Sensors responding to this item rated ICH as Very Easy to use.

Table 17

Frequency of Student Reaction to Overall Ease of Use of ICH

Survey Response Item	Pilot	Sensor	Total
Very Hard	---	----	
Hard	----	----	
Not Easy or Hard	1	1	2
Easy	4	4	8
Very Easy	3	5	8

Subsequent items in this part of the survey probed specific features of the ICH interface and whether they contributed to or detracted from overall usability. In particular, students were asked how usable they found: 1) the checklist-based format; 2) progressive addition of information; 3) knowledge and supporting information links; and 4) color-coded CRM behavior links. Their frequency of responding to these items is displayed below, in the multi-part Table 18. It is apparent from the table each feature was viewed as quite usable, with two-thirds to three-fourths of the students reporting the feature to be either

Easy or Very Easy to use. Again, the tendency was for Sensors to rate the usability of each specific feature higher relative to Pilots. Also, although all four features were rated highly usable, the estimates were somewhat lower for the color-coded CRM behavior links.

Table 18

Frequency of Student Reaction to the Usability of Specific Aspects of ICH

Survey Response Item	Pilot	Sensor	Total
1. Checklist Format			
Very Hard	---	----	
Hard	----	----	
Not Easy or Hard	3	2	5
Easy	4	3	7
Very Easy	2	5	7
2. Progressive Addition of Information			
Very Hard	---	----	
Hard	----	----	
Not Easy or Hard	4	1	5
Easy	3	4	7
Very Easy	3	5	8
3. Knowledge and Supporting Information Links			
Very Hard	---	----	
Hard	----	----	
Not Easy or Hard	3	1	4
Easy	4	3	7
Very Easy	3	6	9
4. Color-Coded CRM Behavior Links			
Very Hard	---	----	
Hard	----	----	
Not Easy or Hard	1	1	2
Easy	7	4	11
Very Easy	1	4	5

Utility Data.

The next section of the survey addressed the usefulness or utility of ICH and its associated information. Students were first asked to rate how useful they found the information provided in ICH. The results to this four-choice item are shown in Table 19. It is evident from the data there was a split between the two crew positions, as all but one of the Sensors rated ICH information as either Moderately Useful or Very Useful, whereas half of the Pilot respondents reported ICH as Not Useful. Given the great experience levels reported for most Pilots relative to Sensor students, this difference in utility ratings is not surprising.

Table 19

Frequency of Student Reaction to Usefulness of ICH Information

Survey Response Item	Pilot	Sensor	Total
Not Enough Information to Answer	1	1	2
Not Useful	4	0	4
Moderately Useful	4	4	8
Very Useful	1	5	6

Students were then asked to rate the usefulness of four aspects of ICH: 1) the Knowledge Information link; 2) Supporting Information link; 3) CRM Behavior links; and 4) the Wrap-up window. The frequency of responses to these items is summarized in the multi-part Table 20. Each feature was rated as either Moderately Useful or Very Useful by at least two-thirds of the respondents, with the Sensors giving every feature much higher utility ratings relative to the Pilots. Overall, the CRM behavior links were rated somewhat lower, whereas the Wrap-up window occupied a middle ground, as a majority of its ratings were in the Moderately Useful category.

Table 20

Frequency of Student Reaction to the Utility of Specific Aspects of ICH

Survey Response Item	Pilot	Sensor	Total
1. Knowledge Information links			
Not Enough Information to Answer	1	0	1
Not Useful	3	0	3
Moderately Useful	4	5	9
Very Useful	2	5	7
2. Supporting Information links			
Not Enough Information to Answer	1	0	1
Not Useful	3	0	3
Moderately Useful	4	6	10
Very Useful	2	4	6
3. CRM Behavior links			
Not Enough Information to Answer	1	1	2
Not Useful	3	0	3
Moderately Useful	5	5	10
Very Useful	1	3	4

Survey Response Item	Pilot	Sensor	Total
4. Wrap-up Window			
Not Enough Information to Answer	0	0	0
Not Useful	2	0	2
Moderately Useful	7	8	15
Very Useful	0	2	2

For the final item in this part of the survey, students were asked whether they would recommend ICH to other aviators. As expected, the responses to this question tracked those obtained for the other utility items; the response frequencies are shown in Table 21. Specifically, a majority of the Sensors were Likely or Very Likely to make this recommendation, whereas the Pilots were more divided in their opinion, with only half falling in either the Likely or Just Not Sure category.

Table 21

Frequency of Student Response to Recommending ICH to Other Aviators

Survey Response Item	Pilot	Sensor	Total
Not Enough Information to Answer	---	1	1
Very Unlikely	2	0	2
Unlikely	2	0	2
Just Not Sure	3	2	5
Likely	3	5	8
Very Likely	0	1	1

Case History Use and Value.

The remainder of the survey asked students questions about which case histories they read (they were asked to read any two of the four modules on the website) and then indicate their interest level of that case history. The frequency of responses to these items is displayed in Table 22, where the first row in each segment indicates the number of students who reported having read the item; the subsequent rows show the interest level expressed by those students who read that module and then their assessment of the value of that case history for their future training.

Table 22

Frequency of Student Reaction to Use and Value of Individual ICH Case Histories

Survey Response Item	Pilot	Sensor	Total
1. Read APPROACH and LANDING case history	0	2	2
Not Interesting	---	0	
Moderately Interesting	---	2	2
Very Interesting	---	0	
Valuable for Student's Future Training	---	0	
Not Valuable	---	0	
Moderately Valuable	---	2	2
Very Valuable	---	0	
2. Read THUNDERSTORM case history	9	6	15
Not Interesting	0	0	0
Moderately Interesting	6	4	10
Very Interesting	2	4	6
Valuable for Student's Future Training			
Not Valuable	2	0	2
Moderately Valuable	4	1	5
Very Valuable	1	4	5
3. Read LOST LINK case history	8	5	13
Not Interesting	2	0	2
Moderately Interesting	2	3	5
Very Interesting	1	2	3
Valuable for Student's Future Training			
Not Valuable	3	0	3
Moderately Valuable	5	2	7
Very Valuable		3	3
4. Read OVER-CONTROL case history	3	4	7
Not Interesting	0	2	2
Moderately Interesting	1	1	2
Very Interesting	2	2	4
Valuable for Student's Future Training			
Not Valuable		1	1
Moderately Valuable	2	0	2
Very Valuable	1	3	4

For the most part, the results in Table 22 mirror the previous pattern of data for usability and utility. In particular, Sensors gave higher ratings for case history interest and training value compared to Pilots. There was also considerable variability in which case histories the students read. Specifically, only two students, both Sensors, read the Approach and Landing case history which, given the mission/in-flight focus of their job duties, is not surprising. On the other hand, the Thunderstorm and Lost Link case histories were read with the highest frequency, as virtually all the Pilots in the sample read both case histories. The interest and training-value ratings were especially high for the Thunderstorm case history, with Lost Link receiving slightly lower ratings by the Pilots.

Level II (Evidence of Learning for ICH)

ICH was implemented as a Web-based, self-study training module which MQ-1 student Pilots and Sensor Operators were to complete during the several-week period following their EA and prior to being scheduled for CO-3 and EPE training sessions. Students could either complete the training in one of the Learning Lab computers or on their own personal laptop. As directed by the CTI CRM Academics instructor, students were to complete at least two of the four case histories residing in the ICH directory.

Given the study design, introduction of ICH to the CRM curriculum constituted part of Spiral 2 in that EA was already inserted into the curriculum as a Spiral 1 intervention. Spiral 2, then, consisted of receiving *both* EA and ICH. Subsequent classes received the Spiral 3 intervention (EA, ICH, and MTT) or the Spiral 4 intervention (EA, ICH, MTT, and GTT). Since the four interventions are experienced in independent sessions, each is subjected to its own Level II analysis and reported separately.

From a logical standpoint, student learning (i.e., Kirkpatrick's Level II) occupies a middle ground between student reaction (Level I – Do they like the instruction?) and transfer of training (Level III – Do they transfer the knowledge and skills they have learned into the operational environment?). In particular, it would seem reasonable to expect students who fail to learn the requisite knowledge-skills-attitudes (KSAs) during the instructional intervention will not be in a position to transfer those unlearned KSAs into the operational environment. Put another way, empirical evidence for student learning (Level II) is a necessary but not sufficient condition for transfer of KSAs to the operational environment (Level III). The purpose of this analysis is to make the case that a significant amount of student learning *did* occur over the course of Spiral 2 training involving ICH.

Logic of the Analysis.
Recall that learning was assessed in the EA course by comparing student scores on a 7-item pretest (taken prior to the start of the course) with that of a comparable 7-item post-test (taken right after the course was completed). Learning was measured directly by comparing each student's post-test score with their corresponding pretest score. A positive difference – meaning they scored higher on the post-test than on the pretest – was a direct reflection of how much they learned as a result of taking the course. In an analysis of these gain scores as reported earlier, it was determined that, on average, Spiral 1 students gained .6 of an item upon taking the course. For a 7-item test, such a gain constituted a performance increase of 9% – a strong indication of learning.

Assessment of learning for ICH was, however, not as straightforward as it was for EA. This was because a pretest/post-test was not an appropriate comparison since much of the ICH learning was about a particular case history which defied pretesting (i.e., it was not reasonable to expect students to know anything about a particular case history prior to using ICH). So what measures could be used to index learning for ICH? First to consider was the anticipated learner outcomes. There were two expected aspects to learning. First, ICH users should have learned *how* CRM principles could be applied to address the problems that arose in the description of the mishap case history. Knowledge items that tested students' understanding of those principles was one index, where their absolute performance (relative to some chance baseline) was an indirect measure of learning, with "chance" meaning the performance level that would have been obtained if students guessed or selected responses in the absence of any true knowledge. With a four-foil multiple choice test as was used, chance performance would correspond to 25% correct. Performance significantly above chance would clearly have been an encouraging sign of learning.

This was certainly not the strongest test of learning, though, since students' knowledge as Pilots and aviators coming into ICH training might have led them to have better performance than a non-aviator, such that chance performance may not have been the best baseline for comparison. So, another reference point was if the average performance level reached 70% correct, a typical requirement to pass academic courses in military training. Clearly, this was a much more stringent test, but it could be argued this was a more valid comparison since this would have been the level required to "pass" a student if he/she took ICH as a true academic course.

Yet another reference point came through comparing the performance of Pilots and Sensors. Specifically, it can reasonably be argued Pilots, by virtue of their greater experience (on average) and flight training, should have performed better than Sensors on any ICH quiz. However, if the performance of Sensors was brought in line with that of Pilots, so that average performance levels were fairly close, this would also have been convincing evidence that actual learning occurred as a result of using ICH.

In addition, students' learning about ICH as a computer-based system (i.e., its interface, features, and functions) could also have been assessed through improved performance as a function of exposure to and experience with the capability. While a performance test of ICH operation could have been embedded into the software, this would likely have added unwanted complexity and degraded the student's learning experience. Alternatively, an inference of performance capability was made by computing the percentage of students who, while originally registering on the Birds of Prey website, failed to complete a review of two ICH modules and their quizzes. As above, this was an indirect measure of learning since other factors may have contributed to student attrition, such as lack of time, boredom, or disinterest. Nevertheless, observing a higher percentage of registered students who completed the two ICH quizzes was a good, albeit, indirect measure of learning that was combined with the performance indices above to triangulate on an inference of degree or amount learning.

Data Collection.
As students worked through a given ICH case history, they got to the end of the checklist and then completed a four-item, four-foil multiple choice quiz. Once they were confident of their answers, they pressed a SUBMIT button which automatically sent an e-mail of their answers to the Birds of Prey website server. Prior to implementing ICH on the website, a relational database was created to store the response data with a sorting feature based on date, user name, crew position, class, and ICH case history. With this organization, it was a straightforward manner to calculate the percentage correct for any combination of these database fields. Interrogation of the database allowed learning assessment using each of the indices described above.

The data described below were obtained from five classes that received ICH. The first was Class 09-09, which served as the ICH beta-test. It was successful from a participation and technology standpoint, so the data were retained and classified as the first Spiral 2 class (since they also received EA). The next two Spiral 2 classes reporting ICH data were Class 09-11 and 09-13. Finally, data are reported from Classes 09-15 and 10-01 which are both Spiral 3 classes, as they received EA and MTT in addition to ICH. Since ICH was experienced separately from the other interventions, it is appropriate to look at all classes that received ICH, regardless of spiral, in one analysis. In the analyses described below, performance both within class and across classes is examined. Note students were free to select the two case histories they reviewed. Consequently, some case histories have a larger sample size than others. Therefore, it was sometimes necessary to combine data across classes to achieve a suitable sample size for analysis.

Quiz Performance Scores as an Index of Learning.
The overall quiz scores for the 4 case histories and 2 crew positions are tallied in Table 23. A total of 87 students, 46 Pilots and 41 Sensors, contributed data for this analysis. Looking first at the totals in the last

row of the table, the average percentage correct, across all case histories and all classes, was 77.2% for Pilots and 60.3% for Pilots. Both percentages were substantially above 25%, the chance performance level for a 4-foil test, and this is borne out statistically. Thus, the cumulative percentage correct for Pilots, 77.2%, was significantly (z = 20.81, p < .001) larger than 25%. The same was true for the Sensors' cumulative percentage correct of 60.3% (z = 14.59, p < .001). (The z-tests reported here are based on the inferences of single proportions as described by Miller and Freund [1965, p. 194]). Clearly, then, the quiz performance for both groups of students well exceeded that which would be expected solely on the basis of chance.

Table 23

Percentage Correct on ICH Quizzes by Crew Position and Case History

	Pilot (N = 46)			Sensor (N = 41)		
ICH Case History	**#Correct**	**Total**	**Pct**	**#Correct**	**Total**	**Pct**
Landing Approach	31	40	77.5	30	44	68.2
Overcontrol	59	76	77.7	35	56	62.5
Thunderstorm	83	132	62.9	86	144	59.7
Lost Link	90	116	77.6	42	76	55.3
TOTAL	263	364	77.2	193	320	60.3

The next benchmark previously mentioned involved comparing average quiz performance against a typical 70% passing mark for an academic class. For the Pilots, this assessment was easy since their average, 77.2%, exceeded the 70% mark. To examine this figure in more detail, the Pilots' average performance scores for each case history and each class are presented in Table 24. Those cells that failed to meet the 70% benchmark are shaded in red. As can be seen, the averages for most classes and most case histories met or exceeded the 70% criterion, although there were some exceptions. This tended to occur for at least one case history in each class, except for Class 09-09, where every case history's average was larger than 70%. With regard to the classes themselves, 4 of the 5 classes had an average quiz score above 70%. The lone exception was Class 09-15, in which the average of 62.5% failed to reach the 70% benchmark by a few percentage points.

Table 24

Percentage Correct on ICH Quizzes for Pilots, By Class and Case History

	Class				
ICH Case History	**09-09**	**09-11**	**09-13**	**09-15**	**10-01**
Landing Approach	91.6	---[a]	91.7	75.0	**50.0**
Overcontrol	87.5	75.0	85.0	**40.0**	95.0
Thunderstorm	71.9	71.4	**45.8**	62.5	**56.2**
Lost Link	87.5	**67.9**	79.2	75.0	87.5
TOTAL	81.6	70.3	72.5	**62.5**	73.2

[a] No Pilot from this class selected the Landing Approach case history.

The quiz averages for Sensor students were lower, as can be seen in Table 23, where the average for every case history was below 70%. However, the averages were still somewhat close to 70%, particularly for Landing Approach and Overcontrol. While the overall average for Sensors, 60.3%, was statistically lower than 70% (z= 3.78, p < .001), the averages for 2 of the case histories, Landing Approach (z = .263, p > .602) and Overcontrol (z = 1.301, p < .11), were not. Given the Sensor students' relative lack of CRM knowledge going into the course, achievement of this level of performance on the ICH quiz was fairly notable evidence of learning.

Finally, the other performance comparison mentioned above, between Pilots and Sensors, was also a fairly positive index of learning. These comparisons are most evident in Table 23, where across the 4 case histories, the average difference in performance between the 2 crew positions ranged from a low of only 3.2% (Thunderstorm) to a high of 22.3% (Lost Link). The overall difference in performance quiz score was 16.9%. Statistically, the average score for Sensors was significantly lower than Pilots (z = 4.68, p < .001). (This is based on the z-test of several proportions described in Miller & Freund [1965, p. 194]). However, the differences in crew position quiz averages did not reach statistical significance for either the Landing Approach (z = .954, p < .18) or Thunderstorm (z = 545, p < .30) case histories. This statistical indifference in crew position for several case histories wass another sign of learning, where the quiz performance of Sensors was at times indistinguishable (at least statistically) from Pilots.

Student Attrition as an Index of Learning.

The other approach to measuring learning was to make an inference from student attrition rates, where the number of students who initially registered on the Birds of Prey website was compared with the corresponding number who actually completed the ICH quizzes. These rates were determined for each class and crew position, and the tallies are presented in Table 25. The bottom row shows the completion rates for the two crew positions across all five classes. The completion rates were high, over 80%, for both Pilots and Sensors. This certainly provides evidence the students were able to navigate the ICH interface, its features, and functions to reach the end of the checklist, access the quiz, provide answers, and submit their quizzes. Looking at the rates for the individual classes, the completion rates exceeded 80% in all cases, other than the Pilots for Class 10-01. This finding is certainly consistent with the inference that students were able to learn the ICH well enough to be competent users of the system.

Table 25

Percentage of Students Completing ICH Training by Class and Crew Position

		Pilots			**Sensors**	
Class	**# registered**	**# completed**	**% completed**	**# registered**	**# completed**	**% completed**
09-09	12	10	83.3	10	5	50.0
09-11	8	8	100	8	8	100
09-13	12	10	83.3	10	9	90.0
09-15	11	11	100	9	8	88.9
10-01	9	7	77.8	11	10	90.0
Overall	52	46	88.5	48	40	83.3

Level III (Transfer of Training for Spiral 2)

Analysis of targeted skills from HF form.

The data set reported here is from the instructor ratings on the Human Factors data sheet implemented at Creech AFB. Recall that these ratings were obtained from two carefully chosen training sessions: the third and final Combined Operations session (CO-3) conducted with an actual aircraft, and the final emergency procedures check (EPE) conducted in the simulator with Standardization/Evaluation (Stan/Eval) raters.

The Spiral 2 comparisons encompassed six classes: 09-08, 09-09, 09-10, 09-11, 09-12, and 09-13. The actual Spiral 2 treatment classes were the odd-numbered ones (i.e., 09-09, 09-11, and 09-13). The other three classes were the non-treatment controls. The Spiral 2 treatment consisted of the EA (which was the lone Spiral 1 treatment) *plus* the Interactive Case Histories (ICH) self-paced computer-based training. Level I and Level II analyses for this intervention were described above. Whereas Class 09-09 constituted a beta-test tryout for ICH, its technical success and adequate participation rate allowed retention of it as an actual Spiral 2 treatment case, giving three Spiral 2 treatment classes for a Level III (training transfer) analysis.

Recall the Spiral 1 analyses were somewhat involved since it was necessary to break out different subgroups of classes because the data collection window for Spiral 1 was so long (due to IRB delays). Fortunately, Spiral 2 comparisons were more direct. They were simply the Human Factors rating data for all three Spiral 2 classes combined for comparison to the data for the three corresponding control classes. For ease of labeling, the three control classes will be referred to as Control 2. Participation rates of 60-80% across the classes gave a fairly substantial sample size (i.e., N = 21 to 29) for computing means and performing confidence interval testing.

Before presenting the results of the mean difference analysis, first how the ratings were distributed across the 5-point (0-4) Human Factors scale is reported. Recall the same 5-point scale was used that Creech AFB instructors used on the training gradesheets, where “0” was essentially unsafe flying and “4” represented exemplary performance. The distribution of the ratings across classes was examined to see how well the instructors used the ends of the scale. Of the 1214 ratings generated across the 6 classes, there were 126 “4” ratings, or 10.4% of the total. While no “0”s were reported, there were 21 “1”s, which comprised 1.7% of the total. Thus, the instructors had just over 12% of their ratings on the 2 ends of the scale, which is a very respectable percentage and is indicative of fairly robust rating scale.

As an aside, the 2 conditions, Spiral 2 and Control 2, were compared to see if there were any differences in how frequently the extreme ends of the scale were used. Only slight differences were found, where Control 2 students received 10.3% “4”s and 2.4% “1”s; this compared to 10.5% and 1.1%, respectively, for Spiral 2 students. Accordingly, whereas the 2 groups received virtually the same number of extremely positive ratings, there was a modest advantage of Spiral 2 students on the low end of the scale, where their percentage was roughly half that of Control 2. It should be recognized, though, these percentages were based on fairly low frequencies, so they should be interpreted with some caution given their inherent instability. However, it is another empirical indication of where the treatment condition had an advantage over the control.

Table 26 presents the means, variances, and sample sizes (N), respectively, for the Spiral 2 and Control 2 classes. The presentation is divided into four segments, corresponding to the CO-3 sessions for Sensors, EPE sessions for Sensors, CO-3 sessions for Pilots, and EPE sessions for Pilots. As with Spiral 1, the data are broken out by crew position since their tasks were so different and because they were rated by different instructors. Within each segment, the data appear in six columns, corresponding to the six dimensions rated on the HF data collection sheet: attention management, task prioritization, COA selection, crew coordination, degree of instructor intervention, and CRM performance.

Table 26

Mean Ratings for the Six Classes Comprising the "Spiral 2 Comparisons" (0-4 Scale)

Sensor CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA Selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 2 (Mean)	2.50	2.37	2.37	2.59	2.74	2.48
Control 2 (variance)	0.34	0.50	0.50	0.42	0.47	0.53
Control 2 (N)	23	23	23	23	23	23
Spiral 2 (Mean)	2.46	2.62	2.42	2.65	2.81	2.52
Spiral 2 (variance)	0.50	0.57	0.57	0.48	0.40	0.68
Spiral 2 (N)	26	26	26	26	26	25

Sensor EPE Sessions

Condition	Attention Management	Task Prioritization	COA Selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 2 (Mean)	2.63	2.42	2.58	2.78	2.76	2.61
Control 2 (variance)	0.51	0.51	0.51	0.63	0.59	0.43
Control 2 (N)	24	24	24	23	23	23
Spiral 2 (Mean)	2.62	2.52	2.57	2.62	2.67	2.57
Spiral 2 (variance)	0.35	0.46	0.26	0.25	0.23	0.26
Spiral 2 (N)	21	21	21	21	21	21

Pilot CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA Selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 2 (Mean)	2.58	2.46	2.58	2.54	2.65	2.62
Control 2 (variance)	0.41	0.50	0.49	0.42	0.40	0.41
Control 2 (N)	26	26	26	26	26	26
Spiral 2 (Mean)	2.62	2.52	2.62	2.64	2.68	2.52
Spiral 2 (variance)	0.39	0.33	0.39	0.39	0.74	0.41
Spiral 2 (N)	29	29	29	28	28	27

Pilot EPE Sessions

Condition	Attention Management	Task Prioritization	COA Selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 2 (Mean)	2.59	2.65	2.69	2.70	3.12	2.44
Control 2 (variance)	0.56	0.42	0.48	0.52	0.91	0.64
Control 2 (N)	27	27	27	27	26	27
Spiral 2 (Mean)	2.68	2.82	2.64	2.68	2.77	2.63
Spiral 2 (variance)	0.60	0.60	0.61	0.60	0.47	0.60
Spiral 2 (N)	28	28	28	28	28	28

To facilitate visual comparison, the variance and N values are displayed in gray within their respective cells so the means are the most prominent measure within each segment. For statistical comparisons, just as in Spiral 1, a confidence interval around the Spiral mean was computed based on the respective variances of the spiral condition and its control counterpart. In essence, a confidence interval was constructed that would cover a 95% probability, by chance, of having another mean within its range. If the control condition mean was outside this confidence interval, then they were considered statistically different. As discussed in the Spiral 1 analysis, this test turned out to be quite conservative since adjustments for the inter-correlations between the six rating dimensions were not made, which could have been done to reduce overall error variance. Such an adjustment was virtually impossible for this data set because of the unequal sample sizes and because there were not both (i.e., EPE and CO-3) sessions worth of data for many of the subjects. Consequently, this was a conservative testing method which means the differences reported were likely to be real and not the result of chance variation.

Visual comparison of the Spiral 2-Control 2 mean differences across the columns and segments revealed fairly small differences, particularly in light of estimated mean error variances on the order of .19-.22. Since the mean difference must be roughly double this value to reach significance (Harris, 1994), there was no mean difference that exceeded a criterion of .38-.44. The largest difference in the table was actually in the opposite direction of that hypothesized, in which the Control 2 mean (3.12) for Instructor Intervention was higher than its Spiral 2 counterpart (2.77) for Pilots in EPE sessions. However, the variance associated with this rating dimension was quite high, where the criterion value for this analysis was .44, indicating a non-significant difference.

While the mean differences were not enough to achieve significance via parametric testing, there was a revealing pattern in the data. Specifically, examining the *pattern* of mean differences in the CO-3 sessions for both Pilots and Sensors, in 10 of the 12 comparisons, the Spiral 2 mean was higher. Using binomial testing procedures, a non-parametric analysis, revealed this would have occurred by chance (assuming each direction was equally likely [i.e., drawing from a binomial population with p = .5]), only 2% of the time (i.e., p < .02) (Miller & Freund, 1965, pp. 212-214). The consistency of the directional differences, despite their small size, was indicative of a slight preference in favor of the treatment (Spiral 2) classes, at least for CO-3. A similar superiority was found for the Spiral 1 HF ratings, as well as with the negative behaviors also recorded on that sheet.

In sum, the results of the non-parametric testing were once again consistent with a general positive advantage of the Spiral 2 classes (i.e., the ones receiving EA and ICH) that appeared to cut across rating dimension and crew position. While they did not reach significance in the parametric sense due to high within-group variance (primarily as a result of having a large number of instructors doing the ratings across students and classes), the trend in differences was certainly consistent with an advantage for the classes receiving treatment. Coupling the fairly positive student reaction data (Level I) to the training with the consistent evidence of learning (Level II), a fairly substantial amount of data supports a highly positive impact of the CRM training interventions.

Analysis of negative behaviors from HF form.

This second analysis examined the percentage of students within a condition, who received a minus on the six to seven behaviors associated with each of the four HF skills. Examples of these behaviors included *effective cross-check*, *cross-check doesn't stagnate*, and *switches attention* under the Avoids Channelized Attention skill. As reported in the corresponding Spiral 1 analysis, note not every instructor used the minus designation, as some simply just scored behaviors as zero (i.e., they left it blank) or a +. Other instructors scored ALL behaviors a +, which was interpreted to mean it was simply observed. On the other hand, the assessment of the HF form was that the instructors who took the scoring process most seriously, scored some behaviors negative, others positive, and the rest neutral. Consequently, the performance was best revealed with this measure by tallying the number of negative behaviors within a class and then converting that to a percentage so all classes were on a common scale.

The analytic strategy is to start with the most aggregated look and then "unpack" the data, by considering more subgroups in subsequent analyses. For the first look, data were aggregated across control and treatment (spiral) conditions, sessions (CO-3 and EPE), and crew positions to get a sense of which behaviors were stronger or weaker than the others. This aggregated tally is shown in Table 27.

Table 27

Percentage of Negative Behaviors across Conditions, Sessions, and Crew Position

Skill/Behavior	Frequency	Percentage
Avoids Channelized Attention		
effective cross-check	6	.029
cross-check doesn't stagnate	4	.020
switches attention	5	.025
adjusts to different cockpits	1	.005
not distracted by radios	16	.078
able to shift attn w/o cues	5	.025
Task Prioritization		
knows high priority task	10	0.049
handle interruptions	7	.034
returns to interrupted task	6	.029
can suspend lower priority task	9	.044
do tasks concurrently	11	.054
aviate-navigate-communicate	6	.029

Skill/Behavior	Frequency	Percentage
Select COA		
considers all options	7	.034
facts vs assumptions	9	.044
avoids hasty decisions	6	.029
doesn't take too long	10	.049
ID potential risks	5	.025
follow-on decisions	8	.039
Crew Coordination		
divide tasks	2	.010
perform team tasks	3	.015
anticipate info needs	10	.049
provides timely data	8	.039
cross-checks others	6	.029
maintain SMM	2	.010
convey SMM	4	.020
N	**204**	

The middle column of the table indicates the frequency or number of students in both control and spiral conditions who received a negatively scored behavior, considering both the CO-3 and EPE sessions together. That frequency was converted to a percentage by dividing each by 204, which was the total number of observations in the sample. Though no statistics were performed, a fairly even distribution of percentages across the behaviors enabled identification of the strongest and weakest behaviors within the Spiral 1 comparison data set. In particular, the lowest percentages of negative behaviors, (i.e., those for which the aggregate percentage was below .02 [2%]) are left-justified. These were the "strongest" behaviors, and included *adjusts to different cockpits*, *divide tasks*, *perform team tasks*, and *maintain SMM*. Note 3 of these stronger behaviors came from the same HF skill, Crew Coordination. The right-justified percentages indicate the behaviors that received the highest percentage of negative scores, in excess of .045 (4.5%). The table shows 5 such behaviors: *not distracted by radios, knows high priority task, do tasks concurrently, doesn't take too long, and anticipates information needs.* These "weakest" behaviors encompassed all 4 HF skills, so no clear trend was apparent.

The second pass through the behavior data peeled away two layers of aggregation, by looking at the percentage of negative behaviors received by students in the three Spiral 2 classes and three Control 2 classes, where separate tallies were provided for the CO-3 and EPE sessions. This breakout is depicted in Table 28.

Table 28

Percentage of Negative Behaviors by Condition and Session

	Control 2 – CO3		Control 2 – EPE		Spiral 2 – CO3		Spiral 2 – EPE	
Skill/Behavior	Freq	Pct	Freq	Pct	Freq	Pct	Freq	Pct
Avoids Channelized Attention								
effective cross-check	2	0.041	3	0.059	1	0.018	0	0.000
cross-check doesn't stagnate	0	0.000	0	0.000	3	0.055	0	0.000
switches attention	3	0.061	0	0.000	2	0.036	0	0.000
adjusts to different cockpits	1	0.020	0	0.000	0	0.000	0	0.000
not distracted by radios	5	0.102	1	0.020	9	0.164	2	0.041
able to shift attn w/o cues	4	0.082	2	0.039	1	0.018	0	0.000
Task Prioritization								
knows high priority task	4	0.082	1	0.020	4	0.018	1	0.020
handle interruptions	5	0.102	0	0.000	3	0.055	0	0.000
returns to interrupted task	3	0.061	2	0.039	3	0.036	0	0.000
can suspend lower priority task	5	0.102	0	0.000	3	0.000	1	0.020
do tasks concurrently	8	0.163	3	0.059	2	0.164	1	0.020
aviate-navigate-communicate	4	0.082	2	0.039	1	0.018	1	0.020
Select COA								
considers all options	3	0.061	2	0.039	2	0.073	0	0.000
facts vs assumptions	4	0.082	3	0.059	1	0.055	1	0.020
avoids hasty decisions	3	0.061	0	0.000	2	0.055	1	0.020
doesn't take too long	3	0.061	3	0.059	3	0.055	1	0.020
ID potential risks	3	0.061	2	0.039	0	0.036	0	0.000
follow-on decisions	6	0.122	1	0.020	1	0.018	0	0.000

	Control 2 – CO3		Control 2 – EPE		Spiral 2 – CO3		Spiral 2 – EPE	
Skill/Behavior	Freq	Pct	Freq	Pct	Freq	Pct	Freq	Pct
Crew Coordination								
divide tasks	2	0.041	0	0.000	0	0.036	0	0.000
perform team tasks	1	0.020	0	0.000	2	0.018	0	0.000
anticipate info needs	3	0.061	2	0.039	4	0.036	0	0.000
provides timely data	3	0.061	3	0.059	1	0.055	1	0.020
cross-checks others	3	0.061	2	0.039	1	0.000	1	0.020
maintain SMM	2	0.041	0	0.000	0	0.018	0	0.000
convey SMM	3	0.061	0	0.000	1	0.000	0	0.000
N	49		51		55		49	

As with Spiral 1, there were 2 statistical tests performed on these data. The first was a simple sign test comparing the number of behaviors (out of 25) that were higher for one condition over the other. This was done for the CO-3 and EPE sessions separately. This comparison is a non-parametric test since it does not make any assumptions about the underlying distribution of data. It is a reasonable test to employ when it seems there are consistent differences in favor of one group over another, regardless of the magnitude of those differences.

Looking first at the CO-3 session, 20 of the Spiral 2 cells had lower percentages than Control 2; 3 were higher; and 2 were ties (the ties were discarded). The probability this breakout arose by chance was computed by comparison to a binomial distribution in which the hypothesized probability was .50 (Miller & Freund, 1965). The resulting p-value was .0013, which was statistically significant. Thus, the Spiral 2 classes received a significantly lower percentage of negative behaviors from instructors.

Regarding the EPE session, there were many fewer negative behaviors scored; this low frequency was prevalent across both conditions. In terms of a binomial sign test, the prevalence of 0's produced a large number of ties. Disregarding these, Spiral 2 classes had 13 behaviors with lower negative percentages compared to Control 2 and 3 behaviors where that directionality was reversed. If a binomial probability distribution (with hypothesized p = .50) is referenced, the difference is again significant, with p = .0106. Thus, in both comparisons, a statistical advantage for the Spiral students over their Control condition counterparts was seen.

The second test compared the Spiral vs. Control percentages on each behavior under the two sessions (CO-3 and EPE). This entailed conducting a substantial number of tests, so a fairly stringent alpha-level was necessary to avoid inflation of Type I error rate (Harris, 1994). In all cases, the percentages were compared using a z-statistic, corresponding to the equation: $z = (p_1 – p_2)/SQRT[(1-\underline{p})*(1/n_1)+(1/n_2)]$. In this equation, p_1 and p_2 were the percentages of the two comparison cells, $\underline{p}$ was the average of the two percentages, and n_1 and n_2 were the number of scores contributing to the percentages in each case. Applying this statistic to all the cell percentages in Table 28, there were two cases that met or exceeded significance. These are denoted by yellow highlighting, and correspond to the *can suspend lower priority task* (z=-2.385) and *make follow-on decisions* (z = -2.097), both in the CO-3 session. In both instances, the Spiral percentage was lower, indicating better performance.

Figure 4 is a combined graphical representation of Table 28 in which the percent of negative comments of each targets' skill were aggregated. The mean of each targeted skill was then calculated and graphed. A

Wilcoxon sign test across these combined data revealed this Spiral 2 reduction was statistically significant (p < .001) for both CO3 and EPE compared to the respective control group.

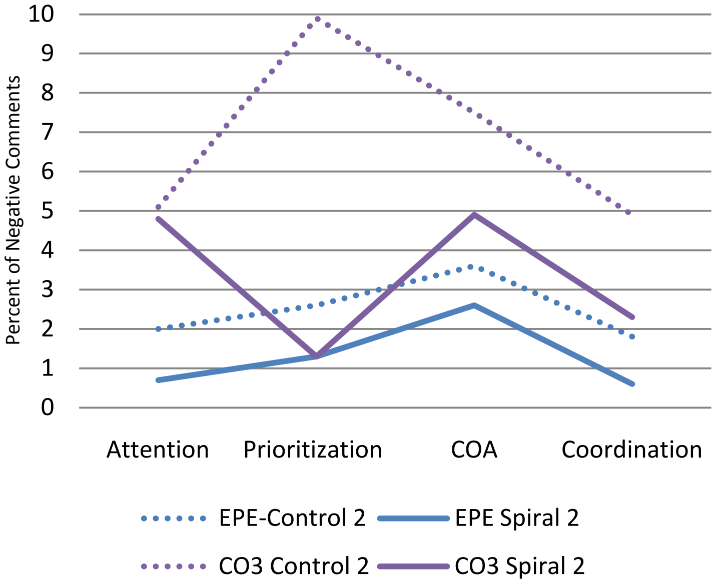


Figure 4: Percent of Negative Comments for Spiral 2

For the final analysis, the percentage of negative behaviors received by the two crew positions was compared. This breakdown is given in Table 29, where only the data from the CO-3 sessions are presented since the EPE session totals were quite small and essentially equivalent for the two crew positions. In addition, the percentages are presented by HF skill, not the specific behaviors themselves, to have sufficient frequencies driving the percentages.

Table 29

Percentage of Negative Behaviors in CO-3 Sessions Received by Crew Position and Condition

	Control 2		Spiral 2	
Human Factors Skill	**Pilot**	**Sensor**	**Pilot**	**Sensor**
Channelized Attention	.077	.026	.076	.038
Task Prioritization	.092	.087	.076	.031
Select Course of Action	.085	.096	.034	.031
Crew Coordination	.045	.072	.034	.023

As can be seen from the table, the Sensors generally had a lower percentage of negative behaviors relative to their Pilot counterparts. However, the differences were not pronounced. This was evident statistically, as only one comparison, that for the Channelized Attention HF skill in the Control 2 classes, achieved significance (z = 1.974, p < .05). Otherwise, the two crew positions received negative behaviors in the same relative proportion.

In closing, once again there was a fairly robust and statistically reliable effect of the Spiral treatment in the behavioral data, which was apparent across multiple HF skills. While the effects were not overwhelming, they were persistent and, importantly, were observed with two different measures: the ratings the instructors provided at the HF skill level and the more in-depth assessment at the behavioral level. The finding of significant Level III transfer of training with a CRM intervention has not been found very often in the literature, so the results were both encouraging (from a practical standpoint) and scientifically quite notable.

Analysis of student gradesheets.
The training items selected for analysis were ones CTI SMEs believed would encompass the skills most likely to benefit from CRM training. These items are displayed in the left column of Table 11, where the corresponding CRM/HF skill areas are presented in the right column. Each item was scored on a 0-4 scale, where most grades were either a "2" or a "3." Students also received an overall grade for the session. Students receiving a "1" on several training items were usually required to take an extra ride (X-ride). Unfortunately, 2 training items that should be particularly sensitive to CRM training, *Mission Checks* and *Tactical Communications*, were deleted from the Pilot's CO-3 gradesheet starting with Class 09-05.

Combining the data from three classes yielded a substantial sample size (N ~ 30) for comparing Spiral 2 and Control 2 performance. With training records in electronic form (starting with Class 09-05), the grades were entered for the training items identified in Table 11 into an Excel spreadsheet, creating a separate tab for Control 2 and Spiral 2. The statistics resident within Excel were used to calculate the means and variances necessary to perform the required analyses. To make the statistical comparisons, the same method was used as in the first look report: 1) the within-group variances were pooled from the conditions being compared in order to formulate a t-test (Hays, 1973); and 2) a conservative Bonferroni criterion was used to control Type I error inflation due to making multiple tests (Harris, 1994).

As in the analysis for Spiral 1, there was considerable "noise" in the data due to having multiple instructors (presumably with different internal criteria) assigning grades across students and classes. This created a larger within-group error variance, making it harder to achieve statistical significance. In addition, there were non-trivial "cohort" effects, in which some classes were simply better or weaker than others because of the makeup of their students (e.g., experience, aviation background, stronger class leader). This, too, made it harder to discern differences due to added noise variance. There is no solution for either problem, since these are a fact of life in doing field research within an operational training squadron. It was for these reasons multiple dependent measures were employed for Level III analysis, so there was "triangulation" on the locus of real effects in the data.

Tables 30-33 present the gradesheet data for the CO-3 sessions (Pilots), S-EP-2 sessions (Pilots), CO-3 sessions (Sensors), and S-EP-2 sessions (Sensors), respectively. For each table, the columns correspond to the various training items, whereas the rows provide the means, variances, and sample sizes (N) for the Control 2 and Spiral 2 conditions. To permit the mean comparisons to stand out more, the cells containing the variance and N statistics are portrayed in light-gray font. Using the same convention adopted in the other analyses, any statistically significant differences in favor of the treatment (Spiral 2) are highlighted in green, whereas the opposite findings (Control 2 superior to Spiral 2) are highlighted in red.

Looking first at Table 30, there was only 1 significant difference in the CO-3 session for Pilots, which unfortunately was in the opposite direction to that desired. Specifically, the Control 2 mean for the CRM training item (2.63) was significantly higher than the Spiral 2 mean (2.34). This achieved significance because, with an estimated confidence interval of .26, the .29 mean difference was outside that range. The rest of the mean comparisons were fairly close and did not differ significantly since anywhere from .24-.30 (depending on the within-group variances) was required for a statistical effect. With 1 exception, the means of the Spiral 2 condition were, disappointingly, lower than their Control 2 counterparts. However, examination of the individual scores revealed that, by chance, one of the Spiral 2 classes had multiple Pilots who required X-rides, so their training item grades (and overall grade) were lower. This is part of the "cohort effect" mentioned earlier.

Table 30

Gradesheet Data and Analysis Results for CO-3 Session, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	12. ATC Comm	17. CRM	18. ORM	19. Flight Discipline	21. EPs	Overall Grade
Control 2 (mean)	2.53	2.34	2.48	2.38	2.63	2.47	2.38	2.53	2.29
Control 2 (VAR)	0.26	0.23	0.26	0.24	0.31	0.26	0.31	0.26	0.41
Control 2 (N)	32	32	31	32	32	32	32	32	31
Spiral 2 (mean)	2.50	2.28	2.38	2.28	2.34	2.38	2.44	2.38	2.24
Spiral 2 (VAR)	0.26	0.21	0.24	0.21	0.23	0.24	0.25	0.24	0.37
Spiral 2 (N)	32	32	32	32	32	32	32	32	34

Table 31 presents the Pilots' data for the other session of interest, S-EP-2. These means were all close in value and, hence, there were no significant differences. Interestingly, the Spiral 2 means tended to be higher, though not significantly so. Examination of the individual gradesheets revealed only a couple of Pilots required X-rides, where the number was comparable between the two conditions. Hence, there did not appear to be any "cohort effect" present in the S-EP-2 data for Pilots.

Table 31

Gradesheet Data and Analysis Results for S-EP-2 Session, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	10. ATC Comm	15. CRM	16. ORM	17. Flight Discipline	19. EPs	Overall Grade
Control 2 (mean)	2.29	2.19	2.23	2.32	2.39	2.16	2.20	2.23	2.16
Control 2 (VAR)	0.21	0.16	0.18	0.23	0.25	0.14	0.17	0.18	0.14
Control 2 (N)	31	31	31	31	31	31	30	31	32
Spiral 2 (mean)	2.41	2.28	2.34	2.22	2.44	2.19	2.22	2.19	2.18
Spiral 2 (VAR)	0.25	0.21	0.23	0.18	0.25	0.16	0.18	0.16	0.15
Spiral 2 (N)	32	32	32	32	32	32	32	32	33

Table 32 presents the Sensor data for the CO-3 session. There was one significant difference in the table, again for the CRM training item and, once again, in the opposite direction desired. Examining the individual gradesheets revealed a number of Sensors required X-rides in the Spiral 2 classes, whereas only 1 Sensor required such remediation in the Control 2 condition. Thus, the cohort effect appeared to complicate the analysis once again. The other mean differences were small and did not approach the .24-.28 value required.

Table 32

Gradesheet Data and Analysis Results for CO-3 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 2 (mean)	2.24	2.21	2.42	2.67	2.29	2.25	2.04	2.21
Control 2 (VAR)	0.19	0.17	0.25	0.23	0.22	0.20	0.04	0.17
Control 2 (N)	25	24	24	24	24	24	24	28
Spiral 2 (mean)	2.29	2.10	2.29	2.33	2.19	2.19	2.10	2.22
Spiral 2 (VAR)	0.21	0.09	0.21	0.23	0.16	0.16	0.09	0.18
Spiral 2 (N)	21	21	21	21	21	21	21	23

The final analysis, for Sensors in the S-EP-2 session, is depicted in Table 33. There was one significant difference, for Airmanship, in which the Spiral 2 mean was significantly higher than the Control 2 mean. Examination of the individual gradesheets revealed this difference appeared despite the fact that the Spiral 2 classes had a number of X-ride-required students, whereas the Control 2 classes did not have any. Thus, the difference was obtained despite a cohort effect working against the Spiral 2 interventions. Interestingly, all the mean differences in the table were in favor of Spiral 2 even if the rest did not reach significance.

Table 33

Gradesheet Data and Analysis Results for S-EP-2 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 2 (mean)	2.17	2.17	2.17	2.45	2.14	2.21	2.10	2.10
Control 2 (VAR)	0.29	0.15	0.15	0.26	0.12	0.17	0.17	0.10
Control 2 (N)	29	29	29	29	29	29	29	29
Spiral 2 (mean)	2.19	2.24	2.43	2.48	2.29	2.29	2.24	2.23
Spiral 2 (VAR)	0.16	0.19	0.26	0.26	0.21	0.21	0.19	0.18
Spiral 2 (N)	21	21	21	21	21	21	21	22

Conclusions and Recommendations.

Overall, the results of the Level I survey presented a fairly clear picture of student reaction to ICH technical access, usability, and utility. First, the Web-based implementation of ICH was successful, as all students reported having no difficulties with accessing ICH or using the system within the bandwidth and connectivity limits at Creech AFB. Second, ICH usability was rated quite highly by most students, with Sensors consistently giving the capability higher marks compared to Pilots. The crew position differences were most apparent in the ratings of ICH usefulness. Sensors tended to rate the usefulness of ICH information and training value of the case histories as moderately high to very high, whereas only half of the Pilots in the sample expressed that opinion.

In terms of future use of ICH, several recommendations should be considered. First, ICH acceptance and use would be higher if it were more closely tied in with the EA course. For example, it would be helpful if the instructor used ICH in class to work through one of the case histories, by pointing out relevant CRM behaviors and key performance breakdowns responsible for the mishap. This would not only give students expert insights into the human factors of the mishap, it would also illustrate how the system can be used to receive a deeper level of CRM training. Second, the differences in reaction to ICH by Pilots and Sensors were no doubt real and deep-seated. In this regard, it is likely the case the ICH approach to case history review is not going to be perceived as necessary or helpful to highly experienced aviators, who are representative of the Pilots in the sample. It is possible as the Pilot student cohort becomes less experienced (e.g., as part of the Air Force Chief of Staff's experiment on using non-rated aviators as UAV Pilots), the perceived value of ICH training by Pilot students will increase. Third, the advisability of adding one or two case histories to the ICH "library" of mishaps should be considered. In particular, it would be helpful to develop several new case histories that deal with in-flight issues and which address problems of higher interest to Pilots (e.g., flight control). Such additions might encourage greater ICH use by Pilot students, thereby increasing its perceived training value. Overall, the results of this Level II analysis present a fairly clear and consistent picture of student learning during ICH training. Although the absence of a pretest/post-test comparison required examination of indirect measures of learning based on absolute performance and student attrition rates, all of the indices examined were supportive of an inference that learning occurred. Thus, absolute quiz performance of both Pilots and Sensors exceeded chance (25%) levels in all case histories and for all classes. Against a more stringent benchmark of 70%, the Pilots' average was above that in four of the five classes. While the Sensors' average was somewhat lower, they were statistically equivalent to 70% in two of the four case histories. Also, student Sensors' performance levels were comparable to Pilots in two case histories, another indication of ICH impact on learning. Finally, student attrition rates were consistently low, under 20%, an indication students were able to master the ICH interface and use its functionality to take the quizzes and demonstrate their understanding of the material.

In retrospect, it would be advisable to include some type of case history-independent pretest students take before looking at their first case history or reviewing the tutorial. The pretest would tap an understanding of basic CRM principles but would not require knowledge of the case history specifics. In this regard, it would perhaps make the most sense to create a pretest that would be applicable to both the EA and ICH. With this approach, the EA post-test and the ICH exit quiz would each be used to index the particular intervention's impact on learning. Additionally, a very simple performance test of ICH capability (e.g., find a checklist, show a CRM behavior link, access a Knowledge Item) could be created that would index the other aspect of ICH learning. Alternatively, a brief survey could be administered to students following ICH use (perhaps after the quiz) asking them to rate their confidence in using various aspects of the system. While not as direct as an actual performance test, it would be easier and quicker to administer and, as evidenced, would likely yield comparable results.

In any event, it is certainly possible to create more direct measures of learning and embed them into a procedure of system use. If the interventions studied in this project receive continued use at Creech AFB or elsewhere, it would be advisable to consider such tests so a methodologically stronger paradigm is put in place to gauge each intervention's impact on student learning. The analysis of the student gradesheets revealed a substantial cohort effect made finding any significant effect of the training interventions quite difficult. Interestingly, there were consistently higher Spiral 2 mean ratings for training items in the S-EP-2 sessions, whereas there was an opposite trend in CO-3. While the cohort effect was responsible for the significant advantage for Control 2, it was not present for the one significant difference in favor of Spiral 2. As discussed in the Spiral 1 analysis, the training gradesheets represented a fairly flawed and "noisy" source of performance data. Not only were they susceptible to obscuring cohort effects, the training items themselves were, at best, only partial reflections of the specific CRM skills impacted with the training interventions.

Finally, note the one difference found in favor of Spiral 2 was with the Sensors. The Level I/II analyses revealed Sensors were especially positive toward receiving additional CRM training, where all of the interventions were highly valued. This indicates there might be greater transfer of training (Level III) with this crew position as well.

SPIRAL 3 Analysis (EA + ICH + MTT)

Level I (Student Critiques of MTT)

The game-based Multi-Task Trainer (MTT) was implemented along with EA and ICH as Spiral 3. MTT was a self-study training module MQ-1 student Pilots and Sensor Operators completed during the several-week period following their EA and prior to being scheduled for CO-3 and EPE training sessions. They were also tasked to complete the ICH self-study modules during this interval. Students could either complete the training in one of the Learning Lab computers or on their own personal laptop.

The critiques were implemented as online surveys, via the SurveyMonkey service, accessed with a link to the main Birds of Prey website. These student critiques, corresponding to a Level I Kirkpatrick analysis, were filled out by each class receiving MTT training. Given the study design for this project, classes receiving MTT training fell under Spiral 3 (EA + ICH + MTT) and Spiral 4 (EA+ ICH + MTT + Gemasim Team Trainer [GTT]). The data described in this report pertain to the two classes that received Spiral 3 training and one that received MTT in beta-test form. Since the latter was fully implemented and was relatively successful from a technology standpoint, their data are included in the analysis. Thus, the Level I data described for MTT came from Classes 09-14 (beta-test), 09-15, and 10-01.

Survey Methodology.
A 39-question student critique survey was developed, adapted for a SurveyMonkey format, and put on the Birds of Prey website as a link. Most questions were forced choice multiple choice (i.e., select only one foil), with several items permitting multiple foil selection and several others asking for comments. During the MTT introduction given by the CTI instructor for EA, students were asked to take the survey once they completed MTT training consisting of 4 familiarization scenarios, 4 training scenarios, and 8 test scenarios. Though the training was put on the student's course schedule for that several week period, the MTT modules were self-study and not required per se. Over the 3 classes, a total of 34 students responded to the survey. These included 22 Pilots and 12 Sensors. While this was not an overwhelming response rate, only about 57%, it was sufficient to perform the Level I analyses reported herein.

To simplify the analysis and presentation, the data from all three classes were combined to create a suitable sample size for interpretation. For most analyses, the data for the two crew positions were broken out separately. SurveyMonkey provides a convenient facility for tallying responses by survey item, and contains a viewing capability by which student comments can be examined for each item for which that option was included. The survey items were organized into four categories: technical issues, usability, usefulness, and training preparation afforded by MTT and the value of its four individual tasks. The survey data are summarized for each category.

Technical Issues.
Because MTT had 4 tasks running simultaneously, the underlying JavaScript programming had to have multiple clocks and counters operating synchronously. This placed considerable demands on the system CPU, which may have experienced difficulties when implemented within the limited bandwidth environment of Creech AFB. Consequently, it was imperative the Level I survey contain multiple probes to ascertain the extent and nature of system lags or other technical problems students experienced. The first item asked if students encountered any problems with "system responsiveness," such as slow page loading times, mouse clicks not working properly, or issues with scoring inaccuracies due to system lags. Half of the sample reported experiencing some type of system non-responsiveness, where that 50% rate was obtained with both Pilots and Sensors. When asked whether this lack of responsiveness impacted their use of MTT, 2 students (1 Pilot, 1 Sensor) indicated it had a "large negative impact," while another 12 respondents (8 Pilots, 4 Sensors) rated it as having a "moderate negative impact."

In examining the student comments, it is clear the most frequent impact was on the students' perception that they lost points, either because of making multiple mouse clicks or because the scores were simply not tallied in time due to system lag. Thus, students indicated the "system was slow to refresh the timer, so some answers would come too late to count." Importantly, the software modifications made after Class 09-15 were specifically designed to address problems that can occur at the end of a trial, so if the subject responded with only one or two seconds before the trial times out, his/her response would not be counted. This issue was addressed and, encouragingly, these types of comments did not appear in the Class 10-01 survey data. Not only did this class report fewer instances of technical problems (20%), the nature of the problems appeared to have a reduced impact on the scores they received.

The next three survey items probed the technical difficulties issue in more detail; students were asked whether they encountered any problems with audio (for the tones) during MTT use. Only 2 (1 Pilot, 1 Sensor) of the 34 respondents reported encountering problems with the audio, with only 1 student (Sensor) citing it as a "moderate negative impact." The students were then asked whether computer access was a problem for using MTT, either due to computer unavailability or the Birds of Prey website server being down. Only one respondent (a Pilot) cited this as an issue, characterizing this problem as having a "large negative impact" on MTT use.

Finally, students were asked if they encountered any error messages during MTT use. Three students, all Pilots, reported receiving such messages. Of these, two indicated the messages as posing a "moderate negative impact" on MTT use. In these cases, note these more serious problems were experienced by students from classes *before* the software changes were made; none of the students from Class 10-01 reported experiencing problems that had either a "large" or "moderate negative impact."

Usability Data.

The first survey item in the usability section asked for students' reactions to the overall look and feel of the MTT. The responses are tabulated in Table 34. Of the 34 respondents, 17 (50%) viewed the MTT as "positive or very positive," with just under half (14 or 41%) rating the MTT's look and feel as "neutral." Only 3 students (9%) rated the interface as "negative." Pilots and Sensors had essentially comparable, and mixed views of the MTT look and feel.

Table 34

Frequency of Student Reaction to Overall "Look and Feel" of MTT

Survey Response Item	Pilot	Sensor	Total
Very Negative	1	0	1
Negative	1	1	2
Neutral	8	6	14
Positive	8	4	12
Very Positive	4	1	5

The next question was how long students spent working through the MTT scenarios, including time spent reading the tutorial. A total of 23 students provided time estimates of their module review. The average amount of time spent using the MTT was just under 42 minutes. Time estimates varied widely, ranging from a reported low of only 2 minutes to a high of 90 minutes. However, the average estimate was virtually identical for Pilots and Sensors, with Pilots spending an average of 41.9 minutes on MTT and Sensors, 41.5 minutes. As with ICH, these estimates were in line with instructor recommendations and indicate (most) students took the assignment seriously.

The next usability item asked students how "overall hard or easy" it was to use MTT. A 5-point scale was again used; their frequency of responding is shown in Table 35. Of 34 respondents, 23 reported the MTT to either be Easy or Very Easy to use. Only 2 students (6%) reported the MTT as "hard" to use.

Table 35

Frequency of Student Reaction to Overall Ease of Use of MTT

Survey Response Item	Pilot	Sensor	Total
Very Hard	0	0	0
Hard	1	1	2
Not Easy or Hard	8	2	10
Easy	9	7	16
Very Easy	5	2	7

Two subsequent items in this part of the survey probed specific features of the MTT interface and whether they contributed to or detracted from overall usability. In particular, students were asked how usable they found the 1) scenario index page; and 2) MTT tutorial. Their frequency of response is displayed in Table 36. It is apparent from the table both features were considered quite usable, with 61-70% reporting the feature to be either "easy or very easy" to use. Only 2 and 1 student, respectively, rated the Scenario Index Page and Tutorial as "hard" to use.

Table 36

Frequency of Student Reaction to the Usability of Two Features of MTT

Survey Response Item	Pilot	Sensor	Total
1. Scenario Index Page			
Not Enough Information to Answer	0	1	1
Very Hard	0	0	0
Hard	2	0	2
Not Easy or Hard	6	3	9
Easy	9	6	15
Very Easy	4	2	6
2. MTT Tutorial			
Not Enough Information to Answer	2	0	2
Very Hard	0	0	0
Hard	1	0	1
Not Easy or Hard	6	1	7
Easy	9	8	17
Very Easy	4	3	7

Utility Data.
The next section of the survey addressed the usefulness of the training provided by MTT and its various components. Students were first asked to rate how useful, overall, they found the MTT training. Their reactions to this four-choice item are shown in Table 37. It is evident from the data, there was a somewhat mixed opinion, as just under one-fourth of the sample reported the training as "not useful" whereas over 60% of respondents viewed MTT training as "moderately useful" or "very useful." Pilots were slightly more positive toward MTT's training utility compared to Sensors.

Table 37

Frequency of Student Reaction to Overall Usefulness of MTT Training

Survey Response Item	**Pilot**	**Sensor**	**Total**
Not Enough Information to Answer	1	3	4
Not Useful	6	2	8
Moderately Useful	12	5	17
Very Useful	3	1	4

Students were then asked to rate the usefulness of the green scorecard that appeared at the end of each Training session trial. The scorecard showed the students how many total points they earned, and also gave a breakdown of points earned by task. The responses to this item are shown in Table 38. As can be seen, the scorecard was viewed quite favorably, as 28 out of 34 respondents (82%) rated it as either "moderately useful" or "very useful."

Table 38

Frequency of Student Reaction to the Utility of the Green Scorecard

Survey Response Item	**Pilot**	**Sensor**	**Total**
Not Enough Information to Answer	0	0	0
Not Useful	4	1	5
Moderately Useful	15	9	24
Very Useful	3	1	4

Besides being a source of informative feedback, the other possible benefit of the scorecard was to serve as a motivating factor to improve performance. Respondents were thus asked whether they found the scorecard to be motivating. Their reactions are shown in Table 39. Slightly more than one-third of the sample (38%) thought the scorecard was "very motivating" whereas only two respondents (both Sensors) considered it "unmotivating." It was a greater motivating factor for Pilots than for Sensors.

Table 39

Frequency of Student Reaction to the Motivating Value of the Green Scorecard

Survey Response Item	Pilot	Sensor	Total
Very Motivating	10	3	13
Neutral	12	7	19
Very Unmotivating	0	2	2

Another aspect of MTT with anticipated training value were the green and red dots appearing within each quadrant during the training scenarios. These were intended to give students immediate and highly specific feedback concerning the correctness of their performance on the individual tasks as the scenario proceeded. The students were specifically asked whether they thought the presence of these dots helped or hurt their performance. Their reactions to this question are shown in Table 40. Over one-half of the subjects (18/34) thought it helped their performance, while slightly less than one-third of the sample (10/34 or 29%) did not notice the dots at all.

Table 40

Frequency of Student Reaction to the Performance Value of the Green and Red Dots

Survey Response Item	Pilot	Sensor	Total
Helped my Performance	11	7	18
Hurt my Performance	4	1	5
Didn't notice the Dots	6	4	10

Another element of MTT that was probed was the training utility of having a single task within each scenario being given a "higher priority" and, as such, worth double the points (both positive and negative) of the other tasks. Student reaction to this item is shown in Table 41. Over half of the sample (18/34 or 53%) thought having a higher priority task within MTT provided "good training." Less than 10% of respondents (3/34 or 9%) felt the higher priority task was a "distraction or not helpful."

Table 41

Frequency of Student Reaction to the Training Value of a Higher Priority Task

Survey Response Item	Pilot	Sensor	Total
Not Enough Information to Answer	1	0	1
Good Training	12	6	18
Neutral	7	5	12
A distraction – not helpful	2	1	3

Another feature of MTT designed to add challenge was shifting the quadrants of the four tasks between scenarios. This was done to "mix up" the students' scan patterns and thereby make the overall attention management task more difficult. The student reaction to the training value of this feature is depicted in Table 42. Slightly more than 40% of respondents thought it made practice "more realistic," whereas a comparable percentage thought it "didn't matter." Half of that percentage (20%) thought it actually "hurt performance and delayed training."

Table 42

Student Reaction to the Training Value of Shifting Task Quadrants between Scenarios

Survey Response Item	Pilot	Sensor	Total
Made the Practice more Realistic	9	5	14
Didn’t Matter	8	5	13
Hurt my Performance – Delayed my Training	5	2	7
Didn’t notice the Shift in Location	0	0	0

Another feature of MTT students were asked about was the clock timer that appeared above the right quadrant in each scenario. Specifically, students were asked if they found it motivating for performance improvement. Their responses to this item are displayed in Table 43. The modal response to this question was “neutral,” with 56% of subjects choosing this response. While only 20% of the sample found the clock timer to be “motivating,” another 12% did not even notice the clock.

Table 43

Student Reaction to the Motivational Value of the Clock Timer

Survey Response Item	Pilot	Sensor	Total
Highly Motivating	0	0	0
Motivating	6	1	7
Neutral	12	7	19
Distracting	1	1	2
Very Distracting	1	0	1
Didn’t Notice the Clock	2	2	4

The last MTT feature probed was the Top Score that was always displayed in the blue bar above the four tasks. This was intended to be motivating in the same way game scores motivate players to continue playing and improve their performance. Responses to this item are displayed in Table 44. Just under 60% of the sample found the top score to be “highly motivating” or “motivating,” with only 1 student reporting the top score to be “distracting.” Another third of the sample rated the top score as “neutral.”

Table 44

Student Reaction to the Motivational Value of the Top Score

Survey Response Item	Pilot	Sensor	Total
Highly Motivating	4	1	5
Motivating	12	3	15
Neutral	5	7	12
Distracting	1	0	1
Very Distracting	0	0	0
Didn’t Notice the Top Score	0	1	1

For the final item in this part of the survey, students were asked whether they would recommend MTT to other aviators. The frequency of responses to this question, shown in Table 45, formed a fairly even distribution which is indicative of a varied opinion by the student respondents. Thus, while only two students said it would be "very unlikely" they would recommend MTT to other aviators, a fairly high percentage, 38% (13/34), were "not sure" of what their recommendation would be. This is indicative of some ambiguity about the true training purpose of MTT and what its targeted skill set was. This issue will be addressed in the concluding section of this analysis.

Table 45

Frequency of Student Response to Recommending MTT to Other Aviators

Survey Response Item	Pilot	Sensor	Total
Not Enough Information to Answer	0	1	1
Very Unlikely	2	0	2
Unlikely	6	1	7
Just not Sure	7	6	13
Likely	4	2	6
Very Likely	3	2	5

Reaction to the Training Preparation of MTT and its Individual Tasks.
The remainder of the survey asked students questions about the structure of training – whether the sequence of Familiarization, Training, and Test scenarios was good preparation – and the training value of the four individual tasks that comprise MTT. Students were first asked whether the Familiarization scenarios (the ones that gave MTT users practice on each task individually) prepared them to receive multi-tasking training. Students were then asked whether they thought the Training scenarios prepared them for the Test scenarios. Their responses to these two questions are shown in Table 46, with the frequency of responding to the questions broken out across columns.

Table 46

Frequency of Student Reaction to Preparation Afforded by Familiarization & Training Scenarios

	Familiarization Scenario			Training Scenario		
Survey Response Item	Pilot	Sensor	Total	Pilot	Sensor	Total
Not enough information to answer	1	1	2	0	1	1
Very well prepared	4	3	7	3	1	4
Well prepared	8	6	14	9	4	13
Neutral	8	2	10	10	5	15
Poorly prepared	0	0	0	0	1	1
Very poorly prepared	0	0	0	0	0	0

For the most part, the results in Table 46 indicate students felt the training structure prepared them fairly well for what was coming next. With regard to Familiarization training, almost two-thirds of the sample (21/34) felt either "very well prepared" or "well prepared" for the subsequent Training scenarios. The numbers were somewhat lower for the Training Scenario's preparation of Test Scenarios, as only 50%

(17/34) selected these two response categories. Nevertheless, only one student (a Sensor) felt "poorly prepared" for the subsequent phase of training/testing. It was clear from the data some students were not sure of how well prepared they were, as 29% (10/34) and 44% (15/34) of students, respectively, selected the "neutral" response category for the Familiarization and Training Scenario preparation questions.

Students were then asked to rate how well they thought the Training scenarios, which were arranged in order of increased difficulty for each phase of training, "matched" their pace of learning. Their frequency of responses to this question is shown in Table 47. These results mirror those in Table 46. Thus, some 50% of students (17/34) thought the training structure was a "great match" or "good match" with their pace of learning. While another 41% (14/34) viewed this match as "neutral," only three students (9%) considered it a "poor match."

Table 47

Frequency of Student Response to How Well Scenario Difficulty Matched Pace of Learning

Survey Response Item	**Pilot**	**Sensor**	**Total**
Not Enough Information to Answer	0	0	0
Great Match	4	0	4
Good Match	10	3	13
Neutral	7	7	14
Poor Match	1	2	3
Very Poor Match	0	0	0

As a final question, students were asked to rate the training value of each of the four individual MTT tasks: memory, addition, audio, and visual search. The responses to these questions are shown in Table 48. To facilitate comparison, the four tasks' frequencies are arrayed across the columns. Note the response choices were not mutually exclusive so respondents could select more than one choice. Consequently, the tallies by task will total more than 34, the number of respondents.

Table 48

Frequency of Student Reaction to Preparation Afforded by Familiarization and Training Scenarios

	Memory Task			**Addition Task**			**Auditory Task**			**Visual Search Task**		
Survey Response Item	**Pilot**	**Sensor**	**Total**	**Pilot**	**Sensor**	**Total**	**Pilot**	**Sensor**	**Total**	**Pilot**	**Sensor**	**Total**
Good training	12	6	18	8	0	8	14	5	19	14	7	21
Not realistic	2	1	3	4	1	5	2	2	4	4	0	4
Frustrating	2	3	5	7	7	14	1	1	2	2	3	5
Challenging	13	4	17	14	6	20	5	0	5	7	2	9
Too easy	0	0	0	0	0	0	5	4	9	1	1	2
Too hard	0	0	0	1	2	3	0	0	0	0	0	0

There was considerable variability in how students reacted to the four tasks. For example, the memory, auditory, and visual search tasks were all viewed as providing "good training" by at least 50% of respondents, with over 60% expressing this opinion about the visual search task. In contrast, the addition task, which was difficult and not necessarily reflective of typical Operator tasks, was considered "good training" by only one-fourth of the sample. On the other hand, no more than five respondents (15%) considered any of the tasks to be "not realistic." There was also a considerable difference in the challenge levels of the tasks, as the memory and addition tasks were considered "challenging" by 50-59% of the sample, compared to only 15-26% for the auditory and visual search tasks. In this regard, no task, save the auditory task, was viewed as "too easy" by more than several students, whereas only the addition task received any responses to the "too hard" option.

Level II (Evidence of Learning for MTT)

Given the study design, introduction of MTT to the CRM curriculum constituted part of Spiral 3, where the EA and ICH were already inserted into the curriculum as Spiral 2 interventions. Spiral 3, then, consisted of receiving the EA, ICH, *and* MTT. Subsequent classes received Spiral 4, consisting of EA, ICH, MTT, and the fourth training intervention, the Gemasim team trainer (GTT). Since the four interventions were experienced in independent sessions, each was subjected to its own Level II analysis and reported separately. In this regard, the Level II analysis results were previously reported for EA and ICH.

Logically, student learning (i.e., Kirkpatrick's Level II) occupies an intervening link between student reaction (Level I – Do they like the instruction?) and transfer of training (Level III – Do they transfer the knowledge and skills they have learned into the operational environment?). In this regard, it would seem reasonable to expect students who fail to learn the requisite knowledge-skills-attitudes (KSAs) during the training intervention will not be able to transfer those unlearned KSAs into the operational environment. Put another way, empirical evidence for student learning (Level II) is a necessary but not sufficient condition for transfer of KSAs to the operational environment (Level III). The purpose of this analysis is to make the case that a significant amount of student learning *did* occur over the course of Spiral 3 training involving MTT.

Logic of the Analysis.
As in previous Level II analysis reports, a direct measure of student learning involved comparison of scores on pretests and post-tests, where the difference or "gain" score yielded a clean measure of how much the student learned. While such a framework was in place for the EA intervention, it was not for ICH and hence several indirect measures of learning were used. The ones used for ICH involved an analysis of absolute performance, in which comparisons to chance performance, established academics passing rates, and between crew positions were all used to "triangulate" how much students learned as a result of receiving ICH. Student attrition, that is, the number of students who registered on the Birds of Prey website but, for whatever reason, did not complete ICH training, was also used.

Just as with ICH, the framework was not available for MTT, in that there were not pretest/post-test comparisons available for a direct assessment. On the other hand, students did take multiple scenarios so that an improvement in scores over scenarios was used as an index of learning. This fact provided two options for examining learning, each with merit. On the one hand, for students who elected to take the same scenario multiple times, their score on the second run-through was compared to the first time they took the scenario. An increase in performance on the second occasion was a good indication they were learning. This method gave a relatively clean measure of learning, although the number of students who opted to repeat scenarios for the purpose of performance improvement was disappointingly low.

The other option involved comparing student scores across the Training and Test scenarios as an index of learning. A problem here was the scenarios were intentionally designed to increase in difficulty, so it could be the case that a student's scores could actually go down due to the difficulty increase, yet they were still learning. However, if their scores remained at some stable level, even with the difficulty increase, then that provided evidence of learning. Unfortunately, there was the added complication that a student's score could decline from one scenario to the next just due to chance and other non-learning factors owing to the scenarios' fast-paced timing. That is, even a moment's distraction could have significant consequences on a student's performance on that scenario. This was, of course, one of the objectives of the intervention – to make students aware that losing focus for even a brief moment had adverse consequences in the cockpit. Nevertheless, these momentary lapses could have dramatic impacts on one's scenario score.

To deal with the impact of momentary lapses, a scoring method was developed. Specifically, the student's *maximum* score was computed across a block of scenarios and expressed as a percentage of the total possible points for the scenario. Since it was feasible for students to actually receive a negative score (i.e., errors and misses outnumbered successes), computing a percentage correctly provided a fairly stiff assessment of learning. Thus, if students obtained a respectable percentage score, such as 50% or higher, then it could be safely concluded they learned some of the requisite attention management skills required to perform the multi-tasking scenarios. Indeed, failure to learn adequately and perform reasonably well resulted in scores near zero and perhaps even negative. Hence, if students obtained a maximum score of 50%, for example, over a block of scenario trials, then that was considered reasonable evidence for learning.

In addition to looking at total scores, it was instructive to examine performance on the four component tasks – auditory, addition, visual search, and memory – that comprised MTT. On each scenario trial, students received two points for a correct response, lost one point for a "miss," and lost two points for an incorrect response (error). In addition, one task was singled out (randomly) to be the high priority task for that trial. Performance on that task was doubled, such that correct responses were worth four, while misses and errors cost students two points and four points, respectively. Since the objective was to maximize one's total score (which was continually displayed and updated in the middle of the screen), it was in the students' best interest to give greater attention to the high priority task.

Finally, as with ICH, the number of eligible students who completed the MTT scenarios was computed as a sign they learned to use the system sufficiently well to complete the training. However, a complication here was posed by system lag problems some students experienced. These issues were described in the Level I analysis of student reaction data to the MTT. Nevertheless, the attrition index was computed to derive some indication of student-perceived learning, though attrition was likely higher due to the technical difficulties of administering MTT over the Web in a poor connectivity environment.

Data Collection.

Each time a student completed a scenario, their scores were automatically sent via e-mail to the Birds of Prey website server. Once on the server, an Excel database housed the scores and permitted an analysis based on class number, crew position, and scenario. The score was also retained on each of the four component tasks to determine which tasks were most difficult for students and posed the most challenge for mastery. Student scores were kept separate based on user name, which allowed identification of which students completed a given scenario more than once.

Three classes provided data for this analysis. The first, Class 09-14, was a control class that received MTT as a beta-test. Since the beta-test was relatively successful technically with a decent participation rate, the data were included in this analysis. The other two classes, 09-15 and 10-01, were true Spiral 3 classes that, in addition to MTT, received EA and ICH. A total of 39 students completed the MTT scenarios, consisting of 21 Pilots and 18 Sensors. That constituted a participation rate of 68% (57 students total in the three classes), which was sufficiently high to warrant the analyses described below.

Student Performance across Scenarios as an Index of Learning.
Depending on the scenario, the total points the student could obtain ranged from 58-64 in the training phase and 88-94 in the test phase. To simplify presentation, each score was divided by the total possible for that scenario to generate a percentage score. In the case of negative points, which happened infrequently, the original score was retained since a percentage score was not very meaningful. If a student received all points in the scenario their percentage was expressed as 1.00.

Table 49 presents the maximum score (as a percentage) each student received in the training phase, the first four test scenarios, and the last four test scenarios. Note the Familiarization scenario scores were omitted from analysis since they were intentionally designed to be easy and student scores did not reveal any evidence of learning there. To preserve anonymity, the student's user name was replaced with a sequence number.

Table 49

Student's Maximum MTT Score (& of Possible Points) by Scenario Phase

Class/Crew Position/ Student No.	Max Score in Training (5-8)	Max Score in Test (9-12)	Max Score in Test (13-16)
	Pilots, 09-14		
Student 1	1.00	0.90	0.69
Student 2	0.83	0.83	0.72
Student 3	0.72	0.66	0.65
Student 4	0.77	0.84	0.74
Average	**0.83**	**0.81**	**0.70**
	Sensors, 09-14		
Student 1	0.80	0.69	0.82
Student 2	0.82	0.83	0.68
Student 3	0.58	0.61	0.63
Student 4	1.00	0.91	0.75
Average	**0.80**	**0.76**	**0.72**
	Pilots, 09-15		
Student 1	0.77	0.80	0.66
Student 2	0.79	0.84	0.77
Student 3	0.95	0.97	0.82
Student 4	0.59	0.33	-0.09
Student 5	0.86	0.85	0.87
Student 6	0.84	0.86	0.80
Student 7	0.89	0.87	0.74
Student 8	0.93	0.48	0.72
Student 9	0.61	0.20	0.23
Student 10	0.94	0.71	0.68
Student 11	0.94	0.87	0.90
Average	**0.83**	**0.71**	**0.65**

Class/Crew Position/ Student No.	Max Score in Training (5-8)	Max Score in Test (9-12)	Max Score in Test (13-16)
Sensors, 09-15			
Student 1	0.80	0.73	0.66
Student 2	0.91	0.91	0.81
Student 3	0.82	0.88	0.72
Student 4	0.83	0.78	0.80
Student 5	0.95	0.98	0.77
Average	**0.86**	**0.86**	**0.75**
Pilots, 10-01			
Student 1	0.57	0.70	0.68
Student 2	1.00	0.90	0.91
Student 3	0.91	0.83	0.90
Student 4	0.79	0.64	0.53
Student 5	0.61	0.68	0.65
Student 6	0.90	0.80	0.72
Average	**0.80**	**0.76**	**0.73**
Sensors, 10-01			
Student 1	0.67	0.66	0.72
Student 2	0.86	0.89	0.83
Student 3	0.74	0.68	0.70
Student 4	0.86	0.74	0.66
Student 5	0.83	0.83	0.77
Student 6	1.00	0.71	0.71
Student 7	0.75	0.81	0.72
Student 8	0.75	0.53	0.36
Student 9	0.63	0.51	0.72
Average	**0.79**	**0.71**	**0.69**

The bottom row in each segment of the table provides the average maximum percentage score for the students in that class and crew position. It is apparent these averages were quite high, ranging from a low of .65 to a high of .86. Comparing the 3 scenario phases, it was clear the last 4 test scenarios were certainly difficult as indicated by somewhat lower average maximum scores. Nevertheless, this was fairly robust evidence the students performed quite well in the scenarios, suggesting considerable learning occurred.

In assessing degree of learning, it was important to look at how the individual students performed, not just the class average. In this regard, virtually every student was fairly successful, with most obtaining a maximum percentage score over .50 in each block of 4 scenarios. There were a few exceptions, however, such as Student 4, a Pilot in Class 09-15, who only reached .33 in the first 4 test scenarios and actually had all negative scores (hence a negative maximum of -0.09) in the last 4 test scenarios. Another student who experienced difficulty was Student 9 in the same class, who only reached a maximum of .20-.23 in the test phase. Also, Student 8, a Sensor in Class 10-01, achieved only a maximum score of .36 in the last 4 test scenarios. However, these individuals were the exceptions, as the other students obtained maximum scores near .50 or, in many instances, much higher, in all 3 blocks of scenarios. Thus, over

90% of the student sample scored well enough in the scenarios to support the view that learning in MTT did indeed occur.

Student Performance in Repeat Scenarios as an Index of Learning.

A more direct index of learning was whether student performance improved when they repeated a scenario. This was because it was not necessary to untangle the effects of increasing scenario difficulty since it was the same scenario being performed. Unfortunately, instances where a given student repeated a scenario (presumably to improve performance) were fairly rare. In the three classes of MTT data, scenarios were repeated only 25 times. These repeat scenarios were done by 12 different students. Despite their low frequency, first-time and second-time scores were still compared to gauge whether learning occurred.

Table 50 presents the scores for each of the 25 scenario repeat instances. The same student numbering scheme was used to identify the participant, where the class number, crew position, and scenario were also indicated. Scores were again reported as a percentage of total possible points for that scenario. The last column displays the difference between the first and second scenario attempts, where a negative score indicated performance actually went down on the second attempt. This only happened 4 times. This effect was statistically significant ($p < 005$), as compared to the null hypothesis that positive and negative scores were equally likely. The average gain score was .16, which corresponds to a 16% improvement between the first and second scenario attempts. This was also fairly strong evidence learning occurred for those students who elected to repeat a scenario.

Table 50

Student's Maximum MTT Score (% of Possible Points) by Scenario Phase

Class/Crew Position/ Student No./Scenario	Score on 1st attempt	Score on 2nd attempt	Difference Score
Student 2, Pilot, 09-14, Scenario 5	0.75	-0.31	-1.06
Student 1, Pilot, 09-14, Scenario 6	0.84	1.00	0.16
Student 1, Pilot, 09-14, Scenario 11	0.72	0.68	-0.04
Student 10, Pilot, 09-15, Scenario 5	0.94	0.23	-0.70
Student 2, Sensor, 09-15, Scenario 5	0.91	1.00	0.09
Student 4, Sensor, 09-15, Scenario 5	0.55	0.89	0.34
Student 1, Pilot, 09-15, Scenario 7	0.77	0.93	0.17
Student 1, Pilot, 09-15, Scenario 8	0.52	0.79	0.28
Student 8, Pilot, 09-15, Scenario 8	0.59	0.86	0.28
Student 1, Sensor, 09-15, Scenario 8	0.47	0.53	0.07
Student 8, Pilot, 09-15, Scenario 13	0.64	0.63	-0.01
Student 5, Sensor, 09-15, Scenario 13	0.33	0.58	0.24
Student 5, Sensor, 09-15, Scenario 14	0.41	0.59	0.18
Student 10, Pilot, 09-15, Scenario 15	0.67	0.86	0.19
Student 6, Pilot, 10-01, Scenario 5	0.38	0.72	0.34
Student 4, Sensor, 10-01, Scenario 5	0.67	0.69	0.02
Student 6, Pilot, 10-01, Scenario 7	0.60	0.67	0.07
Student 4, Sensor, 10-01, Scenario 7	0.70	0.80	0.10
Student 6, Pilot, 10-01, Scenario 9	0.65	0.80	0.15
Student 4, Sensor, 10-01, Scenario 9	0.58	0.72	0.14
Student 6, Pilot, 10-01, Scenario 10	0.62	0.73	0.11
Student 6, Pilot, 10-01, Scenario 12	0.66	0.85	0.19
Student 6, Pilot, 10-01, Scenario 13	0.70	0.87	0.17
Student 4, Sensor, 10-01, Scenario 14	0.15	0.42	0.27
Student 6, Sensor, 10-01, Scenario 14	0.50	0.77	0.27

Individual Task Performance.

To determine the contributions of the individual tasks to overall performance, the average number of misses and errors was computed for each task within the three classes and two crew positions. These were averaged across students and across scenarios within these subgroups. The results are displayed in Table 51.

Table 51

Average Number of Hits and Misses for Each MTT Task, by Class and Crew Position

	Auditory		Calculation		Memory		Visual Search		Sum
Class/Crew Position	**Misses**	**Errors**	**Misses**	**Errors**	**Misses**	**Errors**	**Misses**	**Errors**	
09-14, Pilots	0.58	0.42	0.42	2.31	0.85	0.44	1.52	0.85	7.39
09-14, Sensors	0.85	0.35	0.48	2.19	1.02	0.48	1.00	0.79	7.16
09-15, Pilots	1.32	0.62	0.28	2.19	1.23	0.45	1.04	0.74	7.87
09-15, Sensors	0.70	0.28	0.25	2.56	0.75	0.39	0.57	0.78	6.28
10-01, Pilots	1.01	0.32	1.58	1.58	1.17	0.07	1.58	0.63	7.94
10-01, Sensors	1.18	0.28	0.61	2.43	1.17	0.14	0.79	0.95	7.55
Overall	**0.94**	**0.38**	**0.60**	**2.21**	**1.03**	**0.33**	**1.08**	**0.79**	

Looking first at the bottom row, the overall averages presented a fairly clear picture of which tasks were more difficult for students to master. Specifically, the calculation (3-digit and 4-digit addition) task was by far the most difficult, as its combined miss and error rate of 2.81 was substantially larger than the other tasks. By this combined measure, the auditory task (tone pitch discrimination) and the memory task (serial position memory of letters and numbers) were almost identical in difficulty, as their combined miss and error rates were 1.32-1.36. Occupying a middle ground was the visual search, with a combined rate of 1.88. Note also the patterns of the difficulties varied with the task, as the memory task was more often "missed" (subjects did not have time to either view the original memory set or see the probe digit) than had an incorrect response. This was also the case for the auditory tone discrimination, whereas the opposite pattern was evident for the calculation task.

The rest of the table, indicates there was considerable variation across classes and crew position. Indeed, certain subgroups seemed to exhibit superior performance compared to their cohorts. To illustrate this, the sum of the averages across the four tasks' misses and errors is displayed in the right-most column. This shows the performance for Sensors in Class 09-15 was unusually good, whereas Pilots in Class 10-01 exhibited the opposite pattern. These pronounced cohort effects were shown in other analyses, where there were undoubtedly mediating factors at work. For example, the Pilots in Class 10-01 were older on average than most other classes, which was likely a disadvantage in a fast-paced delivery system like MTT.

Student Attrition as an Index of Learning.

The other approach to measuring learning was to make an inference from student attrition rates, where the number of students who initially registered on the Birds of Prey website was compared to the corresponding number who actually completed the MTT scenarios. These rates were determined for each class and crew position, and the tallies are presented in Table 52. The bottom row shows the completion rates for the two crew positions across all three classes.

Table 52

Percentage of Students Completing MTT Training by Class and Crew Position

	Pilots			Sensors		
Class	**# Registered**	**# completed**	**% completed**	**# Registered**	**# completed**	**% completed**
09-14	9	4	44.4	8	4	50.0
09-15	11	11	100	9	5	55.6
10-01	9	6	66.7	11	9	81.8
Overall	29	21	72.4	28	18	64.2

The completion rates were quite variable; they were much lower for the first class, 09-14, than the other two. This fact was not surprising, however, since this class encountered a larger number of technical issues involving system lag that were frustrating and which, undoubtedly, contributed to the attrition. This issue was discussed at length in the Level I analysis on MTT. An encouraging sign in this regard was the 100% completion rate of Pilots in Class 09-15 and the over 80% completion of Sensors in Class 10-01. This suggests the potential for learning and sustained participation was evident.

Level III (Transfer of Training for Spiral 3)

Analysis of Targeted Skills from HF form.

This initial analysis provides a summary of the first round of Level III data analysis for the Spiral 3 comparison. The data set reported here is from the instructor ratings on the Human Factors data sheet implemented at Creech AFB. Recall these ratings were obtained from two carefully chosen training sessions: the third and final Combined Operations session (CO-3) conducted with an actual aircraft, and the final emergency procedures check (EPE) conducted in the simulator with Standardization/Evaluation (Stan/Eval) raters.

The Spiral 3 comparisons encompassed four classes: 09-15, 09-16, 10-01, and 10-02. The actual Spiral 3 treatment classes were the odd-numbered ones (i.e., 09-15 and 10-01). The other two classes were the non-treatment controls. The Spiral 3 treatment consisted of the EA and ICH (which were the Spiral 2 treatments) *plus* the Multi-task Trainer (MTT), a self-paced computer-based game. Level I and Level II analyses for this intervention were described above. Note both of these analyses included a third class, 09-14, which was a successful beta-test implementation of MTT. However, since the class did not receive the other interventions (EA and ICH), it was not appropriate for inclusion in a Level III analysis. Consequently, only Classes 09-15 and 10-01 were analyzed as Spiral 3 treatment classes.

As with Spiral 2, the comparisons made in Spiral 3 were fairly direct, so it was appropriate to simply combine the results of Classes 09-15 and 10-01, and 09-16 and 10-02, and report the mean ratings of each. For ease of labeling, the two control classes were referred to as Control 3. With participation rates of 60-80% across the classes, this gave a reasonable sample size (i.e., N = 14 to 19) for computing means and performing confidence interval testing.

Before presenting the results of the mean difference analysis, first how the ratings were distributed across the 5-point (0-4) Human Factors scale is reported. Recall the same 5-point scale was used that Creech AFB instructors used on the training gradesheets, where “0” was essentially unsafe flying and “4” represented exemplary performance. The distribution of the ratings across classes was examined to see how often the instructors used the two ends of the scale. Of the 798 ratings generated across the 4 classes (133 student sheets x 6 rating dimensions), there were 91 “4” ratings, or 11.4% of the total. This was just slightly higher (10.4%) than what was observed for Spiral 2. In addition, there were 6 “0”s reported as well as 21 “1”s, which comprise 3.3% of the total. Thus, the instructors had just slightly less than 15% of their ratings on the two ends of the scale, which indicated instructors were quite willing to provide extreme ratings.

In addition, the 2 conditions, Spiral 3 and Control 3, were compared to see if there were any differences in how frequently the extreme ends of the scale were used. Some notable group differences were found on both ends of the scales. First, Control 3 students received a higher percentage of “4”s, 14.9%, compared to the Spiral 3 students, 7.1%. While this was in the opposite direction preferred, ALL six ratings of “0” were also received by Control 3 students. In this case, 3 students accounted for the unsatisfactory ratings, all from the Control classes. On the other hand, the frequency of “1” ratings was virtually the same for the 2 conditions. Once again, there was evidence of very strong cohort effects, in which certain classes appeared to be unusually stronger (in this case, Class 10-02) than others, whereas other classes had a disproportionate number of very low-performing students. This large between-class variability added considerably to the within-group error variance, making it difficult to find statistically significant between-group differences.

Table 53 presents the means, variances, and sample sizes (N) for the Spiral 3 and Control 3 classes. The table is divided into four segments, corresponding to the CO-3 sessions for Sensors, EPE sessions for Sensors, CO-3 sessions for Pilots, and EPE sessions for Pilots. As with Spiral 2, the data were broken out by crew position since their tasks were so different and because they were rated by different instructors. Within each segment, the data appear in six columns, corresponding to the six dimensions rated on the HF data collection sheet: attention management, task prioritization, course of action (COA) selection, crew coordination, degree of instructor intervention, and CRM performance.

Table 53

Mean Ratings for the Six Classes Comprising the “Spiral 3 Comparisons” (0-4 Scale)

Sensor CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 3 (Mean)	2.78	2.78	2.61	2.83	**3.33**	2.89
Control 3 (variance)	0.54	0.54	0.72	0.50	0.59	0.69
Control 3 (N)	18	18	18	18	18	18
Spiral 3 (Mean)	2.50	2.44	2.38	2.88	2.69	2.69
Spiral 3 (variance)	0.53	0.40	0.52	0.92	0.63	0.50
Spiral 3 (N)	16	16	16	16	16	16

Sensor EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 3 (Mean)	2.29	2.35	2.24	2.24	2.41	2.44
Control 3 (variance)	0.72	0.74	0.69	0.82	0.63	0.68
Control 3 (N)	17	17	17	17	17	17
Spiral 3 (Mean)	2.64	2.71	**2.71**	**2.79**	**2.92**	2.64
Spiral 3 (variance)	0.40	0.37	0.22	0.18	0.41	0.40
Spiral 3 (N)	14	14	14	14	13	14

Pilot CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 3 (Mean)	2.68	2.74	2.68	2.84	2.58	2.63
Control 3 (variance)	0.34	0.45	0.45	0.47	0.37	0.58
Control 3 (N)	19	19	19	19	19	19
Spiral 3 (Mean)	2.50	2.44	2.38	2.88	2.69	2.69
Spiral 3 (variance)	0.53	0.40	0.52	0.92	0.63	0.50
Spiral 3 (N)	16	16	16	16	16	16

Pilot EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 3 (Mean)	2.28	2.44	2.39	2.47	2.88	2.52
Control 3 (variance)	0.92	0.73	0.96	0.66	0.99	0.78
Control 3 (N)	18	18	18	18	17	18
Spiral 3 (Mean)	2.44	2.38	2.38	2.38	2.75	2.38
Spiral 3 (variance)	0.53	0.65	0.52	0.52	0.73	0.38
Spiral 3 (N)	16	16	16	16	16	16

To facilitate visual comparison, the variance and N values are displayed in gray within their respective cells so the means are the most prominent measure within each segment. For statistical comparisons, just as in Spirals 1 and 2, a confidence interval around the Spiral mean was computed based on the respective variances of the spiral condition and its control counterpart. In essence, a confidence interval was constructed that would cover a 95% probability, by chance, of having another mean within its range. If the control condition mean was outside this confidence interval, then they were considered statistically different. As discussed in previous analyses, this test turned out to be quite conservative since adjustments for the inter-correlations between the six rating dimensions were not made, which could have been done to reduce overall error variance. Such adjustment is virtually impossible for this data set, however, because the sample sizes were markedly unequal and, more problematic, because there were not both (i.e., EPE and CO-3) sessions' worth of data for many of the subjects. Consequently, this was a conservative testing method which likely understated the group differences present in the data.

The pattern in this data set was somewhat different than the other spirals. First, the Spiral-Control cell differences were much larger for the Sensors than for the Pilots. In fact, examination of the lower half of the table (Pilot data), showed fairly small differences that fluctuated between having the Spiral mean higher and the Control mean higher. Given the large within-cell variances, there were clearly no notable differences here.

For the Sensors, on the other hand, there were some sizeable group differences. In keeping with the convention used in other analyses, the mean differences are color-coded; a positive treatment effect is shaded in green whereas differences in the opposite direction predicted are shaded in red. In the CO-3 session, the Control subjects actually had a significantly higher mean rating for the Instructor Intervention dimension. This was based on the calculation of two times the square root of the estimated variances about the mean (Harris, 1994), which, for this data set, tended to range between .44-.60. On the other hand, three of the Spiral-Control mean differences reached statistical significance in the EPE session, all of which are coded green to reflect a higher mean for the Spiral condition. The dimensions for which this was true included COA selection, Crew Coordination, and Instructor Intervention. Interestingly, ALL six of the rating dimensions in this session showed a superiority of the Spiral over the Control. This was the only segment of the table for which all six rating dimensions went in the same direction. Using binomial testing procedures, a non-parametric analysis, this would have occurred by chance (assuming each direction was equally likely, [i.e., drawing from a binomial population with $p = .5$]), only 2% of the time (i.e., $p < .02$) (Miller & Freund, 1965). The consistency of the directional differences, and their magnitude, was indicative of a slight preference in favor of the treatment (Spiral 3) classes, at least for the EPE session. A similar superiority for the Spirals 1 and 2 Human Factors ratings was found, though it resided primarily in the CO-3 sessions.

In sum, the results of the parametric and non-parametric testing were, once again, mostly consistent with a slight positive advantage of the Spiral 3 classes (i.e., the ones receiving EA, ICH, and MTT) that appeared to reside with Sensors in the EPE session. Other differences in the data set, some favoring the Control classes, were of moderate magnitude in some cases, but were not statistically significant due to unusually high within-group variation. As noted previously, having multiple instructors provide the ratings, a necessity in an operational training squadron, added considerably to the variation in the ratings obtained, and posed a stiff challenge to finding significance differences. Nonetheless, such differences were found that, by and large, favored the Spiral condition over the Control. Consequently, when coupled with the fairly positive student reaction data (Level I) to the training, and the consistent evidence of learning (Level II), a fairly solid set of data continued to accrue to support a positive impact of the CRM training interventions.

Analysis of Negative Behaviors from HF form.

This second analysis examined the percentage of students within a condition who received a minus on the six to seven behaviors associated with each of the four HF skills. Examples of these behaviors included *effective cross-check, cross-check doesn't stagnate,* and *switches attention* under the Avoids Channelized Attention skill. As reported in the corresponding Spiral 2 analysis, note not every instructor used the minus designation, as some simply just scored behaviors as zero (i.e., they left it blank) or a +. Other instructors scored ALL behaviors a +, which was interpreted to mean it was simply observed. On the other hand, the assessment of the HF forms was that the instructors who took the scoring process most seriously, scored some behaviors negative, others positive, and the rest neutral. Consequently, the performance was best revealed with this measure by tallying the number of negative behaviors within a class and then converting that to a percentage so all classes were on a common scale.

The analytic strategy is to start with the most aggregated look and then "drill down" into the data, by considering more subgroups in subsequent analyses. For the first look, data were aggregated across control and treatment (spiral) conditions, sessions (CO-3 and EPE), and crew positions to get a sense of which behaviors were stronger or weaker than the others. This aggregated tally is shown in Table 54.

Table 54

Percentage of Negative Behaviors across Conditions, Sessions, and Crew Position

Skill/Behavior	Frequency	Percentage
Avoids Channelized Attention		
effective cross-check	6	0.05
cross-check doesn't stagnate	4	0.03
switches attention	6	0.05
adjusts to different cockpits	4	0.03
not distracted by radios	11	0.08
able to shift attn w/o cues	3	0.02
Task Prioritization		
knows high priority task	8	0.06
handle interruptions	5	0.04
returns to interrupted task	5	0.04
can suspend lower priority task	9	0.07
do tasks concurrently	6	0.05
aviate-navigate-communicate	5	0.04
Select COA		
considers all options	5	0.04
facts vs assumptions	5	0.04
avoids hasty decisions	6	0.05
doesn't take too long	12	0.09
ID potential risks	2	0.02
follow-on decisions	4	0.03

Skill/Behavior	Frequency	Percentage
Crew Coordination		
divide tasks	3	0.02
perform team tasks	8	0.06
anticipate info needs	5	0.04
provides timely data	7	0.05
cross-checks others	7	0.05
maintain SMM	3	0.02
convey SMM	1	0.01
N	**132**	

The middle column of the table indicates the frequency or number of students in both control and spiral conditions who received a negatively scored behavior, considering both the CO-3 and EPE sessions together. That frequency was converted to a percentage by dividing each by 132, which was the total number of students in the 4 classes for which there were EPE and CO-3 data. Though no statistics were performed, a fairly even distribution of percentages across the HF skills enabled identification of the strongest and weakest behaviors within the Spiral 3 comparison data set. In particular, the lowest percentages of negative behaviors (i.e., those for which the aggregate percentage was below .02 [2%]) are left-justified. These were the "strongest" behaviors, and they corresponded to *able to shift attention without cues* and *conveys mental models*. The right-justified percentages indicate the behaviors that received the highest percentage of negative scores, in excess of .9%. There were 2 such behaviors: *not distracted by radios* and *doesn't take too long to select a COA*. Interestingly, these behaviors were also identified as the weakest behaviors in the Spiral 2 analysis.

The second pass through the behavior data peeled away two layers of aggregation, by looking at the percentage of negative behaviors received by students in the two Spiral 3 classes and two Control 3 classes, where separate tallies were provided for the CO-3 and EPE sessions. This breakout is depicted in Table 55.

Table 55

Percentage of Negative Behaviors by Condition and Session

Skill/Behavior	Control 3 – CO3		Control 3 - EPE		Spiral 3 – CO3		Spiral 3 - EPE	
	Freq	**Pct**	**Freq**	**Pct**	**Freq**	**Pct**	**Freq**	**Pct**
Avoids Channelized Attention								
effective cross-check	1	0.027	3	0.083	1	0.034	1	0.033
cross-check doesn't stagnate	1	0.027	1	0.028	1	0.034	1	0.033
switches attention	2	0.054		0.000	2	0.069	2	0.067
adjusts to different cockpits	2	0.054		0.000	2	0.069		0.000
not distracted by radios	5	0.135		0.000	6	0.207		0.000
able to shift attn w/o cues		0.000	1	0.028		0.000	2	0.067

	Control 3 – CO3		Control 3 - EPE		Spiral 3 – CO3		Spiral 3 - EPE	
Skill/Behavior	**Freq**	**Pct**	**Freq**	**Pct**	**Freq**	**Pct**	**Freq**	**Pct**
Task Prioritization								
knows high priority task	1	0.027	3	0.083	2	0.069	2	0.067
handle interruptions	3	0.081		0.000	1	0.034	1	0.033
returns to interrupted task	2	0.054		0.000	2	0.069	1	0.033
can suspend lower priority task	2	0.054	3	0.083	3	0.103	1	0.033
do tasks concurrently	2	0.054	2	0.056	1	0.034	1	0.033
aviate-navigate-communicate	2	0.054	2	0.056		0.000	1	0.033
Select COA								
considers all options	1	0.027	1	0.028	2	0.069	1	0.033
facts vs assumptions	1	0.027		0.000	1	0.034	2	0.067
avoids hasty decisions	1	0.027	2	0.056	2	0.069	2	0.067
doesn't take too long	3	0.081	3	0.083	3	0.103	3	0.100
ID potential risks		0.000		0.000		0.000	2	0.067
follow-on decisions	1	0.027	1	0.028	1	0.034	1	0.033
Crew Coordination								
divide tasks	2	0.041	0	0.000	0	0.036	0	0.000
perform team tasks	0	0.000	1	0.028	1	0.034	1	0.033
anticipate info needs	3	0.081	4	0.111	1	0.034	0	**0.000**
provides timely data	2	0.054	0	0.000	1	0.034	2	0.067
cross-checks others	1	0.027	3	0.083	2	0.069	1	0.033
maintain SMM	2	0.054	1	0.028	1	0.034	3	0.100
convey SMM	2	0.054	0	0.000	1	0.034	0	0.000
N	37		36		29		30	

As with Spiral 2, there were 2 statistical tests performed on these data. The first was a simple sign test comparing the number of behaviors (out of 25) that were higher for one condition over the other. This was done for the CO-3 and EPE sessions separately. This comparison is a non-parametric test since it does not make any assumptions about the underlying distribution of data. It is a reasonable test to employ when it seems there are consistent differences in favor of one group over another, regardless of the magnitude of those differences.

Looking first at the CO-3 session, 15 of the Control 3 cells had lower percentages than Spiral 3; 8 were higher; and 2 were ties (the ties were discarded). The probability this breakout arose by chance was computed by comparison to a binomial distribution in which the hypothesized probability was .50 (Miller & Freund, 1965). The resulting p-value was .06, which was near but did not reach an alpha = .05 level of significance. Note the total number of negative behavior frequencies in this data set was quite low, so the probabilities themselves were not stable. However, it is clear there was no evidence of a positive impact of the Spiral 3 treatment on this measure, and if anything, the effect was in the opposite direction to that expected.

Regarding the EPE session, there were not many negative behaviors scored, so these frequencies were also quite low and the associated probabilities unstable. In terms of a binomial sign test, the prevalence of 0’s produced more ties. Disregarding these, Control 3 classes had 14 behaviors with lower negative percentages compared to Spiral 3 and 7 behaviors where that directionality was reversed. If a binomial probability distribution (with hypothesized $p = .50$) is referenced, the difference again misses the alpha = .05 significance level with $p = .06$. Thus, in both comparisons, a marginal (though not significant) statistical advantage for the Control 3 students over their Spiral 3 condition counterparts was seen.

The second test compared the Spiral vs. Control percentages on each behavior under the two sessions (CO-3 and EPE). This entailed conducting a substantial number of tests, so a fairly stringent alpha-level was necessary to avoid inflation of Type I error rate (Harris, 1994). In all cases, the percentages were compared using a z-statistic, corresponding to the equation: $z = (p_1 – p_2)/SQRT[(1-\underline{p})*(1/n_1)+(1/n_2)]$. In this equation, p_1 and p_2 were the percentages of the two comparison cells, $\underline{p}$ was the average of the two percentages, and n_1 and n_2 were the number of scores contributing to the percentages in each case. Applying this statistic to all the cell percentages in Table 55, only one case was found that met or exceeded significance. This is denoted by the green highlighting, and corresponds to *anticipate info needs* (z=-1.961) in the EPE session. In this instance, the Spiral 3 percentage was lower, indicating better performance.

is a combined graphical representation of Table 55 in which the percent of negative comments of each targets’ skill were aggregated. The mean of each targeted skill was then calculated and graphed. A Wilcoxon sign test across these combined data revealed this Spiral 3 was not statistically significant in the reduction of negative behaviors. This might be due in part to the fact that the control group was unusually strong.

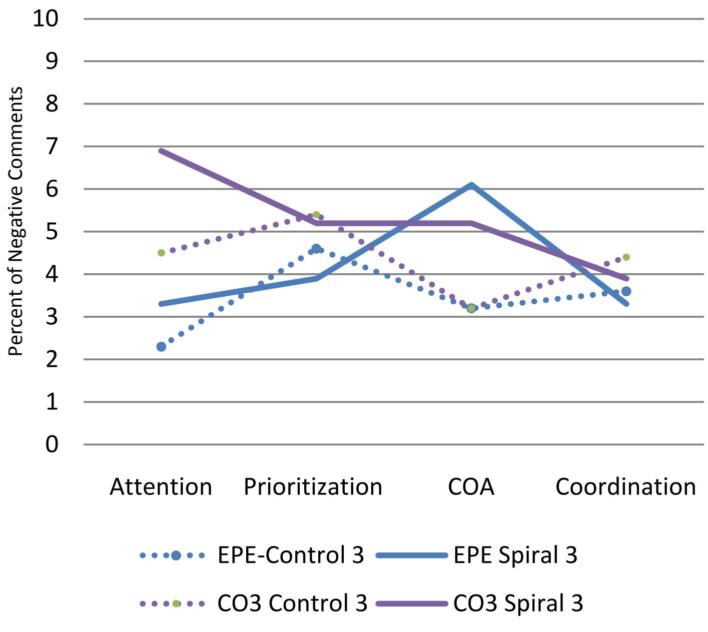


Figure 5: Percent of Negative Comments for Spiral 3

For the final analysis, the percentage of negative behaviors received by the two crew positions was compared. This breakdown is given in Table 56, where the data from the CO-3 and EPE sessions were combined to increase the cell frequencies. In addition, the percentages are presented by Human Factor skill, not the specific behaviors themselves, to have sufficient frequencies driving the percentages.

Table 56

Percentage of Negative Behaviors in CO-3/EPE Sessions by Crew Position and Condition

	Control 3		**Spiral 3**	
Human Factors Skill	**Pilot**	**Sensor**	**Pilot**	**Sensor**
Channelized Attention	0.059	0.014	0.052	0.049
Task Prioritization	0.081	0.019	0.068	0.019
Select Course of Action	0.041	0.023	0.078	0.031
Crew Coordination	0.054	0.037	0.063	0.012
N	222	216	192	162

As can be seen from the table, the Sensors generally had a lower percentage of negative behaviors relative to their Pilot counterparts. However, the size of the differences varied considerably across the HF skill areas and conditions. The differences that reached statistical significance are indicated by yellow highlighting. Within the Control 3 condition, the Sensors had a significantly lower percentage of negative behaviors with regard to Channelized Attention ($z = 2.511$, $p < .01$) and Task Prioritization ($z = 2.977$, $p < .001$). For the Spiral 3 condition, the Sensors had a lower percentage of negative behaviors for Task Prioritization ($z = 2.251$, $p < .01$) and Crew Coordination ($z = 2.516$, $p < .01$).

In closing, there was a mixed picture regarding the impact of the Spiral 3 treatment on the frequency of negative behaviors assigned by instructors on the HF forms for sessions CO-3 and EPE. On the one hand, the non-parametric evidence provided marginal support for a superiority of the Control 3 classes, as they tended to have a small but consistent advantage across sessions and HF skill areas. On the other hand, the one parametric-based statistical difference was obtained in favor of Spiral 3 – this one located in the *perform team tasks* behavior of Crew Coordination. However, note all of these analyses referred to fairly low frequencies of assigned negative behaviors which gave rise to probabilities that were not very stable. In part, this was due to the relatively low overall frequencies of observations since there were only two classes' worth of data for each condition, where the rate of data sheet return varied between 60-80%. In addition, only some of the instructors actually assigned negative behaviors in their data sheets, which also resulted in low frequencies.

With regard to other aspects of the data, there were several trends that mirrored those obtained in Spiral 2. First, the behavior *distracted by radios* was consistently a problem area as both Sensors and Pilots received negative behaviors on this task with considerable frequency. Second, the Sensors tended to receive significantly fewer negative behaviors on their HF scoresheets. This was evident in both types of sessions and in both treatment and control conditions. This difference, which has occurred in all the spiral analyses, was perhaps a reflection of instructional practice more than anything else.

Analysis of Student Gradesheets.

The training items selected for this final analysis were ones CTI SME's believed would encompass the skills most likely to benefit from CRM training. These items are displayed in the left column of Table 11, where the corresponding CRM/HF skill areas are presented in the right column. Each item was scored on a 0-4 scale, where most grades were either a "2" or a "3." Students also received an overall grade for the session. Students receiving a "1" on several training items were usually required to take an extra ride (X-ride). Unfortunately, 2 training items that should be particularly sensitive to CRM training, *Mission Checks* and *Tactical Communications*, were deleted from the Pilot's CO-3 gradesheet starting with Class 09-05.

Combining the data from two classes yielded a respectable sample size (N ~ 20) for comparing Spiral 3 and Control 3 performance. With training records in electronic form (starting with Class 09-05), the grades were entered for the training items identified in Table 11 into an Excel spreadsheet, creating a separate tab for Control 3 and Spiral 3. The statistics resident within Excel were used to calculate the means, variances, and N necessary to perform the required analyses. To make the statistical comparisons, the same method was used as in the first look report: 1) the within-group variances were pooled from the conditions being compared in order to formulate a t-test (Hays, 1973); and 2) a conservative Bonferroni criterion was used to control Type I error inflation due to making multiple tests (Harris, 1994).

As in the analyses for Spirals 1 and 2, there was considerable "noise" in the data due to having multiple instructors (presumably with different internal criteria) assigning grades across students and classes. This created a larger within-group error variance, making it harder to achieve statistical significance. In addition, there were non-trivial "cohort" effects, in which some classes were simply better or weaker than others because of the makeup of their students (e.g., experience, aviation background, stronger class leader). This, too, made it harder to discern differences due to added noise variance. There is no solution for either problem, since these are a fact of life in doing field research within an operational training squadron. It was for these reasons multiple dependent measures were employed for Level III analysis, so there is "triangulation" on the locus of real effects in the data.

Tables 57-60 present the gradesheet data for the CO-3 sessions (Pilots), S-EP-2 sessions (Pilots), CO-3 sessions (Sensors), and S-EP-2 sessions (Sensors), respectively. For each table, the columns correspond to the various training items whereas the rows provide the means, variances, and sample sizes (N) for the Control 3 and Spiral 3 conditions. To permit the mean comparisons to stand out more, the cells containing the variance and N statistics are portrayed in light-gray font. Using the same convention adopted in the other analyses, any statistically significant differences in favor of the treatment (Spiral 3) are highlighted in green, whereas the opposite finding (Control 3 superior to Spiral 3) is highlighted in red.

Looking first at Table 57, there was only 1 significant difference in the CO-3 session for Pilots, highlighted in green, which was in the predicted direction. Specifically, the Spiral 3 mean for the Flight Discipline training item (2.33) was significantly higher than the Control 3 mean (2.10). As with the other segments of the table, the within-group variances for the CO-3 session were somewhat smaller, allowing differences on the order of .23-.25 to be significant. None of the other mean differences were statistically significant. However, of the 8 mean differences, 7 were higher for the Spiral 3 condition relative to the Control 3 counterpart.

Table 57

Gradesheet Data and Analysis Results for CO-3 Session, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	12. ATC Comm	17. CRM	18. ORM	19. Flight Discipline	21. EPs	Overall Grade
Control 3 (mean)	2.53	2.34	2.35	2.30	2.40	2.20	**2.10**	2.20	2.15
Control 3 (VAR)	0.26	0.22	0.24	0.22	0.25	0.17	0.09	0.17	0.13
Control 3 (N)	20	20	20	20	20	20	20	20	20
Spiral 3 (mean)	2.58	2.17	2.46	2.38	2.54	2.42	**2.33**	2.29	2.25
Spiral 3 (VAR)	0.25	0.14	0.26	0.24	0.26	0.25	0.23	0.22	0.28
Spiral 3 (N)	24	24	24	24	24	24	24	24	24

Table 58 presents the Pilots' data for the other session of interest, S-EP-2. These means were all close in value and, hence, there were no significant differences. Although the Control 3 condition tended to have higher means than their Spiral 3 counterparts, the size of the differences was quite small, so none were even close to reaching significance.

Table 58

Gradesheet Data and Analysis Results for S-EP-2 Session, Pilots

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Airmanship	10. ATC Comm	15. CRM	16. ORM	17. Flight Discipline	19. EPs	Overall Grade
Control 3 (mean)	2.52	2.33	2.19	2.19	2.48	2.19	2.19	2.24	2.23
Control 3 (VAR)	0.26	0.23	0.16	0.16	0.26	0.16	0.16	0.19	0.18
Control 3 (N)	21	21	21	21	21	21	21	21	22
Spiral 3 (mean)	2.41	2.14	2.14	2.18	2.41	2.27	2.14	2.14	2.09
Spiral 3 (VAR)	0.25	0.12	0.12	0.16	0.25	0.21	0.12	0.12	0.09
Spiral 3 (N)	22	22	22	22	22	22	22	22	22

Table 59 presents the Sensor data for the CO-3 session. Because the within group variability was rather high, the criterion for a significant mean difference ranged from .28 to .36. None of the Control 3 – Spiral 3 differences reached this value. However, 6 of the 8 means were in favor of the Spiral, and in several cases, the mean difference was substantial, accounting for about two-thirds of the required statistical difference.

Table 59

Gradesheet Data and Analysis Results for CO-3 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Airmanship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 3 (mean)	2.33	2.38	2.43	2.48	2.33	2.29	2.19	2.21
Control 3 (VAR)	0.23	0.25	0.36	0.36	0.23	0.21	0.26	0.17
Control 3 (N)	21	21	21	21	21	21	21	24
Spiral 3 (mean)	2.50	2.22	2.67	2.72	2.28	2.39	2.28	2.29
Spiral 3 (VAR)	0.26	0.18	0.24	0.21	0.21	0.25	0.21	0.22
Spiral 3 (N)	18	18	18	18	18	18	18	17

The final analysis, for Sensors in the S-EP-2 session, is depicted in Table 60. Within-group variability was considerably lower in the sim sessions, and consequently, several of the differences reached significance. Unfortunately, both were in the opposite direction to that predicted, as Control 3 students received higher ratings on Airmanship and Flight Discipline. The other differences were fairly close, though they tended to favor Control 3 except for Admin Checks.

Table 60

Gradesheet Data and Analysis Results for S-EP-2 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 3 (mean)	2.25	2.33	**2.63**	2.42	2.26	**2.58**	2.29	2.28
Control 3 (VAR)	0.37	0.23	0.42	0.34	0.20	0.25	0.39	0.38
Control 3 (N)	24	24	24	24	23	24	24	25
Spiral 3 (mean)	2.13	2.38	**2.13**	2.25	2.25	**2.25**	2.13	2.13
Spiral 3 (VAR)	0.12	0.25	0.12	0.20	0.20	0.20	0.12	0.12
Spiral 3 (N)	16	16	16	16	16	16	16	16

Conclusions and Recommendations.

Overall, the results of the Level I survey presented a fairly clear picture of student reaction to MTT technical access, usability, utility, and training preparation. First, the Web-based implementation of MTT had some technical difficulties; almost half of the respondents experienced some type of "negative impact." As noted earlier, the software was modified to mitigate system delays and with Class 10-01, the "negative impact" rate was reduced to 20%. Periodic system lags were frustrating for students and decreased their confidence in and use of the system. It also undoubtedly colored their perception of the overall training value and benefit of the system. It is unfortunate the limited bandwidth at Creech AFB, coupled with the high processor demands of MTT's multiple clocks, combined to create these periodic delays. Despite the technical problems, MTT received fairly respectable usability ratings. Fifty percent of the sample rated the interface as "positive," with only 9% giving it a "negative" rating. The index page and tutorial were considered pluses for system usability, with two-thirds of the sample rating these features as "easy to use." MTT's training utility ratings were surprisingly high, given the technical difficulties some students encountered. Over 60% of the respondents viewed the system as "moderately or very useful" for training. Elements that enhanced this training value included the scorecard, green and red dots, use of a higher priority task, and shifting the task quadrants between scenarios. The clock timer and top score indicators were not as well-received; the clock timer was viewed as "motivating" by only one-fifth of respondents. Alternatively, respondents gave higher marks to the training structure provided by the progressive increase in difficulty in the scenarios. With regard to the individual tasks, the memory, auditory, and visual tasks were considered "good training" by 50% or more respondents, with the addition and memory tasks viewed as "challenging" by at least half of the sample. Assessments of "unrealistic" were low, suggesting attempts to make the tasks relevant to Predator Operator tasks were largely successful.

In terms of future use of MTT, several recommendations should be considered. First, more class time (i.e., at the end of EA when the introduction to MTT is given) should be devoted to ensuring all students are aware of the structure and long-term goal of MTT training. In particular, the scenarios were designed to truly challenge the student such that each scenario could be exercised more than once, perhaps even multiple times, so that students' scores show a progressive increase. Nevertheless, it was clear from survey data the average time spent on MTT (~41 minutes) was such that students did not experience a given scenario more than once. This was due, in part, to the technical difficulties described earlier. However, for the cognitive benefits of attention management to accrue, more MTT scenario repetitions will undoubtedly be needed. To this end, a brief description was added on the Scenario Index page describing "what this training is designed to accomplish." This was created prior to Class 10-03,. Because MTT is not as obviously linked to enhanced CRM skills as are the EA and ICH, a reminder to students of the purpose of this unusual type of training would be advisable.

Finally, the display of the top score and clock timers on the MTT scenario screens should be reviewed. These items were designed to enhance student motivation and feedback; these items were found to be effective in other applications. Their lowered impact in this context may be due to several factors, including location, size, and possibly the students' expectations of the training they were to receive.

This Level II analysis presented fairly strong evidence of student learning during MTT training. While data were sparse for the most direct measure (improvement in performance during repeated scenarios) there was, nevertheless, an average increase of 16% for those students who elected to repeat a given scenario. As well, the absolute levels of maximum performance within a block of scenarios was consistently high for almost all students, again further evidence students were acquiring the requisite skills to successfully perform the MTT tasks. Although student attrition (failure to finish the training once started) was higher than desired, the technical difficulties associated with system lag, which was particularly a problem with the beta-test class, was largely responsible for the loss of students over time. Student performance on the individual MTT tasks were consistent with reactions in the Level I student critique survey. The calculation (addition) task was by far the most difficult as evidenced by a higher incidence of misses. This was presaged in the survey, where students complained of the task's difficulty and lack of realism. In that analysis, modifications to the task were recommended that will both modulate its difficulty and make it more aviation-relevant.

At this point in the project, the primary concern with MTT was the reluctance of students to repeat scenarios to maximize performance. While most students performed well, there was still room for improvement, particularly for the last 4 test scenarios. Here, average maximum performance was around .70, suggesting more repetitions would yield further improvement. Several steps can be taken in this regard. First, a notice was placed on the Birds of Prey MTT directory, in the form of a rollover, which explained the importance of scenario repetition and the need to make attention management as automatic as possible. Second, CTI instructors ensured scenario repetition is encouraged and emphasized as an objective of training. Finally, although it is outside the scope of the this project, it would be advisable in the future to modify the software so students receive a recommendation to repeat a scenario if their score is below a certain level, (that threshold would depend on which scenario they are performing and how many times it has been repeated [if any]). Providing system prompts to encourage scenario repetition would remind students of the need to practice and make it easier for them to perform. As part of this modification, the colored links in the directory (that presently only change color once a scenario has been played) would be altered. Instead, the directory should also indicate the number of times the scenario has been played by a given user (based on the user name entered during log-in).

In closing, the training gradesheet analysis for the Spiral 3 comparisons provided a mixed picture that should probably be scored as a "tie." On one hand, Spiral 3 students enjoyed an advantage on the CO-3 sessions for Pilots and Sensors. In the case of the former, eight of the nine means were larger for Spiral 3, with one difference (Flight Discipline) reaching significance. For the CO-3 session with Sensors, none of the differences reached statistical significance, with six of the eight means larger for Spiral 3. The opposite pattern was evident in the S-EP-2 sessions. For Pilots, though none of the differences reached significance, eight of the nine mean differences were in favor of Control 3. For Sensors, two mean differences were statistically significant (Airmanship, Flight Discipline) and in favor of Control 3. Also, six out of seven of the means were larger for Control 3 relative to Spiral 3. Since the instructor cadre was completely different for the aircraft and sim sessions, it was difficult to discern whether these differences reflected true performance effects or whether it was a combination of cohort effects (i.e., certain classes were stronger than others) and different grading criteria being used by the instructors. As seen in the other spiral comparisons, the gradesheet data were probably the most "flawed" and weakest form of proficiency data; hence, these results should be taken less seriously.

SPIRAL 4 Analysis (EA + ICH + MTT + GemaSim)

Level I (Student Critiques of GemaSim)

This analysis summarizes the reactions of student Pilots and Sensor Operators to the GemaSim Team Training (GTT) simulation beta-tested at Creech AFB in January 2010 (Class 10-03) and then administered again to 2 successive classes, 10-05 and 10-06, in March and April 2010, respectively. Due to the success of the beta-test implementation, and the very high student reaction to the training, data from the Class 10-03 were retained for this analysis. A total of 55 students (20 in Class 10-03, 19 in 10-05, and 16 in 10-06) participated in GTT over the 3 classes.

GemaSim Training Methodology.
Recall GemaSim is the fourth and final CRM training intervention. As such, it is part of Spiral 4, which includes EA, ICH, and MTT. The GTT is a self-contained, local area networked simulation that allows a group of 4 students, working on 2 laptops, to plan for and conduct a 30-minute "deep space" mission. In the implementation used here, the 4-person team competed with another 4-person team, also using 2 laptops. Pairs of students worked on each laptop, with systems specifically designed so students must share information to operate their systems successfully (e.g., keep batteries charged, navigate, land on planets, avoid black holes). Points were awarded based on completing experiments while on the planets, answering questions, completing checklists, and performing other duties during the 30-minute period. A critical requirement, and one that caused considerable stress and workload, was the mission crew must return to base within 30-minutes, otherwise the mission was a "failure."

The version of GTT used here was a streamlined and much shortened version of GemaSim that has been deployed in Europe. It appears this project was the first use of GTT within the United States. To permit GTT implementation within the students' compressed training schedule, a total of 2½ hours was devoted to the training. This consisted of: a 20-minute briefing by a CTI instructor; a 10-minute planning session for the first mission; a 30-minute familiarization flight; a 20-minute debrief; a 10-minute break; a 10-minute planning session for the second operational mission; a 30-minute operational mission; and a final 20-minute debrief, followed by an open-ended period where students completed a 2-page course critique sheet and had any remaining questions answered.

In addition, students were asked (and all complied) to complete a 15-minute CBT prior to the class, where they were given some basic information on the operation of GTT (systems, controls, mission objectives, etc.). GTT is intentionally designed to stress students' workload so they are forced to work together to share information, shed tasks, reprioritize tasks, make decisions quickly, communicate clearly, avoid channelized attention, and generally work together as a crew. Since these objectives line up quite well with the four Human Factors skills targeted for the project, GTT would seem a perfect fit for these purposes. As the data will show, this assumption turned out to be correct.

While the GTT system was designed with great care, considerable flexibility was given to the four-student team regarding their own internal organization of tasks and duties. In essence, one laptop was utilized for piloting while the other was for navigation. The two students at each laptop decided who controlled the mouse and who was responsible for monitoring the display and reviewing the hardcopy checklists available. There were two hardcopies for each four-person station, one for operational checklists and one describing how the experiments were to be conducted once on a planet. Thus, students were forced to share information (and even the checklists) if they were to be successful.

Each four-person team had two CTI SMEs observing the students interact with the system and communicate with one another. Thus, a total of four SMEs were needed to cover the two teams. Each SME was responsible for scoring the CRM behaviors of two students, either the two manning the pilot station or the two working with the navigation station. For scoring, the SMEs used the same HF sheets as were used to collect the Level III data in the CO-3 and EPE sessions. The intent was to collect CRM behavioral ratings for the familiarization flight and operational missions, where a comparison of the two sets of ratings yielded a measure of Level II learning.

Because the total session length for GTT was at least 2½ hours, 2 sessions of GTT were scheduled for each class, with 1 in the morning and the other in the afternoon. To handle a class of 20 students, this required 10 students in each session, with 2 students serving as observers in each session. (Class 10-05 had 19 students participate, so 3 of the 4 "teams" had a fifth observer while the fourth team did not. Since Class 10-06 had 16 students participate, there were no observers for any of the 4 teams.) Observers were told to take notes on the team's performance and offer critiques during the debrief. They were not allowed to help the students or engage with the system in any fashion. The training experience of the observers was somewhat limited, and was reflective of their critiques of the training session. Anticipating this limitation, an attempt was made to assign more senior Pilots and Sensors to the observer position. Also, Pilots and Sensors were intentionally paired (to the extent possible) on teams so they would have opportunities to work together and share information. The casual observation of this arrangement indicated this was a very successful and highly productive approach, and is something that should be repeated in future GTT training implementations.

Structure of the Critique Sheet and Methods of Analysis.

To assess student reaction to GTT, a modified form of the EA course critique sheet was used.. Seven new items were rated by students on a five-point Likert scale: "strongly agree" (i.e., the highest rating), "agree," "neutral," "disagree," or "strongly disagree." A non-rated option of Not Applicable was also available. The data were entered into an Excel worksheet, where statistical calculations (means, variance) were performed. To support quantitative analysis, these responses were converted into a numeric scale by assigning "strongly agree" to 5, "agree" to 4, "neutral" to 3, "disagree" to 2, and "strongly disagree" to 1. Responses of "not applicable" were disregarded in any calculation of average ratings. The seven items were:

1. *I used CRM skills in this training.*
2. *I felt I was challenged in this training.*
3. *I learned from this experience.*
4. *This training helped develop my CRM skills.*
5. *I see benefit in this type of training.*
6. *I enjoyed this type of training.*
7. *I prefer this type of training over other CRM training I have experienced.*

Six items from the EA critique sheet were also included. These asked for comments concerning: what one or two things students liked about the course; what one or two things were needed to improve the course; and in what ways did the training improve attention management skills, task prioritization skills, course of action selection/decision-making, and crew coordination. As with the other intervention critiques, student responses were anonymous, so there was no way to link student reaction data to subsequent Level II or Level III performance. The comments for each student were also entered into an Excel worksheet for qualitative content analysis.

As each GTT class' data were collected, a separate worksheet tab was created, where quantitative and qualitative analyses were performed on each class separately. Interpretive analyses were also performed to compare the classes on an item-by-item basis to discern trends. The results of these analyses for the three GTT classes are described in the following subsections.

Data Quantity and Quality.

Before listing the results, information is provided about the quantity and quality of data collected. First, all of the 55 students spent considerable time filling out the critique sheets, resulting in complete data from all 55. In several cases, students scored a particular item as N/A, resulting in a reduced N when that item's average was computed. However, this only occurred 7 times out of 385 possibilities (i.e., 7 rating items x 55 students), or less than 2% of the time. This resulted in a virtually complete set of rating data for performing statistical analysis.

With regard to comments, the student responses on the critique sheets were remarkable. Typically in studies of this type, if 75% of the subjects provide at least 1 comment, the qualitative data obtained are considered sound. For the 3 classes receiving GTT, all 55 students provided comments, for a 100% response rate. Such a response rate is amazing. Moreover, examination of the critiques reveals students did not simply make 1 or 2 comments, as typically is the case. Rather, almost every student made multiple comments, where many provided comments to all 6 items. For example, when asked to state 1 or 2 things they liked about the course, 52 of the 55 students (95%) supplied a comment. Such a response rate is frankly unheard of. Moreover, if the number of comments provided per student is calculated, the average is slightly over 5 (out of six items). In fact, 32 of the 55 students (58%) provided comments for all six items. The high rate of response is indicative of very positive student reactions to the training.

Rating Data.

Using the numeric conversion scheme described above, the average Likert-rating was computed for each of the first seven items on the critique sheet. Table 61 presents the average ratings for each of the three classes. The overall average rating across all seven items, appears (in bold font) in the bottom row of the table; the corresponding overall averages for each critique item appear in the right-most column.

Table 61

Average Likert-Ratings for the First 7 Items on the GemaSim Critique Sheet

Critique Item	**Class**			**Item Average**
	10-03	**10-05**	**10-06**	
1. I used CRM skills in this training	4.9	4.8	4.9	4.9
2. I felt I was challenged in this training	4.9	4.6	4.8	4.8
3. I learned from this experience	4.9	4.6	4.6	4.7
4. The training helped develop my CRM skills	4.9	4.7	4.6	4.7
5. I see benefit in this type of training	4.9	4.7	4.5	4.7
6. I enjoyed this type of training	4.9	4.7	4.1	4.6
7. I prefer this type of training over others	4.7	4.7	4.3	4.6
OVERALL CLASS AVERAGE	**4.9**	**4.7**	**4.5**	4.7

Examination of the overall class averages in the right-most column of Table 61 reveals the GTT ratings are extremely high. Indeed, the overall average rating is 4.7, only .3 below the maximum rating of 5. Clearly, the students' reactions to the GTT were quite positive. This view is further seen in the individual item averages, where the lowest was 4.6, only .4 below a perfect rating. To provide a statistical yardstick to compare average differences, the average standard error about the mean was computed for each class, and then that standard error was averaged. With this method, a difference between averages of .3, or almost a third of a unit difference, is considered statistically meaningful. By this reckoning, the first five item averages do not differ statistically from a perfect rating of 5.0.

The middle columns of the table show how the item ratings varied across the three classes. While all averages were high, there were some notable differences, where the ratings were lower for Class 10-06 relative to the two other classes. In particular, students' responses to the items concerning enjoyment and preference of this type of training were significantly lower in Class 10-06 (though still high overall) than the other two classes. Average ratings on the other five items did not differ very much across the three classes.

In looking at the individual student ratings, 26 of the 55 students, or 47%, gave a "5" rating to all seven items. Only eight students (15%) assigned any items a rating of "3" or lower. One of these students, who assigned lower values to items 6 and 7 (enjoy this type of training, prefer this type of training), served as an observer on his particular team. Since being an observer is clearly a more passive role than what most students would want in a hands-on training session, the lower rating is not surprising. This issue will be addressed again in analysis of the comments below. Several students had suggestions about how the issue of an assigned observer might be addressed in future training sessions. On the other hand, most of the lower ratings came from Class 10-06, where five students expressed either neutral ("3") or slightly negative ("2") attitudes toward the enjoyment of this training.

In sum, the training was viewed quite favorably by virtually all the students. Indeed, even the students who assigned neutral or slightly negative ratings to some of the items, still assigned ratings of "5" or "4" to the items concerning the challenge and learning benefits of the training. Clearly, this type of hands-on CRM training is highly thought of by students, both in terms of its enjoyment value, as well as the learning they believe they are receiving.

Comment Data.

Following the Likert-items, the critique sheet asked for student comments indicating one or two good things about the training, one or two things that needed improving, and how the training helped/or could help each of the four targeted HF skills: attention management, task prioritization, COA selection, and crew coordination. Below, the most frequent (modal) comments for each question are analyzed, as well any other comments that provide unique and critical insights concerning the team training. Note qualitative content analysis is far from an exact science, so some interpretation is required. However, the concrete nature of the team training content and the questions asked make most of this interpretation fairly straightforward.

One or Two Good things about the Training.

As noted above, 52 of the 55 students provided comments to this question. A brief paraphrase of the main categories of comments is provided, with the number of students giving that comment indicated in parentheses. Note that because students provided more than one answer to this question, the numbers add up to more than 52:

Forced focus on teamwork and communication (14)

Hands-on/interactive (12)

Fun and engaging (9)

Challenging in a good way (9)

Educational (7)

Shows effects of time stress on individual and team performance (4)

Non-aircraft specific mission and scenarios (3)

Instructors (1)

One or Two Things about the Training that Need Improvement.
When asked how the GemaSim training could be improved, 50 of the 55 students made comments. The responses to this question are given below, where again the numbers of students falling in that category are given in parentheses. As will be evident, there is a wide range of suggestions offered for improvement, where some are mutually exclusive. Also, a number of students indicated nothing should be changed:

Nothing (8)

More task and system familiarization up front (6)

Less instruction, more hands-on (3)

Have more missions (3)

Make more like the Predator (2)

More training time (2)

Correct some of the book answers (2)

Simplify the CBT (2)

Explain CRM principles during the game (1)

Let the observer play (1)

Pocket checklist (1)

Let crews see each others' displays (1)

Add MICs (mission intelligence coordinator) to the crew (1)

Dual control on each computer (1)

Course Impact on Attention Management.
For this and the remaining items, the analysis focused mainly on extracting comments that offered useful insights regarding students' expression of what they learned about the targeted skill. These statements provided the best and most direct indication of what students took out of the course and, hopefully, applied later in the training curriculum. The focus in this analysis is not on the quality of word-smithing or clarity of expression, but rather, on what students indicated they were learning.

Forty-five of the 55 students offered fairly insightful responses to the question of how the course helped their attention management skills. These included the following, some of which were paraphrased (and in some cases, combined) for readability and/or conciseness:

It helped to address the important issues vs chaff.

It helped get me to looking around to different things every couple of minutes.

Not get locked on one task

Had to be aware of things I couldn't control

Helped instill importance of being logical in making decisions

Given lots of unrelated information together that had been sorted through

I now understand channelized attention a lot better.

It identified my strong and weak areas.

Helped to ID areas where I need to improve

Emphasizes that good listening is a key to good teamwork

Created a requirement to get input from 3 other people and 2 computers

Makes me focus on doing things right the first time

It was good review and practice in attention management.

Showed importance of crew comm. and personality traits as well as teamwork

Showed importance of good cross-check and speaking up

I expect this training will help me in the future.

The level of stress helped my attention management skills.

There was a lot to focus on.

Warnings were realistic and attention-grabbing.

Taught me not to get fixated on one thing

It helped to look at prioritization.

It makes you aware of all the skills needed to accomplish a task.

Good exercise in how to prioritize

The first attempt showed we need to constantly monitor the ship for the second attempt.

Required continuous scan of unfamiliar instruments

No system knowledge or manual made me pay more attention

It helped be more vocal and apprehensive in learning.

Told me where to focus first based on mission goals

Vast improvement over the Winnebago

It helps me to learn what skills other guys have and incorporate those into everyday ops.

Highlighted areas that needed improvement and allowed opportunity to use those skills

Decreased task channelization, channelization awareness, prevent channelized attention

Things moved so quickly I had to be disciplined about my crosscheck.

It enforced the message of not having tunnel vision to coordinate with your team and succeed.

It helped me prioritize what needs more attention.

Learned through observing others

Seeing issues from the first mission helped me focus on the second mission and keep priorities straight.

Learning how, via the game, not getting channelized as much on the second run

Course Impact on Task Prioritization.
Thirty-nine students provided comments to the question concerning GTT impact on task prioritization. As above, all of the comments were positive, and included the following:

> *Enough simple tasks that it was easily highlighted for me when I didn't prioritize*
>
> *It made me choose which things were most important to my mission.*
>
> *I understand better how to make decisions that align with the shared mental model.*
>
> *It helped by debriefing where I was weak and how to improve.*
>
> *It made me think about my day-to-day ops a sensor operator.*
>
> *Exercises forced me to prioritize which tasks were most important and line them up properly*
>
> *The more iterations the crew did, the better their prioritization – one more iteration would cement these ideas.*
>
> *As a team we realized ignoring the tasks would benefit us more.*
>
> *Made me realize how to task prioritize better*
>
> *Showed must just how catastrophic it can be to focus on the wrong thing first*
>
> *Case studies about past mistakes definitely helped.*
>
> *Don't get channelized attention.*
>
> *We had to balance priorities in terms of battery life, times, points, etc.*
>
> *Survival of crew and completion of mission came before additional tasking.*
>
> *Aircraft and safety comes first.*
>
> *After the mission I realized an improvement.*
>
> *Precise communication to understand what is the most critical task*
>
> *It makes you think more about task prioritization.*
>
> *Shows me how some tasks were much more important than others and to focus on those first*
>
> *Great example of what we need to think about before each mission*
>
> *Helped me to realize that there were many things I was instinctively keeping track of (power, navigation)*
>
> *Lots of tasks to accomplish as a team*
>
> *Learned by observing*
>
> *It reinforced my line of thinking.*
>
> *Learning from my mistakes during the first mission and fixing them before the second was helpful.*
>
> *After the first mission, you learn what are the top priorities and discuss what to do.*
>
> *Having the first run vs the second run really helped out.*
>
> *Determine what is critical right now*
>
> *Once we learned what was important we were able to prioritize.*

> *It helps recognize priority breakdown methodology much like the altitude chamber helps us recognize hypoxia.*

Course Impact on COA Selection.

Thirty-two of the 55 students gave comments to the question concerning training impact on COA selection. These comments, in condensed form, are summarized below:

> *It helped us in our prebriefing skills.*
>
> *Helped me to realize how time critical things can be a times (making decisions quickly and pressing forward)*
>
> *Had to make decisions as a team*
>
> *Made us focus on the score*
>
> *Showed me how to use the crew to take care of problems while I continued to fly*
>
> *Seeing the mistakes I made first hand helped more than anything.*
>
> *We had to act on info to decide whether to execute or knock it off based on what info I believed to be correct.*
>
> *Learned to solicit more info when the time is available*
>
> *Helped me to ID what tasks were needed most in the situation*
>
> *Must have someone in charge to make the final call – but is still critical to listen to most junior input*
>
> *It helped me to demonstrate the decision train and how one COA can lead to another and change a whole mission.*
>
> *I saw the improved COAs available by paying attention to all crew members.*
>
> *We had a lot of times where we had to choose between mission failure or getting home.*
>
> *Utilize all resources before making a decision and be flexible*
>
> *Forget about bad decisions and move on.*
>
> *Understand how to use the time properly to make on-time decisions.*
>
> *Had to prioritize tasks vs fuel remaining*
>
> *Beginning to work together and delegate*
>
> *We made several decisions to figure out our COA, especially when we had unexpected events occur.*
>
> *Allowed me to query everyone in the group before making decisions*
>
> *Determine what is critical right now.*
>
> *Help recognize what priorities have broken down*
>
> *Helped to identify the more critical functions/tasks/routes of action*

Course Impact on Crew Coordination.

Finally, 45 of 55 students provided comments to the question concerning training impact on crew coordination. In some cases, students also used this item, the last block on the sheet, to offer their summary opinions of the course. An attempt was made to separate those from the coordination-specific comments, which are listed first:

The value of crew coordination cannot be measured – it is a value that is the greatest asset for success.

We had to have good crew coordination in order to complete the tasks and mission.

It helped to be able to establish an effective/highly tasked crew in a very small timeframe.

It taught the importance of farming out tasks.

We had to work as a team.

I need serious work on this and interacting with several people on unfamiliar tasks forces interaction.

I definitely feel I have better crew coordination now that I have heard what my fellow coworkers have said.

I understand better how to integrate as a crew.

It made me think of making sure the task at hand is being done correctly.

Helped me to open my mouth and talk

I can see the difficulty in developing a shared mental model and I will be cognizant of my own biases and assumptions.

Communication improvement

Provided briefing time and tasks that required coordination and comm. between team members

Be assertive and speak up when you see a problem.

A lot crew coordination occurred to execute this simulation and thus my skills improved.

Checklist discipline and closing checklists

A realistic approach to melding four different crew members into one effective decision-making team

The need to maintain open communication with other crew members

Talking with each other more

The communication factor

Allowed hands on experience

Now I begin to understand why pilots say and do some of what they say/do when we get in the GCS.

I now see what a pilot might expect (incorrectly) from a MIC from a CRM perspective given MIC's lack of CRM training.

By listening to what all the instructors had to say and by observing teams

Helped me practice asking crew members for information and give information

It's all about communication.

Helped get me in a mindset to communicate with my teammates

Improved comm. flow, talking when appropriate, listening when appropriate

Required to coordinate due to the system setup

Be directive and clear.

Solicit inputs from anyone who has information.

Need to rely on using my SO better

Had to work together to get all the info – one of the most useful and effective training I've ever had. Well done!

Great crew coordination exercise because everyone was under pressure – made us work together

Forced us to plan the mission and keep us talking – I really liked this because we actually get feedback on how we do

Forced us to work as a team to accomplish the mission

Doing this training earlier in the syllabus might serve dual purposes by getting to know each other plus the CRM piece.

I recommend the observer to be experienced with CRM.

Great training aid

It improved my problem solving and communication skills.

The overall difference between the 1st and 2nd attempts was good to see, particularly in light of the debrief.

Initial run may cause some discouragement

Please remove or edit the trick question on "28 days in a month."

Crew communication might have been negative training – need more intro upfront about what good comm. looks like

Might want to experiment with assigning a leader

Would help to define crew positions first

Level II (Evidence of Learning for GTT)

This analysis summarizes the results of student learning that occurred as a consequence of receiving GTT. This training was administered to three classes: 10-03, the beta-test class, during the first week of January 2010; 10-05 in February, and 10-06 in late March. Note Class 10-05 was a distinct group because it was part of the USAF's "beta-test" experimental assessment of non-rated Pilots for the Predator program. As such, these students had little prior tactical and flying experience, making them very distinct from the other classes. While Class 10-05 was omitted from the Level III (transfer of training) assessment for this reason, they were nonetheless a valid class for gauging student reaction to GTT and measuring the extent of learning that occurred over the course of GTT. Thus, Class 10-05's data were included in the Level I analysis. As well, Class 10-05's learning data were included in the analyses reported herein. In this regard, all three classes were treated as a Spiral 4 class. Recall Spiral 4 refers to a full receipt of the four CRM training interventions: EA, ICH, MTT, and GTT.

Logic of the Analysis.
In the analysis of Level I student critiques, a description of the GTT and the methodological details concerning its implementation were given. Students were trained in teams of four, where two students worked on the same laptop. The four-student team competed against another four-person team to see who received the highest score during a 30-minute space mission. Two missions were conducted during the training. A fifth student served as an observer on some of the teams but did not take part in the mission. The data collected were on the four-student team, where two sessions were conducted for each class. Thus, there were Level II learning data from a total of 48 students, (12 4-person teams).

To measure student learning, each student's performance on the first mission was compared with their performance on the second mission. The total score received was a team score, so it was necessary to have other indices tied more directly to the individual student. For that reason, two CTI SMEs observed each team, rating students using the Human Factors (HF) data sheet used for the Level III assessments. Thus, the SMEs assigned a five-point (0-4 scale) rating on the four Human Factors skill areas: Attention Management, Task Prioritization, Course of Action (COA) Selection, and Crew Coordination. They also assigned a rating to the degree of Instructor Intervention (where 4 = no intervention required) and Crew CRM Performance. Besides the category rating, the SMEs also checked whether six to seven CRM behaviors were evident during the mission, assigning it a + if it was positive and – if it was negative. The absence of a check was taken as a neutral. As well, there was space to make comments describing the basis for the rating.

Using the above method, there were four indices of learning that were subjected to analysis. The first was comparing the team's score on the two missions. Since this was a team score, there were 12 data points, one for each of the 12 teams trained. Evidence of learning appeared if the total score was significantly higher on the second mission compared to the first.

The second measure compared each student's SME-supplied Human Factor's ratings on the two missions. For this measure, there were 16 students in each class and ratings for the six areas covered on the HF data sheet. Learning was reflected in significantly higher ratings on the second mission compared to the first. Separate analyses were performed on the three classes.

The third measure compared the number of CRM behaviors scored positively or negatively on the two missions. Again, a greater preponderance of positive behaviors in the second mission compared to the first revealed evidence of learning. As above, there were 16 pairs of scores for each of the three classes, where the classes were analyzed separately.

Finally, a qualitative assessment of the SME comments described both the extent of the problems experienced in the initial mission and the nature of those problems. Comparison to the second mission indicated whether these problems were rectified through student/team learning, and which problems remained even at the end of the second mission. Comments were tallied, for a quantitative index of learning, and content analyzed to pinpoint specific areas of strength and weakness. For this analysis, the focus was on the qualitative aspects of the comment data.

Learning Index 1 – Team Mission Performance.

The first measure of learning came from comparing the total points each team scored during the two missions. These data are displayed in Table 62, where, within a class, the four teams were labeled A through D. In particular, Teams A and B competed against each other in the first session for the class, whereas C and D were competitors in the second session.

Table 62

Team Game Performance as an Index of Learning

	Class 10-03		Class 10-05		Class 10-06	
Team	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**
1A/2A	294	2114	1770	998	374	1526
1B/2B	1150	2186	324	1306	836	2030
1C/2C	752	1856	1066	1886	1224	2364
1D/2D	208	1904	610	2250	434	1536
MSN AVG	**601**	**2015**	**942**	**1610**	**717**	**1864**

It is evident from the table mission performance improved substantially for all but one of the teams (Team A in Class 10-05) between the first and second mission. As seen from the bottom row, the average increase was anywhere from 700 to 1500 points, all fairly large increases. The one exception was Team A in Class 10-05, whose performance actually declined by 800 points. This unusual decrease was a joint combination of two factors: 1) the team's exceptionally high performance on the first mission (over 500 points higher than the next closest time); and 2) experiencing a very difficult black hole encounter generated by the game on a stochastic (chance probability) basis. Discounting this one team, all other teams improved substantially, regardless of their initial level of performance. A paired t-test was performed on these data to see if performance on the second mission was significantly higher than the first mission (Miller & Freund, 1965). The results confirmed the visual inspection, with t = 24.633, df = 11, p < .001. Thus, there was very strong evidence of learning based on this measure.

Learning Index 2 – SME Ratings of CRM Categories.
The second measure of learning entailed comparing the SME-supplied ratings of the six CRM categories (four Human Factors skills, instructor intervention, CRM performance) for each student on the two missions. The mean ratings for the 16 students in each class (where the students from all four teams were treated as a single group) on the two missions for the six categories are displayed in Table 63.

Table 63

Average SME Ratings for the CRM Categories

	Class 10-03		Class 10-05		Class 10-06	
CRM Category	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**
Attention Management	1.19	2.06***	1.06	2.06***	1.06	1.94***
Task Prioritization	1.44	2.56***	0.94	1.94***	1.38	2.50***
COA Selection	1.13	2.31***	1.50	1.81	1.44	2.25***
Crew Coordination	1.50	2.25**	1.31	1.81*	1.38	2.06**
Instructor Intervention	2.33	3.67**	--[a]	--	--	--
Crew CRM Performance	1.19	2.31***	1.25	1.94**	1.38	2.25***
Mission Average	**1.46**	**2.53**	**1.21**	**1.91**	**1.33**	**2.20**

*p<.05. **p<.01. ***p<.001

[a] This measure was dropped from further analysis because the instructors were required on occasion to intervene due to technical factors other than student performance, rendering the rating invalid.

Looking first at the bottom row of the table, it was evident the average SME rating, collapsed across CRM categories, increased considerably between the first and second mission. This was seen for all 3 classes, although the size of the increase varied with class. In particular, the overall average SME rating was more than a full point higher on the second mission compared to the first mission for Class 10-03; the corresponding difference was .70 for Class 10-05 and .87 for Class 10-06. Moving up the table, there was a notable increase in the SME average rating for each of the 6 CRM categories in each of the 3 classes. The size of the difference varied considerably, with the smallest increase being .31 (COA Selection, Class 10-05). Most of the increases, though, were at least .5 (half a scale unit) or higher. Examination of the individual student data revealed most, if not all, students exhibited an increase in rating from the first to the second mission. The significance of these differences was tested by computing a paired t-test for each category and each class, separately.

Looking first at Class 10-03, the t-tests revealed strong statistical support for learning effects in all six CRM categories. In particular, the following results were obtained: Attention Management (t = 7.00, p < .001, df = 15), Task Prioritization (t = 6.26, p < .001, df = 15), COA Selection (t = 8.73, p < .001, df = 15), Crew Coordination (t = 3.50, p < .003, df = 15), Instructor Intervention (t = 4.30, p < .0012, df = 11) (It was not possible to obtain Instructor Intervention ratings for four students, resulting in a loss of four degrees of freedom), and CRM Performance (t = 4.70, p < .001, df = 15). Thus, there was a significant increase in SME ratings for every CRM category, consistent with an interpretation that there was a strong learning effect present throughout the data.

Turning to Class 10-05, while the increases in SME ratings from the first to the second mission were not quite as large as in 10-03, they nonetheless were sizable in most cases. The following results of t-testing were obtained: Attention Management (t = 4.90, p < .001, df = 15), Task Prioritization (t = 3.87, p < .0015, df = 15), COA Selection (t = 1.78, p < .10, df = 15), Crew Coordination (t = 2.24, p < .04, df = 15), and CRM Performance (t = 3.70, p < .001, df = 15). Curiously, the increase in SME rating for COA selection was rather small and failed to reach significance.

Finally, the SME rating increases for Class 10-06 were also substantial, falling in between the large increases noted for Class 10-03 and the somewhat smaller ones for Class 10-05. The results of t-testing on the individual subject data for Class 10-06 were as follows: Attention Management (t = 5.65, p < .001, df = 15), Task Prioritization (t = 5.58, p < .001, df = 15), COA Selection (t = 5.98, p < .001, df = 15), Crew Coordination (t = 3.15, p < .007, df = 15), and CRM Performance (t = 5.65, p < .001, df = 15). In sum, the results of t-testing with the second index provided very strong evidence that learning occurred during the GTT.

Learning Index 3 – SME Observations of CRM Behaviors.
The third measure of learning involved tallying the number of positive (+), neutral (0), and negative (-) observations the SMEs made to the 6 to 7 CRM behaviors under each of the 4 HF skills. A total score was then computed for each student in each skill. A negative score would indicate more negative behaviors were scored than positive. An average CRM behavior score across the 16 students was then computed. These average behavior scores are depicted in Table 64 for each of the 3 classes.

Table 64

Average SME-Supplied CRM Behavior Scores for the Four Human Factors Skills

	Class 10-03		**Class 10-05**		**Class 10-06**	
HF Skill	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**	**1st Mission**	**2nd Mission**
Attention Management	-1.38	1.44***	-1.25	1.63**	-1.19	1.50**
Task Prioritization	-1.00	2.25***	-1.19	1.69**	-0.88	1.88**
COA Selection	-0.75	2.56***	-0.63	0.88*	-0.69	2.06**
Crew Coordination	-0.53	1.93**	-0.69	1.69*	-0.81	1.88*
Mission Average	**-0.92**	**1.80**	**-0.94**	**1.47**	**-0.89**	**1.83**

**p*<.05. ***p*<.01. ****p*<.001

Looking at the last row, the average CRM behavior score in the first mission for all 3 classes was actually negative, -.89 to -.92, indicative of an average of almost 1 negative CRM behavior per category. During the second mission, students' performance improved dramatically, so the average CRM behavior score was substantially higher, reaching 1.5 to almost 2 positive behaviors per skill. Paired t-tests were performed on the 4 HF skills; the results for each class are presented below.

Starting with Class 10-03, as with the previous index, the t-tests revealed a strong learning effect in three of the Human Factors skills: Attention Management (t = 4.15, p < .001, df = 15), Task Prioritization (t = 4.82, p < .001, df = 15), and COA Selection (t = 5.00, p < .001, df = 15). A somewhat smaller, and only marginally significant effect was noted for Crew Coordination (t = 2.66, p < .018, df = 15). On balance, though, there was fairly strong evidence for learning between the first and second GTT missions.

Turning to Class 10-05, there were notable effects of learning with this third index, though the size of the differences were smaller than with the other two classes. Specifically, the following results were found with t-testing: Attention Management (t = 3.42, p < .004, df = 15), Task Prioritization (t = 3.55, p < .003, df = 15), COA Selection (t = 2.26, p < .04, df = 15), and Crew Coordination (t = 2.44, p < .03, df = 15). Thus, the somewhat smaller learning effects observed in this class with the SME ratings were mirrored in the CRM behavior scores.

Finally, the results from the t-tests with Class 10-06 showed fairly sizeable learning effects in all four Human Factors skill areas. Specifically, the following results were found: Attention Management (t = 3.36, p < .004, df = 15), Task Prioritization (t = 3.17, p < .006, df = 15), COA Selection (t = 3.81, p < .002, df = 15), and Crew Coordination (t = 2.45, p < .03, df = 15). Nevertheless, in examining the absolute levels of performance, be it the SME ratings or their observations of CRM behaviors, it was clear there was certainly room for further improvement, both for the teams and for individual students. Thus, students would no doubt have benefited from experiencing a third mission, where it is likely further learning would occur. Unfortunately, time constraints on the training schedule did not permit a third mission. As shown with the student critiques, several of the students indicated additional missions would have been beneficial (as well as fun).

Learning Index 4: SME Comments.
Analysis of the first three measures yielded clear, unequivocal evidence for strong learning effects in the data. Thus, there was a significant increase in the team's game score, SME ratings of HF proficiency, and positive CRM behaviors between the first and second GTT mission. For the final learning index, the SME comments were examined to identify more specifically what the students learned, either about the

GemaSim game itself or about ways to work together better as a team. The major themes documented in the comments for each CRM category are summarized below.

With regard to attention management, the SMEs noted a tendency for students in the first mission to get channelized attention, or “focus lock,” on one particular task to the exclusion of other duties. Often, this was charging the ship’s batteries, a fairly involved task. Sometimes this occurred at the expense of more pressing or important tasks, such as navigation, landing on a planet, or conducting the planetary experiment. By the second mission, students were more “proactive” in their crosscheck, with vastly improved situational awareness. They were also better able to check multiple sources of information in shorter amounts of time, indicating a more flexible attention management strategy.

The comments on task prioritization revealed a similar theme. In the first mission, tasks tended to be performed one at a time, with minimal multi-tasking. Sometimes, students spent too much time trying to answer a question, which was worth fewer points, than staying on course, completing an experiment, or returning to the base on time. They also concentrated, narrowly, on their individual tasks rather than addressing the collective tasks required of the team. By the second mission, tasks were shared more efficiently, where students were better at prioritizing tasks based on the number of points they would yield. They were also more likely to develop a task management plan during planning for the second mission, a strategy that would yield high dividends later in the mission.

For COA selection, SMEs commented during the first mission that poor decisions were made concerning routing (e.g., not going directly to a planet), timing (i.e., returning to base within the time limit), or waiting for others on the team to make a decision rather than offering the suggestion themselves. Some of the students failed to take an active role when decisions were made concerning critical COAs. By the second mission, students managed to allocate decision-making duties more efficiently, where there was a clear understanding of how individual decisions would impact the team’s total points. It was also more likely most or all of the four team members would make active contributions of information during COA selection.

Crew coordination improvement was clearly evident in the SME comments. In the first mission, the typical SME comments concerned students focused on individual, personal tasks at the expense of the team’s performance, failing to offer information to the team that only they had, asking questions but failing to persist when their question was not answered, doing someone else’s task by mistake, or reading checklists without waiting for confirmation of the items being performed. By the second mission, most of these problems were resolved in all the teams, where a more efficient arrangement of duties was allocated across team members. As well, students were engaged in efficient crosschecks, backing up each others’ tasks, with a much freer yet efficient flow of information among the four team members. The flexibility of the GemaSim gaming environment to let teams work through task assignments and establish priorities in their own ways was cited as a plus by all participants.

Level III (Transfer of Training for Spiral 4)

Analysis of Targeted Skills from HF form.

This initial analysis provides a summary of the first round of Level III data analysis for the Spiral 4 comparisons. The data set reported here is from the instructor ratings on the Human Factors data sheet implemented at Creech AFB. Recall these ratings were obtained from two carefully-chosen training sessions: the third and final Combined Operations session (CO-3) conducted with an actual aircraft, and the final emergency procedures check (EPE) conducted in the simulator with Standardization/Evaluation (Stan/Eval) raters.

The Spiral 4 comparisons encompassed four classes: 10-02, 10-03, 10-04, and 10-06. The Spiral 4 treatment classes were 10-03 and 10-06; classes 10-02 and 10-04 were the controls. Note Class 10-05 was excluded from this analysis because it was a "beta-test" class that was part of the USAF's experiment in using non-rated aviators for Predator Pilots. This class was excluded in the Level III analysis since these Pilots were far less experienced than their counterparts in the other classes.

The Spiral 4 treatment consisted of all four training interventions created for this project: EA, ICH, the MTT, and the GTT.

As with Spirals 2 and 3, the comparisons made in Spiral 4 were fairly direct, so it was appropriate to combine the data from Classes 10-02 and 10-04, and 10-03 and 10-06, and report the mean ratings of each. For ease of labeling, the two control classes will be referred to as Control 4. With participation rates of 80-100% across the classes, this gave a reasonable sample size (i.e., N = 16 to 21) for computing means and performing confidence interval testing.

Before presenting the results of the mean difference analysis, first how the ratings were distributed across the 5-point (0-4) Human Factors scales is reported. Recall the same 5-point scale is used that Creech AFB instructors used on the training gradesheets, where "0" was essentially unsafe flying and "4" represented exemplary performance. The distribution of the ratings across classes was examined to see how often the instructors used the 2 ends of the scale. Of the 858 ratings generated across the 4 classes (148 student sheets x 6 rating dimensions, with 30 ratings lost due to incomplete data), there were 113 "4" ratings, or 13.2% of the total. This was slightly higher than what was observed for Spiral 2 (10.4%) and Spiral 3 (11.0%). In addition, there were 5 "0"s reported, as well as 14 "1"s, which comprise 2.2% of the total. Thus, the instructors had 15.5% of their ratings on the 2 ends of the scale, which indicates instructors were quite willing to provide extreme ratings.

In addition, the Spiral 4 and Control 4 conditions were compared to see if there were any differences in how frequently the extreme ends of the scale were used. Some notable group differences were found on both ends of the scales. First, Control 4 students received a higher percentage of "4"s, 15.7%, compared to the Spiral 4 students, 10.7%. A statistical test of binomial proportions revealed this difference to be significant (p < .031) (Boersma, 2010). While this difference was in the opposite direction preferred, ALL five ratings of "0" were also received by Control 4 students. In this case, 2 students accounted for the unsatisfactory ratings, both from the Control classes. A test of binomial proportions showed this difference to be statistically significant (p < .022), which is in favor of Spiral 4. In addition, the frequency of "1" ratings was higher in Control 4 (2.4%) compared to Spiral 4 (0.9%). This difference failed to reach statistical significance (p < .090), although it was trending in that direction. Once again, there was evidence of very strong cohort effects, in which certain classes appeared to be unusually weaker (in this case, Class 10-06) than others, whereas other classes had a disproportionate number of very high-performing students. This large between-class variability added considerably to the within-group error variance, making it difficult to find statistically significant between-group differences.

Table 65 presents the means, variances, and sample sizes (N) for the Spiral 4 and Control 4 classes. The table is divided into four segments, corresponding to the CO-3 sessions for Sensors, EPE sessions for Sensors, CO-3 sessions for Pilots, and EPE sessions for Pilots. As with Spirals 2 and 3, the data are broken out by crew position since their tasks were so different and because they were rated by different instructors. Within each segment, the data appear in six columns, corresponding to the six dimensions rated on the HF data collection sheet: attention management, task prioritization, course of action (COA) selection, crew coordination, degree of instructor intervention, and CRM performance.

Table 65

Mean Ratings for the Six Classes Comprising the "Spiral 4 Comparisons" (0-4 Scale)

Sensor CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 4 (Mean)	2.60	2.73	2.80	2.67	3.13	2.87
Control 4 (variance)	0.97	0.78	0.74	0.67	0.55	0.84
Control 4 (N)	15	15	15	15	15	15
Spiral 4 (Mean)	2.35	2.71	2.41	2.53	2.76	2.59
Spiral 4 (variance)	0.24	0.22	0.51	0.26	0.44	0.38
Spiral 4 (N)	17	17	17	17	17	17

Sensor EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 4 (Mean)	2.28	2.33	2.22	2.22	2.44	2.42
Control 4 (variance)	0.57	0.59	0.54	0.65	0.38	0.54
Control 4 (N)	18	18	18	18	18	18
Spiral 4 (Mean)	2.50	2.50	2.56	2.56	2.73	2.50
Spiral 4 (variance)	0.27	0.27	0.26	0.26	0.21	0.27
Spiral 4 (N)	16	16	16	16	15	14

Pilot CO-3 Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 4 (Mean)	2.70	2.70	2.75	2.75	2.70	2.55
Control 4 (variance)	0.33	0.43	0.51	0.51	0.43	0.58
Control 4 (N)	20	20	20	20	20	20
Spiral 4 (Mean)	2.90	2.76	2.81	2.81	2.75	2.80
Spiral 4 (variance)	0.49	0.49	0.56	0.46	0.72	0.69
Spiral 4 (N)	21	21	21	21	20	20

Pilot EPE Sessions

Condition	Attention Management	Task Prioritization	COA selection	Crew Coordination	Instructor Intervention	CRM Performance
Control 4 (Mean)	2.55	2.65	2.70	2.60	3.10	2.75
Control 4 (variance)	1.00	0.87	0.75	0.99	0.94	1.04
Control 4 (N)	20	20	20	20	20	20
Spiral 4 (Mean)	2.48	2.52	2.48	2.52	2.93	2.52
Spiral 4 (variance)	0.56	0.56	0.56	0.56	0.81	0.56
Spiral 4 (N)	21	21	21	21	21	21

To facilitate visual comparison, the variance and N values are displayed in gray within their respective cells so the means are the most prominent measure within each segment. For statistical comparisons, just as in Spirals 1-3, a confidence interval around the Spiral mean was computed based on the respective variances of the spiral condition and its control counterpart. In essence, a confidence interval was constructed that would cover a 95% probability, by chance, of having another mean within its range. If the control condition mean was outside this confidence interval, then they were considered statistically different. As discussed in the previous analyses, this test turned out to be quite conservative since adjustments for the inter-correlations between the six rating dimensions were not made, which could have been done to reduce overall error variance. Such an adjustment was virtually impossible for this data set, however, since the sample sizes were markedly unequal and, more problematic, there were not both (i.e., EPE and CO-3) sessions’ worth of data for many of the subjects. Consequently, this was a fairly conservative testing method which likely understated the group differences present in the data.

The pattern in this data set was different from what was observed with the other spirals. Specifically, for two segments of the table (Sensors in the CO-3 sessions and Pilots in the EPE sessions) the control mean ratings were higher than their spiral counterparts for all six HF dimensions. Conversely, the mean ratings in the other two segments (Sensors for EPE sessions and Pilots for CO-3 sessions) showed the opposite effect. Despite the consistency in the directionality of the comparisons, none managed to meet statistical significance (i.e., based on the calculation of 2 x the square root of the estimated variances about the mean [Harris, 1994], which for this data set ranged between .38-.58). This failure was due primarily to the fairly high within-cell variances (~.50-1.00) for the various control group cells. These variances were a result of the greater number of extreme scores (i.e., 0, 1, 4) described earlier. Since the variances of the control group must be pooled with the spiral conditions, the result was a substantial error variance that dwarfed any of the mean differences.

Then data were tested to determine whether the within-group variances were significantly different between the Control and Spiral classes using Bartlett’s test of homogeneity of variances (Arsham, n.d.). Separate tests were performed on each segment of the table (i.e., CO-3 sessions for Sensors, EPE sessions for Sensors, CO-3 sessions for Pilots, and EPE sessions for Pilots). The results of the test revealed the Control 4 condition had significantly higher variances in the CO-3 sessions for Sensors (chi square = 19.919, p < .046, df = 5) compared to the Spiral 4 subjects. Though the test did not reach significance, there was also a trend toward higher variances in the Control 4 condition for the EPE sessions with Sensors (chi square = 14.117, p < .227, df = 5). In contrast, the hypothesis of equal variances was supported for the CO-3 sessions with Pilots (chi square = 4.620, df = 5, p < .052). Finally, the test of homogeneity of variances for the EPE sessions with Pilots was inconclusive (chi square = 7.345, df = 5).

The results of the instructor rating analysis yielded a mixed picture for the Spiral 4 comparisons. While instructors were using the full range of the scale, there was a pattern in the data that obscured the detection of statistically significant effects. Specifically, the Control 4 classes exhibited a substantially greater range in ratings, receiving a significantly larger number of both “4” and “0” ratings compared to the Spiral 4 classes. This, in turn, produced higher within-group variances for the Control group cells, particularly in CO-3 sessions with Sensors, as well as a moderate effect in the EPE sessions for Sensors. No evidence for higher variances in the Control group was found for either session type involving Pilots. As noted previously, having multiple instructors provide the HF ratings, a necessity in an operational training squadron, added considerably to the variation in the ratings obtained, and posed a stiff challenge to finding significance differences. While some modest, but consistent, effects were found in favor of the treatment classes in Spirals 1-3, these were not evident in Spiral 4. Instead, consistent but statistically non-significant differences in mean ratings were found, such that the Spiral 4 ratings were higher for Sensors in the CO-3 sessions and Pilots in the EPE sessions; the opposite effect was noted in the EPE sessions for Sensors and CO-3 sessions for Pilots. In the absence of any compelling explanation for this pattern of differences, it was likely to be due to spurious (random) factors.

Analysis of Negative Behaviors from HF form.

This second analysis examined the percentage of students within a condition who received a minus on the six to seven behaviors associated with each of the four HF skills. Examples of these behaviors included *effective cross-check, cross-check doesn’t stagnate*, and *switches attention* under the Avoids Channelized Attention skill. As reported in the corresponding Spiral 2 and 3 analyses, not every instructor used the minus designation, as some simply just scored behaviors as zero (i.e., they left it blank) or a +. Other instructors scored ALL behaviors a +, which was interpreted to mean it was simply observed. On the other hand, the assessment of the HF forms was that the instructors who took the scoring process most seriously, scored some behaviors negative, others positive, and the rest neutral. Consequently, the performance was best revealed with this measure by tallying the number of negative behaviors within a class and then converting that to a percentage so all classes were on a common scale.

As with the other spirals, the analytic strategy is to start with the most aggregated tabulation and then “drill down” into the data, by considering more subgroups in subsequent analyses. For the first pass through, data were aggregated across control and treatment (spiral) conditions, sessions (CO-3 and EPE), and crew positions to get a sense of which behaviors were stronger or weaker than the others. This aggregated tally is shown in Table 66.

Table 66

Percentage of Negative Behaviors across Conditions, Sessions, and Crew Position

Skill/Behavior	Frequency	Percentage
Avoids Channelized Attention		
effective cross-check	6	0.05
cross-check doesn't stagnate	4	0.03
switches attention	6	0.05
adjusts to different cockpits	4	0.03
not distracted by radios	11	0.08
able to shift attn w/o cues	3	0.02

Skill/Behavior	Frequency	Percentage
Task Prioritization		
knows high priority task	8	0.06
handle interruptions	5	0.04
returns to interrupted task	5	0.04
can suspend lower priority task	9	0.07
do tasks concurrently	6	0.05
aviate-navigate-communicate	5	0.04
Select COA		
considers all options	5	0.04
facts vs assumptions	5	0.04
avoids hasty decisions	6	0.05
doesn't take too long	12	0.09
ID potential risks	2	0.02
follow-on decisions	4	0.03
Crew Coordination		
divide tasks	3	0.02
perform team tasks	8	0.06
anticipate info needs	5	0.04
provides timely data	7	0.05
cross-checks others	7	0.05
maintain SMM	3	0.02
convey SMM	1	0.01
N	**132**	

The middle column of the table indicates the frequency or number of students in both control and spiral conditions who received a negatively scored behavior, considering both the CO-3 and EPE sessions together. That frequency was converted to a percentage by dividing each by 132, which was the total number of students in the 4 classes for which there were EPE and CO-3 data. Though no statistical tests were performed on these data, a fairly even distribution of percentages across the HF skills enabled identification of the strongest and weakest behaviors within the Spiral 4 comparison data set. In particular, the lowest percentages of negative behaviors, (i.e., those for which the aggregate percentage was below .02 [2%]) are left-justified. These were the "strongest" behaviors, and they corresponded to *able to shift attention without cues* and *conveys mental models*. The right-justified percentages indicate the behaviors that received the highest percentage of negative scores, in excess of 8%. There were 2 such "weakest" behaviors: *not distracted by radios* and *doesn't take too long to select a COA*. Interestingly, these behaviors were also identified as the weakest behaviors in the analyses of Spirals 2 and 3.

In the second pass through the behavior data, the data were collapsed across behaviors within each of the 4 HF skills, computing the percentage of negative behaviors associated with the 2 crew positions, the 2 sessions (EPE and CO-3), and experimental condition. This yielded a total of 32 cell percentages and 16 (experimental vs. control) comparisons. Their breakout is depicted in Table 67. Underneath each percentage is the sample size (N) on which the percentages were based. This N corresponds to the number of subjects in that cell, multiplied by the number of possible behaviors (either 6 or 7) tallied. Below that is a row that represents the outcome (in terms of probability) of the statistical test, where each pair of percentages was compared using a binomial testing procedure (Miller & Freund, 1965).

Table 67

Percentage of Negative Behaviors by Crew Position, Condition, and Session

Session:	**CO-3**		**EPE**		**CO-3**		**EPE**	
Condition:	**Control**	**Spiral**	**Control**	**Spiral**	**Control**	**Spiral**	**Control**	**Spiral**
Crew Position:	**P**	**P**	**P**	**P**	**SO**	**SO**	**SO**	**SO**
Skill/Behavior								
Avoids Channelized Attention	.100	.032	.092	.063	.022	.059	.029	0
N	120	126	120	126	90	104	104	102
significance (p-level)	**.030**		.408, ns		.205, ns		.084, ns	
Task Prioritization	.083	.048	.092	.071	.022	.049	.029	.010
N	120	126	120	126	90	102	104	102
significance	.256, ns		.562, ns		.323, ns		.322, ns	
Select COA	.083	.048	.092	.071	0	.049	.029	0
N	120	126	120	126	90	102	104	102
significance	.256, ns		.562, ns		**.033**		.084, ns	
Crew Coordination	.086	.048	.086	.048	.019	.041	.024	0
N	140	147	140	147	105	119	126	119
significance	.194, ns		.194, ns		.324		.090, ns	

As with the previous Spiral analyses, significant differences in favor of the experimental group are highlighted in green, whereas those in favor of the control group are highlighted in red. As can be seen in the table, there were one of each, where Pilots from the experimental group had a lower percentage of negative Attention Management behaviors relative to their counterparts from the control group. Conversely, Sensors from the control group had fewer negative behaviors in the Select COA skill compared to their experimental group counterparts.

Beyond these two differences, a closer examination of the table reveals most of the cell differences tended to favor the experimental group, and indeed, their (non-significant) probability values were typically lower, suggestive of an overall lower rate of negative behaviors from students in the experimental group, particularly among Pilots. To assess this pattern, a simple sign test was performed, comparing the number of cell comparisons, out of 16, in favor of the experimental condition. This is a non-parametric test since it does not make any assumptions about the underlying distribution of data. It is a reasonable test to employ when it seems there are consistent differences in favor of one group over another, regardless of the magnitude of those differences.

Of the 16 comparisons, 12 yielded lower negative percentages for the experimental condition. The probability of getting an extreme outcome of this by chance was .038, which was within the range of significance. Thus, it seems reasonable to conclude the probabilities of negative behaviors were lower for students from the experimental condition compared to those from the control condition.

Exploring this finding further, the above analysis was repeated, only this time it included the data from Spiral 1 (the EA intervention) in the comparison. The percentage of negative behavior data is depicted in Figures 6 and 7 for the EPE and CO-3 sessions, respectively. Graphically, the control group percentages tended to be higher than either spiral in both sessions. A Wilcoxon sign test was applied to the combined data from the two sessions that compared Spiral 1 against the control. The test revealed this Spiral 1 reduction was statistically significant (p < .035). This translated into fewer errors or problems by students who received the additional training (EA) compared to those who did not. A second test was then performed, this time comparing Spiral 1 against Spiral 4. Consistent with the graphics, Spiral 4 students exhibited significantly fewer negative behaviors compared to Spiral 1 (also p < .035). Thus, there was benefit to adding the other three training interventions (ICH, MTT, GTT) to create Spiral 4, as it resulted in fewer negative behaviors compared to Spiral 1, when only EA was provided as the training intervention.

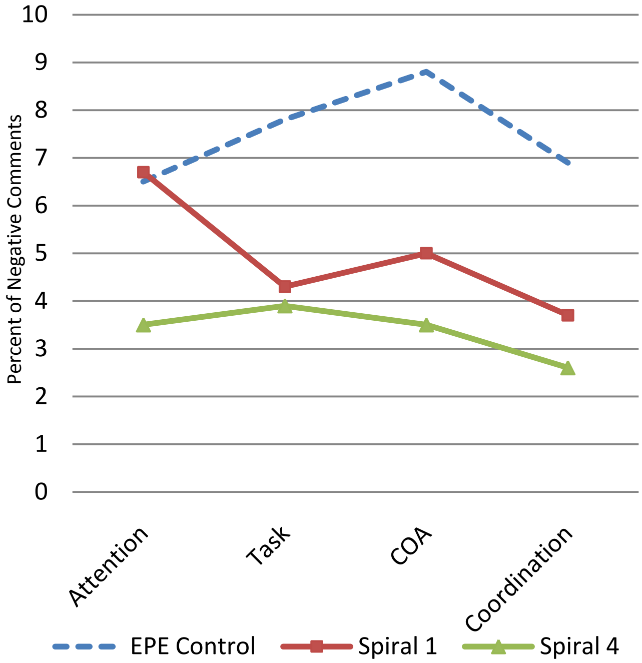


Figure 6. Percent of Negative Comments during the EPE Session

There was considerable support for a significant benefit of the spiral treatments when looking at negative behaviors (as provided by instructor comments) as Level III data. This was most evident when the behavioral data were "rolled up" within a HF skill and the composite percentages were compared between conditions. Using a simple sign test, a higher number of cell comparisons were in favor of Spiral 4 relative to Control 4. In addition, the magnitude of the percentage differences tended to be in favor of the Spiral condition. This tendency held up with the Wilcoxon sign test (which rank orders the percentages), where Spiral 4 not only had significantly fewer negative behaviors compared to the Control, but also had fewer negative behaviors compared to Spiral 1, which in turn exhibited better performance than the Control. This ordering of Control, then Spiral 1, followed by Spiral 4 was an important

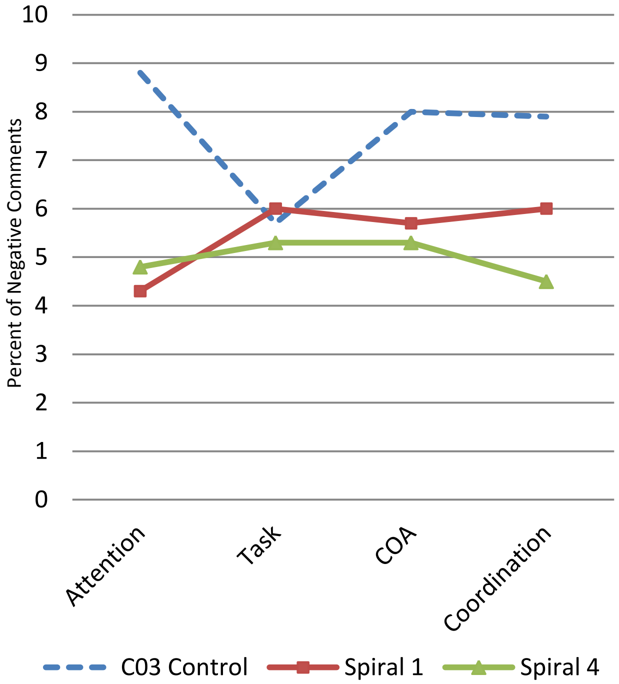


Figure 7. Percent of Students Receiving Negative Comments during the CO-3 Session

finding as it indicated not only were the treatment interventions having an impact on Level III behavior, but there was added benefit from including the three interventions beyond the EA in Spiral 1. That is, beyond EA, there was still room for additional improvement upon receiving the other three treatments (particularly GTT).

Analysis of Student Gradesheets.
In the third analysis of the Level III data for the Spiral 4 comparisons, the dependent measures were selected training item ratings from the Creech training gradesheets during the third CO-3 session and the S-EPE-2 session. (Prior to Class 09-04, the analysis focused on SIM-14 as the final simulator evaluation session, but starting with 09-05, SIM-14 was split into two sessions: S-EP-1 and S-EP-2. S-EP-2 was covered since that was the session where Pilots and Sensors demonstrated their emergency procedure handling skills.) Recall Spiral 4 consisted of all four training interventions: EA, ICH, MTT, and GTT. The classes covered under this comparison were 10-02 and 10-04 for Control 4 and 10-03 and 10-06 for Spiral 4.

The training items selected for analysis were ones CTI SME's believed would encompass the skills most likely to benefit from CRM training. These items are displayed in the left column of Table 11, where the corresponding CRM/HF skill areas are presented in the right column. Each item was scored on a 0-4 scale, where most grades were either a "2" or a "3." Students also received an overall grade for the session. Students receiving a "1" on several training items were usually required to take an extra ride (X-ride). Unfortunately, 2 training items that should be particularly sensitive to CRM training, *Mission Checks* and *Tactical Communications*, were deleted from the Pilot's CO-3 gradesheet starting with Class 09-05.

Combining the data from two classes yielded a respectable sample size (N ~ 20) for comparing Spiral 4 and Control 4 performance. With training records in electronic form (starting with Class 09-05), the grades were entered for the training items identified in Table 11 into an Excel spreadsheet, creating a separate tab for Control 4 and Spiral 4. The statistics resident within Excel were used to calculate the means, variances, and N necessary to perform the required analyses. To make the statistical comparisons, the same method was used as in the first look report: 1) the within-group variances were pooled from the conditions being compared in order to formulate a t-test (Hayes, 1973); and 2) a conservative Bonferroni criterion was used to control Type I error inflation due to making multiple tests (Harris, 1994). Thus, trying to achieve an experiment-wise alpha level of .05 required a fairly stringent value of .001 (i.e., .05/50 tests) for each individual test.

As mentioned in the third-look analyses for Spirals 1-3, there was considerable "noise" in the data due to having multiple instructors (presumably with different internal criteria) assigning grades across students and classes. This created a larger within-group error variance, making it harder to achieve statistical significance. In addition, there were non-trivial "cohort" effects, in which some classes were simply better or weaker than others because of the makeup of their students (e.g., experience, aviation background, stronger class leader). This, too, made it harder to discern differences due to added noise variance. There is no solution for either problem, since these are a fact of life in doing field research within an operational training squadron. It was for these reasons multiple dependent measures were employed for Level III analysis so there was "triangulation" on the locus of real effects in the data.

Tables 68-71 present the gradesheet data for the CO-3 sessions (Pilots), S-EP-2 sessions (Pilots), CO-3 sessions (Sensors), and S-EP-2 sessions (Sensors), respectively. For each table, the columns correspond to the various training items whereas the rows provide the means, variances, and sample sizes (N) for the Control 4 and Spiral 4 conditions. The bottom row in the table provides the results of the t-test comparisons for the two means, where the resulting t-value, degrees of freedom, and probability value are depicted. To permit the mean comparisons to stand out more, the cells containing the variance and N statistics are portrayed in light-gray font. Using the same convention adopted in the other analyses, any statistically significant differences in favor of the treatment (Spiral 4) are highlighted in green, whereas the opposite findings (Control 4 superior to Spiral 4) are highlighted in red.

Looking first at Table 68, there were no significant differences in the CO-3 session for Pilots. This was evident from examining the bottom row, where none of the probabilities approached the .001 level, in either direction (i.e., experimental higher than control or the reverse). The lowest probability value observed was for the ATC communication item, where the experimental students had a somewhat higher rating relative to the controls (t=1.447, df=36, p<.156). None of the other mean differences were even close to this probability value. Thus, there was virtually no evidence for any effect within this table for the CO-3 sessions with Pilots.

Table 68

Gradesheet Data and Analysis Results for CO-3 Session, Pilots

Condition/ Statistic	**1. Mission Planning**	**2. Admin Checks**	**8. Air-manship**	**12. ATC Comm**	**17. CRM**	**18. ORM**	**19. Flight Discipline**	**21. EPs**	**Overall Grade**
Control 4 (mean)	2.35	2.25	2.30	2.30	2.40	2.20	2.20	2.20	2.10
Control 4 (VAR)	0.24	0.20	0.22	0.22	0.22	0.17	0.17	0.17	0.09
Control 4 (N)	20	20	20	20	20	20	20	20	20
Spiral 4 (mean)	2.44	2.22	2.33	2.11	2.28	2.17	2.11	2.06	2.11
Spiral 4 (VAR)	0.38	0.18	0.24	0.10	0.21	0.15	0.10	0.06	0.22
Spiral 4 (N)	18	18	18	18	18	18	18	18	18
t-value, df, p-level	t =.5008, df = 36, p < .620	t =.2116, df = 36, p < .834	t=.1928, df=36, p<.848	t=-1.447, df=36, p<.156	t=.7963, df=36, p<.431	t=-.2306, df=36, p<.819	t=-.7491, df=36, p<.459	t=1.255, df=36, p<.937	t=.0781, df=36, P<.937

Table 69 presents the Pilots' data for the other session of interest, S-EP-2. Most of these means were close in value and, hence, there were no significant differences. Interestingly, there was 1 comparison, between experimental and control group students on the Airmanship item, where the probability for the t-test reached .02, which was fairly notable. This difference was in favor of the experimental group. None of the other differences were even close to this probability value.

Table 69

Gradesheet Data and Analysis Results for S-EP-2 Session, Pilots

Condition/ Statistic	**1. Mission Planning**	**2. Admin Checks**	**8. Air-manship**	**10. ATC Comm**	**15. CRM**	**16. ORM**	**17. Flight Discipline**	**19. EPs**	**Overall Grade**
Control 4 (mean)	2.21	2.05	2.00	2.00	2.16	2.00	2.00	2.00	2.00
Control 4 (VAR)	0.18	0.05	0	0	0.14	0	0	0	0
Control 4 (N)	19	19	19	19	19	19	19	19	19
Spiral 4 (mean)	2.20	2.05	2.25	2.05	2.10	2.10	2.10	1.95	1.95
Spiral 4 (VAR)	0.17	0.05	0.20	0.05	0.20	0.09	0.20	0.05	0.16

Condition/ Statistic	1. Mission Planning	2. Admin Checks	8. Air-manship	10. ATC Comm	15. CRM	16. ORM	17. Flight Discipline	19. EPs	Overall Grade
Spiral 4 (N)	20	20	20	20	20	20	20	20	20
t-value, df, p-level	t=-.0747, df=37, p<.941	t=0, df=37, p<1	t=2.436, df=37, p<.020	t=.9723, df=37, p<.337	t=-.4534, df=37, p<.653	t=1.452, df=37, p<.155	t=.9745, df=37, p<.336	t=.9723, df=37, p<.337	t=-.5445, df=37, p<.589

Table 70 presents the Sensor data for the CO-3 session. Because the within group variability was rather high, none of these differences reached significance either. However, 2 of the items, Airmanship and ORM/Safety, exhibited a notable difference, though non-significant. In both cases, the mean difference was in favor of the control condition, where the significance levels were quite a bit higher (.073 for Airmanship and .004 for ORM) for several of the training items relative to the others.

Table 70

Gradesheet Data and Analysis Results for CO-3 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 4 (mean)	2.38	2.25	2.31	2.50	2.38	2.38	2.19	2.22
Control 4 (VAR)	0.25	0.20	0.36	0.40	0.25	0.25	0.30	0.18
Control 4 (N)	16	16	16	16	16	16	16	18
Spiral 4 (mean)	2.41	2.06	2.24	2.24	2.00	2.18	2.00	2.12
Spiral 4 (VAR)	0.26	0.06	0.19	0.19	0	0.15	0	0.11
Spiral 4 (N)	17	17	17	17	17	17	17	17
t-value, df, p-level	t=.1705, df=31, p<.866	t=1.527, df=31, p<.137	t=-.3851, df=31, p<.073	t=-1.383, df=31, p<.177	t=-3.137, df=31, p<.004	t=-1.29, df=31, p<.207	t=-1.431, df=31, p<.162	t=-.774, df=33, p<.445

The final analysis, for Sensors in the S-EP-2 session, is depicted in Table 71. Within-group variability was considerably lower in the sim sessions, and consequently, several of the differences approached significance. Unfortunately, both were in the opposite direction to that predicted, as Control 4 students received higher ratings on Airmanship and Flight Discipline. As seen from the table, the other differences were fairly close, though they, too, tended to favor Control 4.

Table 71

Gradesheet Data and Analysis Results for S-EP-2 Session, Sensors

Condition/ Statistic	1. Mission Planning	2. Admin Checks	10. Air-manship	14. CRM	15. ORM	16. Flight Discipline	19. EP	Overall Grade
Control 4 (mean)	2.21	2.16	2.42	2.42	2.17	2.37	2.21	2.16
Control 4 (VAR)	0.29	0.14	0.37	0.26	0.15	0.25	0.29	0.25
Control 4 (N)	19	19	19	19	18	19	19	19
Spiral 4 (mean)	2.00	2.06	2.06	2.29	2.12	2.06	2.00	2.00
Spiral 4 (VAR)	0.13	0.06	0.06	0.22	0.11	0.06	0	0
Spiral 4 (N)	17	17	17	17	17	17	17	17
t-value, df, p-level	t=-1.356, df=34, p<.184	t=-.9365, df=34, p<.356	t=-2.279, df=34, p<.029	t=-.7929, df=34, p<.433	t=-.409, df=33, p<.685	t=-2.317, df=34, p<.027	t=-1.604, df=34 p<.118	t=-1.317, df=34, p<.197

The extremely positive reactions of students to the GemaSim team trainer were quite encouraging and underscored the success of the three class implementations. Hence, any future revisions should be more "fine-tuning" rather than anything else. Obtaining average reactions on many Likert items statistically equivalent to a maximum rating of "5" was indeed remarkable; it will be impossible, of course, to improve on that. Looking at the comments, it was evident the students very much liked the hands-on aspects of the training, the opportunity to work as a team, and that the non-aircraft, space mission orientation of GemaSim was fun and engaging. Judging by the students' comments and reactions it was highly likely the core CRM principles of interest – attention management, task prioritization, decision-making, crew coordination – were all tapped in high degrees during the training. Just from casual observation, and from listening to the debriefs, there appeared to be a number of areas where students gained CRM-relevant insights based on their experience with GTT.

In terms of recommendations for future changes to GTT, there really is not much that should be altered. If possible, it would be ideal if no student had to serve as an observer since there is (probably) less instructional benefit from that position given its passive and non-interactive nature. Also, there were some notable problems with several of the book-task questions that came with the system, where students (rightly) complained about the confusing nature of the item. Some of the more egregious questions were eliminated from use after the beta-test (Class 10-03). However, other questions should be scrutinized for accuracy and wording, and modifications made prior to future system use. Finally, several students requested there be additional instruction concerning the system specifics (e.g., battery charging) to ensure they are up to speed before the first mission gets underway. It is likely these issues can be addressed with some minor changes to the spin-up CBT that students experience on the Web before the training session begins.

It was quite rewarding to have observed and participated in a training intervention so well-received and highly-lauded by an entire class of students. Seeing a unanimous positive reaction to a classroom training exercise, of any sort, is unprecedented. Watching the students interact with the training simulation, and seeing both how much fun they had and how much they appeared to learn, it was quite evident this intervention is a "winner," with enormous potential for supporting CRM training in the future.

Conclusions and Recommendations.

The Level II analysis presented a fairly clear and consistent picture of student learning during GTT. Whether measuring the improvement in performance between the first and second missions in terms of game score, SME ratings, or observed CRM behaviors, the statistical evidence was strongly positive. This learning was also reflected in the SME comments on the data sheets, where improvements in key behaviors were noted in all the CRM categories. Students were quite specific on their critique sheets about what they learned as they were highly stressed and task-loaded in the first mission. In general, the GTT environment created multiple opportunities for students to experience the consequences of poor teamwork. as well as the benefits that accrue when members work effectively as a team.

Based on the learning data, there is little to suggest any tangible ways to improve the game experience in subsequent implementations. As noted previously, it would be ideal to give students a third mission, since there was clearly still room for improvement in many of their CRM behaviors. However, the compressed training schedule the students now have makes it impractical to expand the training time for this intervention. Based on the results of this analysis and the Level I analysis, the GTT experience was very successful. Not only did students like the training a great deal, they also learned important lessons about the importance of teamwork in the context of a game that, while not emulating Predator functionality, stimulates many of the cognitive and behavioral functions required in a tactical aviation setting.

The Level III results of the training gradesheet analysis for the Spiral 4 comparisons provided limited evidence for the effect of the training interventions. As in the other Spiral analyses, the Creech gradesheet data were beset with large within-group variability, making it difficult to discern any significant differences. The evidence presented in this analysis could best be scored as a "push for the control group," as there were more differences that favored Control 4 relative to Spiral 4. However, none were statistically significant against the conservative .001 level. Besides the large error variability in the data, there also appeared to be notable cohort effects across classes, making identification of any true effect very difficult, if not impossible, to determine. Fortunately, there were other dependent measures, particularly the negative behavior index described in the second-look reports, which provided a more sensitive index of training intervention impact.

Level IV (Operational Impact) Analysis

Analysis of Electronic Survey

This analysis provides the results of the Level IV (organizational impact) survey administered during June-July 2010. The purpose of the survey is to assess the collective impact of the four training interventions on the students' CRM skill proficiency, from the standpoint of the gaining units. That is, when asked to compare the CRM skills of their students currently with those from more than 18 months ago (the duration of the study), would qualified instructors with first-hand familiarity with student performance rate the current period higher than the previous one? If they do, that will be one piece of evidence suggesting the training was having a positive impact on the gaining unit at an organizational level. However, since half of the subjects participating in the study did not receive the interventions, the gaining units must be probed further, by querying knowledgeable instructors regarding the CRM skills of each study participant by name. While such queries on an individual basis are not, strictly speaking, typically part of a Level IV analysis, it is necessary in this case to determine that, if an organizational impact is identified, it not be strictly due to the control subjects' superior performance.

With these considerations in mind, the survey consisted of 3 parts. The first part asked the raters for some demographic background concerning their crew position, flight hours, and type of aircraft flown. The second part consisted of 5 questions asking for relative ratings of CRM skill proficiency for students in the past 18 months (the study period) compared to the previous periods. Four- or 5-point rating scales were used in each case, where the questions differed in the type of "reflection" (general assessment, CRM-specific training, review of organizational records) requested of the rater. The third part asked the survey respondent to rate individual students, listed by name, on a 10-point scale (1 = lowest and 10 = highest), on each of the 4 CRM skills covered by the training: attention management, task prioritization, course of action (COA) selection, and crew coordination. Only students presently at the gaining unit were named on the survey, where raters were "blind" to the condition (treatment vs control) the students were in. If the instructor did not have knowledge of a given student's skill level, he/she was to indicate "do not know" (DNK).

Because of the present high OPTEMP for RPA operations, as well as high levels of Air Force-wide "survey fatigue," the survey was administered electronically on a not-to-interfere basis. Specifically, the Zoomerang Web-based survey service was used to administer the Level IV survey. Gaining units were identified via e-mail, and letters of introduction and explanation of the survey were sent to the commanders of each unit. As an electronic survey capability, Zoomerang provides interactive branching so respondents only saw the students who were getting mission-qualified at their unit (rather than the list of all 481 students in the study); this made each survey no more than two to three pages for each respondent. Unfortunately, the short timeframe of the survey period and the need to have a "small footprint" on the various gaining units limited the ability to collect very many data during this administration. In particular, data were only collected from two units, the California Air National Guard (ANG) and the Texas ANG, with six respondents total, one from the TX ANG and five from the CA ANG. Their results are described below.

Rater Demographics.

All 6 raters were experienced Pilots, with half having 4 to 8 years of experience and the other half with more than 8 years of experience. One rater was from the TX ANG; the other 5 were from the CA ANG. For Predator hours, 1 rater had 51-100 hours, 1 had 1001-1500 hours, and 4 had more than 1500 hours. Prior aircraft experience was mixed, with 1 indicating fighter, 2 citing transport/airlift, 1 stating helos, 1 RPA, and the sixth rater listing Boeing commercial aircraft (707, 727, 737, 777) experience.

Survey Results – Overall CRM Skill.
The five survey questions concerning CRM skill in general were slanted differently to give the raters a variety of bases to reference when making their ratings. Table 72 presents the survey questions, their rating scales, distribution of responses, and a concluding interpretation.

Table 72

Summary of Overall CRM Skill Results

Question	Rating Scale	Distribution of Responses	Interpretation
CRM training in 11RS in the past 18 months is better than before.	1 - strongly disagree 2 - disagree 3 - neutral 4 - agree 5 - strongly agree 6 - DNK[a]	2 DNK 3 neutral 1 strongly agree	Net +
On average, 11RS graduates over the past 18 months have significantly better CRM skills than previous graduates.	1 - strongly disagree 2 - disagree 3 - neutral 4 - agree 5 - strongly agree 6 - DNK	1 DNK 4 neutral 1 strongly agree	Net +
How well do you think recent MQ-1 FTU graduates in the last 18 months are trained with respect to CRM and their ability to perform on an RPA team?	1 - poor 2 - fair 3 - good 4 - excellent 5 - DNK	1 fair 3 good 2 excellent	Net ++
Based on your knowledge of unit training records and stan/eval data, how would you characterize the CRM training program's impact on your flying program?	1 - very negative 2 - negative 3 - neutral 4 - positive 5 - very positive 6 - DNK	3 neutral 2 positive 1 excellent	Net ++
From your experience with MQ-1 FTU graduates, how would you characterize the frequency/severity of "problem" students over the past year?	1 - marked decline 2 - decline 3 - neutral 4 - increase 5 - marked increase 6 - DNK	1 increase 4 neutral 1 marked decline	Net +

[a] DNK – does not know

With six responses, there are simply not enough data to perform statistics on the response distributions, though it is possible to extract some qualitative trends. As can be seen from the distribution of responses in the third column, the six raters made mostly neutral or DNK responses, with a few exceptions in each question. Importantly, the exceptions were always positive. In the interpretation column, the interpretation of CRM training quality in the past 18 months (during the period of the study) was either slightly positive (Net +) or somewhat more positive (Net ++). Thus, for all five overall CRM training questions, the net response was a positive for students entering mission qualification training during the current study period.

Survey Results – CRM Skill Ratings by Individual Students.
The remaining questions on the survey queried respondents concerning the CRM proficiency level, on a 10-point scale of the named students at their unit, for each of the 4 human factors skills of interest. Because of a low survey return rate, only 19 named students received ratings, with 10 coming from the experimental condition and 9 from the control. This is only 4% of the 481 students who participated in the study from classes 08-16 through 10-06. Nonetheless, it is enough to perform t-tests comparing the average ratings between the 2 conditions for each of the 4 human factors skills, but because the sample size is low, it was decided not to further subdivide them into Pilot and Sensor for a more detailed comparison.

Table 73 presents the ratings obtained for each named student with whom the raters were familiar. Students presently at the 2 ANG units who received DNK responses from the raters were excluded. Furthermore, note the ratings in italics, those for students 5/6-10 for the experimental condition and 7-9 for the control, were based on the median of the ratings provided by the 5 CA ANG raters. As can be seen, there is a consistent superiority of the mean ratings for the control students, on the order of about .3-.5 scale unit. However, statistical testing using independent t-tests (Harris, 1994) revealed none of the differences were significant. In fact, the t-values were fairly small, indicative of no meaningful differences. Examination of the median values revealed they were basically the same between the conditions, suggesting the distributions were skewed (non-normal). Looking further, the mean differences were typically the result of one outlier value, a value eliminated when the median (rather than the mean) is calculated. In short, there appears to be equivalence in the named-student ratings on each of the four human factors skills.

Table 73

Distribution of CRM Proficiency Ratings for Experimental and Control Students

	Attention Management		**Task Prioritization**		**COA Selection**		**Crew Coordination**	
Student	**EXP**	**CON**	**EXP**	**CON**	**EXP**	**CON**	**EXP**	**CON**
1	9	10	9	10	10	10	8	10
2	8	10	9	10	10	10	9	9
3	4	8	6	9	5	9	7	8
4	8	8	*9*	9	*3.5*	9	*9*	8
5	8	7	*4*	6	*9*	8	*4*	8
6	*8*	5	*8*	*5*	*7*	5	*9*	8
7	*3*	*4*	*7*	*8*	*7*	*5*	*8*	*6*
8	*9*	*8*	*9*	*7*	*9*	*8*	*6.5*	*8.5*
9	*8*	*6.5*				*7*		*7*
10	*6*							
Mean	7.1	7.4	7.6	8.0	7.6	7.9	7.6	8.3
Median	8.0	8.0	8.5	8.5	8.0	8.0	8.0	8.5
t statistic	.306		.406		.314		1.00	
degrees of freedom	17		14		15		15	
probability	.763		.691		.758		.332	
significance	ns		ns		ns		ns	

Note. All numbers in italics were calculated as the median of the multiple raters from the TX ANG.

Analysis of Archival Information

This analysis provides the results from two sources of archival information used to gauge the impact of the four CRM training interventions on organizational performance. The impact of an intervention on the organization receiving that intervention constitutes a Level IV analysis within the four-tier model of Kirkpatrick (1996, pp. 54-59). That is, does the organization which receives the students trained with the interventions in question perform better since those students have entered? The results of a Level IV analysis in which the source data were survey responses from instructors and officers of the gaining unit were summarized. To a certain extent, these data can be viewed as *subjective* since they are based on the opinions – though well-informed – of stakeholders from the gaining unit. In this analysis, the results are reported from an analysis of *objective* data, that is, hard empirical data routinely collected and which do not depend on any person’s opinion for their observation.

Two measures are analyzed in this report. The first are instances of Hazardous Air Traffic Reports (HATRs). Presently, these mostly occur in theater since the majority of a wing’s flying takes place there. The focus was on those HATRs where the MQ-1 aircraft was at fault, as opposed to other non-Predator causes, such as air traffic control (ATC), another aircraft, or some deficient procedure. HATRs are a good measure to look at since 1) they occur with some frequency; 2) are clearly reflected in Pilot/Sensor Operator proficiency; and 3) their impact on the organization is important, as they can be a harbinger of even more serious events.

The second measure is the time it took IQT graduates to become mission ready (MR). That is, do students who have received the CRM training interventions exhibit faster rates (as measured by number of days or number of sorties) of becoming mission qualified compared to students who did not receive such training? Faster rates of becoming MR would be a clear benefit for the gaining organization; hence, it is a potentially useful Level IV measure.

Logic of the Level IV Analysis.
At the outset, note it is very difficult to conduct a "clean," controlled Level IV analysis for several reasons. First, there are many factors besides the intervention itself that typically occur at the same time as students join a gaining unit and which can either mask or confound the effects of the intervention. In this case, several factors conjoined on the study, notably syllabus changes and increases in OPTEMPO (e.g., increases in number of orbits flown), that make it nigh impossible to attribute any changes in the measured data solely to the interventions. Second, an intervention's organizational impact does not occur instantaneously, but rather, accrues gradually over time. Since the observation window for the Level IV analysis is quite restricted, covering only a few months after the last class received the interventions, there was not much time for the objective measures to "reveal themselves." In short, finding a significant, immediate effect in a Level IV analysis is a tall order, so one's expectations must be adjusted accordingly.

The basic design logic of the Level IV analysis is an interrupted time series (Glass, 1997). This is depicted conceptually in Figure 8. The measures of interest, be they HATRs, time to become MR, or whatever, are collected continuously over time; that is what makes it a time series. The "interruption" is the intervention, where the organization is presumably different after the interruption occurs. Thus, if the graph of the dependent variable (i.e., the time series) exhibits an abrupt shift after the interruption, then that effect might be attributed to the intervention. However, the abrupt shift is difficult to achieve in practice since the intervention itself tends to be introduced gradually. Indeed, the impact has been depicted as being gradual in Figure 8, since if there is an impact, it will not be all or none (like a light switch). In particular, as more intervention-trained students enter the gaining unit, that unit's organizational performance should improve gradually over time, reflecting the increasing proportion of its Pilots/Sensors as having received the CRM training. This is all theoretical, of course, and it is difficult to pinpoint these effects solely on the intervention as was mentioned above.

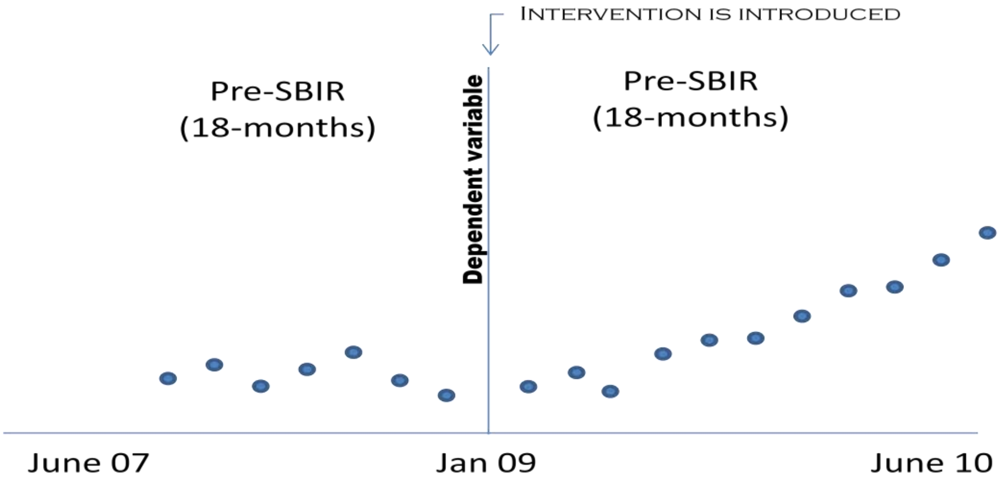


Figure 8. Hypothetical depiction of the intervention's impact on organizational performance

The design logic depicted above implies the data collection should be treated as a "before" and "after" comparison. In the present case, this means comparing the measures of interest obtained BEFORE the intervention with a comparable span of time AFTER the intervention was introduced. Considering the two measures of interest, this would correspond to comparing the number of HATRs that occur in the 18-month period before the intervention was introduced (which is benchmarked as Jan 09) with those that occur in the corresponding 18-month period after that date. Similarly, the average time required to become MR for students in the 18-month period before January 09 would be computed with the corresponding average for students who become MR after that date.

In each case, of course, there are many factors that can be confounded with the introduction of the intervention. For time to MR, a complicating factor will be the training syllabus in place. Specifically, if the syllabus is changed to require more flights to achieve MR, then naturally that figure will work to a disadvantage (i.e., in terms of seeing a positive impact of the interventions). While the average time to achieve MR might still be shorter because there are fewer students who need extra rides, it will be necessary to "travel uphill" to see an effect. Similarly, for HATRs, the operating tempo will have a major impact. In particular, as flight hours per month increase, the frequency of HATR occurrence will naturally increase, all other things being equal. Moreover, as combat operations expand in theater, there will be a great preponderance of threats during airspace maneuvers, which are also likely to increase HATR frequency. In sum, the ability to infer a positive impact in a Level IV analysis is not easy, particularly with the short time period the interventions were available to "transform" students before they join their flying wing. With these caveats in mind, the methods used to collect the Level IV measures and the results obtained will be examined.

Method.
All the MQ-relevant HATRs that occurred in Afghanistan and Iraq between 2005 and June 2010 were compiled. The analysis was restricted to the June 2007 to June 2010 period indicated in Figure 8. The "mishap one-liners" were reviewed in each case, and a "count" was registered if there was a potential for the problem to have been Pilot-related. If the cause was clearly an ATC or other aircraft infraction, those were not counted. The review was done independent of date to ensure impartiality of inclusion in the pre-SBIR bin (June 2007 to December 2008) versus the SBIR bin (January 2009 to June 2010).

A second source of HATR data was obtained from the 432nd Wing's own, most recent quarterly flight safety report. This report tracks HATRs, along with other indices (e.g., Class A-E mishaps), as a way to gauge their overall flight safety trends. Though based on fewer data than the in-theater HATR reports, these trends nonetheless provide a more focused measure since they only include mishaps attributable to the wing itself.

For the other measure, a SBIR research staff member went to Creech AFB and, coordinating through the operations group commander, obtained access to 110 gradebooks from the 15th Reconnaissance Squadron (15RS). These gradebooks contain the start and end dates of students' mission qualification training (MQT) experience to become MR. For this analysis, only duration of upgrade training and number of flight sorties were recorded, as student identity (other than crew position) was not recorded. Students whose training experience started before January 09 were categorized as "pre-SBIR" while those who matriculated after January 09 were placed in the "SBIR" category. Comparable numbers of Pilots and Sensors were included in the sample. Training duration was defined as the number of calendar days to become MR, including weekends, holidays, and stand-down days in the total. Thus, it encompassed the first MQT activity to the final checkride.

Measure 1 – HATRs.
For the first analysis, the number of Operator-related HATRs in the pre-SBIR (June 07 – Dec 08) and SBIR periods (Jan 09 – June 10) were tallied separately. The tallies were essentially identical, with 20 from the pre-SBIR period and 21 for the SBIR period. Although the actual rates might vary somewhat if transformed as a frequency per hours flown, it is doubtful such a transformation would change the basic finding of no difference. Looking at the reasons for the HATR, there was also not much difference

between the two reporting periods. Thus, the most common reasons include failing to notice airspace violations, operating without clearance, or violating altitude separation restrictions.

The latest trend analysis report from the 432nd Wing, showed a major program was implemented to reduce hazard air traffic reporting. In particular, the Wing reported a dramatic decrease from FY09 (which is a mix of pre-SBIR and SBIR period) to FY10 in the number of HATRs. The Wing's program, which placed greater emphasis on mid-air collision avoidance (MACA) and HATR avoidance, was obviously confounded with the CRM training interventions. Given the strong emphasis the Wing has placed on this program, it was difficult to attribute the reduction to anything other than their prevention program. Hence, it is likely any effects of CRM training intervention on organizational performance have yet to surface.

Measure 2 – Time to become MR.
Time to become fully mission qualified after arrival in gaining units was originally included in the plan to assess impacts of the training interventions on the larger organization. Unfortunately, several changes in Air Force practices occurred during this study that made such comparisons difficult to interpret. First, at least four upgrade syllabi were used in the 15RS during the subject timeframe, with different numbers of upgrade sorties in at least three of them. Second, there were several different IQT syllabi used in the 11RS during this time. Third, and most important, the pre-SBIR students were all rated with at least one previous operational tour and many Sensors had previous flight experience. The SBIR Pilots were mostly recent SUPT graduates, plus some non-rated students, and the Sensors were predominantly right out of initial technical training. Thus, the two populations were quite different, with experience favoring the pre-SBIR period. As a result, these measures were not used in the comparisons.

Conclusions.
Although the survey returns were modest in number, the trends observed were consistent with a moderately positive interpretation of present CRM training by the participating instructors. This was seen in all five CRM training-related questions posed to the survey respondents, where several questions (concerning training impact on their program) resulted in fairly positive assessments. Comparisons of CRM ratings for students identified by name revealed no differences between control and experimental students, indicating that the positive assessment could not be attributed to instructors having experience with only the control subjects. While it would be best to collect additional survey data, the picture thus far is a positive one concerning the impact of the CRM training on gaining units. A follow-up survey should be conducted~~,~~ after a few months, so the observation window is longer, allowing more time for gaining unit instructors to have experience with a much larger portion of the study sample. But thus far, the picture painted by the admittedly very limited Level IV survey is positive with respect to CRM training impact.

As noted at the outset, performing a Level IV analysis is quite difficult due to the large number of factors that can intervene, confound, and/or obscure the impact of any training intervention. The intended objective measures examined in this report (number of Predator HATRs and time for IQT graduates to achieve MR upgrade status in their gaining units) did not exhibit the desired superiority of the SBIR period (Jan 09 – June 10) compared to the corresponding pre-SBIR period (June 07 – Dec 08). While objective, these were very "noisy" measures since they were subject to wide fluctuations due to factors beyond control, and which had nothing to do with the type of training students were receiving. Consequently, despite the lack of statistical effects in the direction desired – better organizational performance in the SBIR period relative to the pre-SBIR period – there were clearly extenuating circumstances. Chief among these were the turbulence in the IQT and MQT syllabi during the period, as well as the greater levels of experience of students during the pre-SBIR period. Moreover, even the slight improvement in the HATR incidence for the 432nd Wing could not be attributed to the CRM training interventions since the Wing had in place an emphatic program intended to reduce HATRs and other mishaps. While it is preferable to think the enhanced CRM training during the SBIR period facilitated these programmatic emphases, no measures were in place with the level of resolution to support such an interpretation.

In closing, the case for a Level IV impact of the CRM training program has yet to be made. The data collection, due to the time constraints of this project, occurred very early in the career timelines of students. It would be prudent to repeat this analysis in a year or so, at which time all the graduating CRM students will have been in their gaining units for a sufficient duration to see an impact. Also a comparison between experimental vs. control students in the time to become MR should be attempted. Nevertheless, as seen in survey analysis, the opinions of the gaining unit decision-makers and instructors were certainly encouraging for the potential of the enhanced CRM training to yield tangible performance dividends within their organization.

SUMMARY AND RECOMMENDATIONS

In the bigger picture, looking across all levels of analysis, there was consistently positive student reaction (Level I) to the new types of training as well as consistent evidence of learning (Level II). Students preferred the GTT the most followed by EA.

Preference for ICH and MTT varied by crew position with Pilots preferring the MTT and Sensor Operators preferring ICH. Sensor Operators, generally lacking in aviation-related knowledge and experience, but thrust in a critical crew position, found the ICH an easy method to look at all the factors involved in a mishap and how it relates to doing their job. Pilots with operational flying experience have this knowledge from safety briefings and presentations and would naturally find this method repetitive to what they previously have accomplished. MTT training was more useful for the Pilots as it practiced those skills they would need to directly operate the Predator, more so than the Sensor Operators who can only operate Predator mission systems.

While there were sometimes inconsistent results for Level III effects for specific interventions, overall there was a fairly solid set of data to support a positive impact of both EA and the full combination of all four interventions. Overall, across 540 students, those that had the training interventions had fewer negative evaluation comments during the evaluated missions. They showed an improvement in documented human factors skills. This should translate into fewer errors, better teamwork, reduced mishaps and more effective missions over time. Although there were several confounding variables that affected Level IV data, there were some positive Level IV effects at the unit level through opinions of the gaining command. This is encouraging for the potential of the enhanced CRM training to yield tangible performance dividends within their organization, as Predator crewmembers would arrive at their unit better prepared to fly more effective combat missions than their predecessors. Over a longer period of time, this could translate into a reduction of the Predator mishap rate, although it will take several years before the data are available for such a comparison. Training effects on mishap data in a growing weapons system often take years before any noticeable trend is evident and was not possible given this short-term investigation.

Another observation from the study was the effectiveness of the training for Pilots and Sensor Operators. A top-level look at the data across all training interventions shows there were more positive training effects for Sensor Operators compared to Pilots. This would correlate directly with the general aviation experience level of each aircrew member. The current USAF approach is a one-size-fits-all approach for Cockpit Resource Management training when the results clearly show Sensor Operators would benefit more from a unique training program than Pilots. Level I feedback comments from Pilots attending mandatory initial CRM training indicate the training is generally repetitive from what they have already received and they would prefer shorter, more Predator-specific training.

These results raise several new questions:

The data showed the most effective intervention was the culmination of all four training interventions. Can GemaSim training work as well when implemented as a stand-alone intervention? It could provide some significant benefits to improve CRM training and ultimately aircrew performance.

Should existing USAF CRM training be modified? Based on feedback from EA, consideration should be made for the USAF to modify the initial CRM academics to include immersive multi-media interactive case studies and introduce sixth generation CRM training as part of the initial CRM training in the Predator. Early positive feedback from EA was incorporated in the 2010/2011 Predator CRM Continuation Training course given to operational aircrews. Given the positive response for interactive case histories, there could be some benefit for use as continuation training or continuing education programs similar to those in other professional fields (legal, medical, etc.). There appears to be some merit in using the Multi-Task Trainer to enhance aircrew skill, although its value may be limited, perhaps as part of a CBT "spin-up," prior to receiving GemaSim team training.

Should the USAF explore the use of the GTT for small crew-based weapons systems as part of initial training? GemaSim provided and encouraged the participants to use the CRM skills they learned in their current baseline CRM. The opportunity for pilots and sensor operators to interact directly with one another early in training was viewed favorably by virtually all students. The use of this intervention provided the unique stressful simulation that would force the participants to experience weaknesses and strengths of their CRM skills. The debriefing of the events provided the "capstone of learning" by bridging the gap between learning and experience. Use of the GemaSim trainer for larger crew aircraft or fighter aircraft would need to be validated for effectiveness before implementation.

As the USAF moves to increasingly larger amounts of RPA and crews with significantly less or no aviation experience, relying on traditional training methods may not be the optimal path to produce the most capable Predator or RPA aircrew members. Aircrew manning has often lagged the ability and desire to operate more RPA orbits and non-traditional sources of aircrews are now part of the process to fill RPA requirements. Will this type of training allow Mission Coordinators, non-aviator members of the crew, to be able to perform more effectively? With validated Level III results of new training interventions that improved human performance, these interventions will better prepare our Predator aircrews to maximize their effectiveness and do it safer than was possible in the past. A new training paradigm is ready for the challenges of the future.

REFERENCES

Air Force Safety Center (2009, December). *Aviation statistics: RQ001 UAS mishap history* [Data file]. Retrieved from http://www.afsc.af.mil/shared/media/document/AFD-080114-108.pdf

Anderson, John D. (2004). *Inventing flight: The Wright Brothers and their predecessors*. Baltimore, Maryland: Johns Hopkins University Press.

Arsham, H. (n.d.). *Homogeneity of multi-variances: The Bartlett's test* [Statistical test]. Available from Merrick School of Business, University of Baltimore Web site, http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm

Boersma, P. (2010, October). *Binomial proportions* [Statistical test]. Available from Phonetic Services, University of Amsterdam Web site, http://www.fon.hum.uva.nl/Service/Statistics/Binomial_proportions.html

Campbell, D.T. & Stanley, J.C. (1996). *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation design and analysis issues for field settings*. Chicago: Rand-McNally.

Crew Training International, Inc. (2010). *MQ-1 Predator training from 1 Oct 1999 to 30 May 2010* [Data file]. Retrieved from http://www.cti-crm.com/caf

Elsmore, T. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instruments & Computers, 26,* 421-426.

Glass, G.V. (1997). Interrupted time series quasi-experiments. In R.M. Jaeger (Ed.), *Complementary methods for research in education* (2nd ed., pp 589-608). Washington DC: American Educational Research Association.

Harris, R.J. (1994). *ANOVA: An analysis of variance primer*. Itasca, IL: Peacock Publishers.

Hays, W.L. (1973). *Statistics for the social sciences* (2nd ed). New York: Holt-Rinehart-Winston.

Helmreich, R.L., Merritt, A.C., & Wilhelm, J.A. (1999). The evolution of Crew Resource Management. *International Journal of Aviation Psychology*.

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, NY: McGraw-Hill.

Kirkpatrick, D.L. (1996). Great ideas revised. *Training & Development*, January, 54-59.

Miller, I. & Freund, J.E. (1965). *Probability and statistics for engineers* (pp. 212-214). Englewood Cliffs, NJ: Prentice-Hall.

Nullmeyer, R.T., Herz, R.A., Montijo, G.A., & Leonik, R. (2007). *Birds of Prey: Training solutions to human factors problems*. Proceedings of the 28th Interservice/Industry Training Systems and Education Conference. Orlando, FL.

Nullmeyer, R.T., Spiker, V.A., Montijo, G.A., & Kaiser, D.S. (2008). *Training interventions for reducing flight mishaps*. Proceedings of the 29th Interservice/Industry Training Systems and Education Conference. Orlando, FL.

Nullmeyer, R.T., Stella, D., Montijo, G.A., & Harden, S.W. (2005). *Human factors in Air Force flight mishaps: Implications for change*. Proceedings of the 26th Interservice/Industry Training Systems and Education Conference. Orlando, FL.

O'Toole, K.J., Hughes, T., and Musselman, B.T. (2006) *USAF Aviation Safety: FY 2005 in review*. http://afsafety.af.mil/SEF/Downloads/FY05_Life_Sciences_Safety_Review(AsMA06).ppt

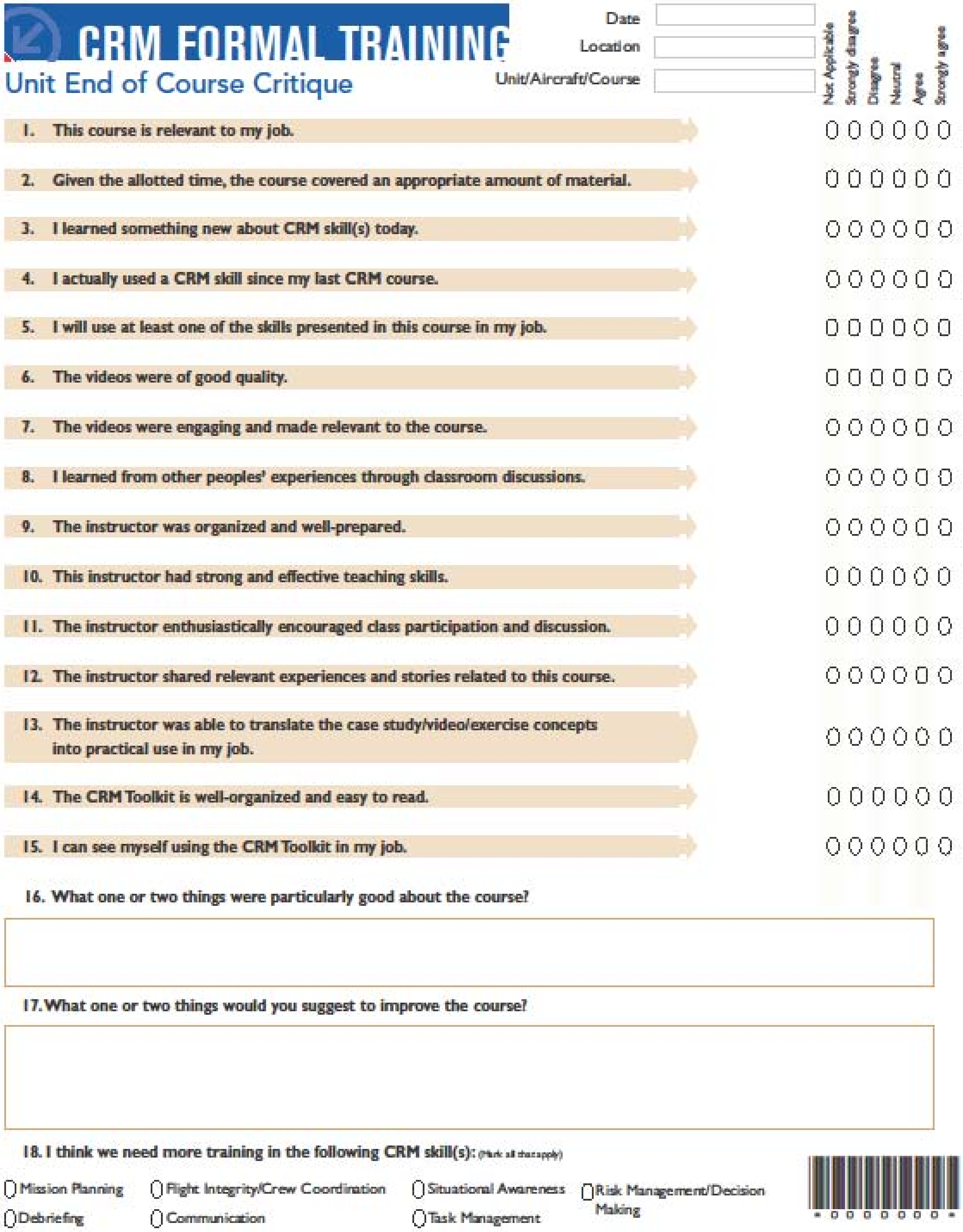
Office of the Secretary of Defense (2006, Feb 6). *The Quadrennial Defense Review Report*. http://www.globalsecurity.org/military/library/policy/dod/qdr-2006-report.pdf

Salas, E., Wilson, K.A., Burke, S.C., & Wightman, D.C. (2006). Does crew resource management training work? An update, an extension, and some critical needs. *Human Factors*, *48* (2), 392-412.

Spiker, V.A., Hunt, S.K., & Walls, W.F. (2005, July). *User reaction to annotated approach* (CNAP Technical Memorandum). San Diego, CA: Commander Naval Air Force, US Pacific Fleet.

Tvaryanas, A.P., Thompson, W.T., & Constable, S.H. (2005). *U.S. Military Unmanned Aerial Vehicle Mishaps: Assessment of the role of human factors using Human Factors Analysis and Classification System (HFACS).* HSW-PE-BR-TR-2005-0001.

APPENDIX A



CRM FORMAL TRAINING

Unit End of Course Critique

Date

Location

Unit/Aircraft/Course

	Not Applicable	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. This course is relevant to my job.	○	○	○	○	○	○
2. Given the allotted time, the course covered an appropriate amount of material.	○	○	○	○	○	○
3. I learned something new about CRM skill(s) today.	○	○	○	○	○	○
4. I actually used a CRM skill since my last CRM course.	○	○	○	○	○	○
5. I will use at least one of the skills presented in this course in my job.	○	○	○	○	○	○
6. The videos were of good quality.	○	○	○	○	○	○
7. The videos were engaging and made relevant to the course.	○	○	○	○	○	○
8. I learned from other peoples' experiences through classroom discussions.	○	○	○	○	○	○
9. The instructor was organized and well-prepared.	○	○	○	○	○	○
10. This instructor had strong and effective teaching skills.	○	○	○	○	○	○
11. The instructor enthusiastically encouraged class participation and discussion.	○	○	○	○	○	○
12. The instructor shared relevant experiences and stories related to this course.	○	○	○	○	○	○
13. The instructor was able to translate the case study/video/exercise concepts into practical use in my job.	○	○	○	○	○	○
14. The CRM Toolkit is well-organized and easy to read.	○	○	○	○	○	○
15. I can see myself using the CRM Toolkit in my job.	○	○	○	○	○	○

16. What one or two things were particularly good about the course?

17. What one or two things would you suggest to improve the course?

18. I think we need more training in the following CRM skill(s): (Mark all that apply)

○ Mission Planning
○ Flight Integrity/Crew Coordination
○ Situational Awareness
○ Risk Management/Decision Making
○ Debriefing
○ Communication
○ Task Management

0000000

APPENDIX B

Advanced Focus Academics Pre Survey (v 8.2)

Name: ___________________________ Class #:_____________ Crew position:____________

The purpose of this survey is to evaluate the training for research purposes only.
This will not be a reflection of your individual performance.

Circle the best answer for each question.

1. Which of the following is <u>not</u> considered to be a defense against threats?
   a. Very experienced crewmembers.
   b. Alarms.
   c. Effective Squadron Operating Procedures.
   d. Effective use of CRM skills.

2. All of the crewmembers were intently involved with a new checklist procedure. As the crewmembers were accomplishing the procedure, they did not notice the altitude warnings on the HUD as the aircraft descended below the minimum safe altitude. Which situational awareness warning signs are most likely present and how should crewmembers respond?
   a. ISO recognizes confusion. “IP, my SO and I are still not clear on this procedure.”
   b. IP recognizes fixation. “Pilot, we have spent too much time on this procedure; check our aircraft flight path.”
   c. Pilot recognizes information overload. “SO, clean off some of the data on the sensor display.”
   d. UIP recognizes ambiguity. “IP, I don’t see the proper readouts on the HUD.”

3. The pilot put the gear down and started the gear bump down check. The SO was unfamiliar with the check, in fact, the pilot was the only crewmember experienced with this procedure. Therefore, the IP, UIP and ISO also became involved with listening to and observing how to perform this task. As the crewmembers were accomplishing the bump down procedure, they did not notice the altitude warnings on the HUD as the aircraft descended below the minimum safe altitude. Of the actions listed, which one would have the best chance of helping the crew avoid the aircraft descending below MSA?
   a. The pilot should tell the SO to remove unnecessary information from the display.
   b. The pilot should maintain his focus on flying the aircraft and let the ISO train the SO.
   c. The pilot should recognize that the aircraft did not start leveling off at the MSA, and announce the information to the crew as soon as possible to prevent the aircraft from descending further.
   d. The pilot should ask the UIP or IP to back him up on monitoring the aircraft flight data as he is instructing/accomplishing the bump down procedure.

4. Which of the following best describes a threat?

   a. An event that is considered an error, caused by a crewmember, that must be identified and corrected to prevent an undesirable outcome.
   b. An event that is considered an error which impacts the success and/or the safety of the mission.
   c. An event that requires the crew's attention and response if mission requirements are to be met and/or safety maintained.
   d. An event that causes one of the crewmembers to make an error.

5. Your crew is flying a mission above FL 230 and is redirected to a known combat zone to gather imagery. You notice several thunderstorms as you approach the area of interest. The thunderstorms are situated such that you might be able to get between them and get the required imagery. You estimate the gap between the thunderstorms to be approximately 12-18 nm. Which of the following would be the best course of action for the pilot to select to maintain safety and execute the mission?

   a. Direct the SO to determine the exact distance between thunderstorms, then decide whether they can conduct the combat mission or not.
   b. Solicit information on alternative courses of action from the crew, then make and announce the decision, and brief and execute the plan.
   c. Decide whether there is sufficient space between thunderstorms to conduct the combat mission, then clearly direct how to accomplish the mission.
   d. Determine the course of action, then tell the crew how to execute the mission.

6. The aircraft is heading toward a mountain that appears higher than your current altitude. The crew consists of an experienced MQ-1/former fighter pilot and a SO fresh out of FTU. The pilot commands a right turn to avoid the mountain and does not disengage the airspeed and heading hold functions. Which of the following would be the best response from the SO?

   a. SO should say, "Pilot, that ridgeline looks pretty high to me."
   b. SO should say, "Pilot the aircraft looks like it is co-altitude with the mountain."
   c. SO should say, "Pilot, did you intend to leave the airspeed and heading hold engaged?"
   d. SO should say, "Pilot, disengage airspeed and heading hold."

7. Which of the following statements is true?

   a. It is possible to have an error-free flight.
   b. Experienced crews can expect to make fewer errors than inexperienced crews.
   c. Error can happen to anyone at any time.
   d. Anticipating threats prior to the mission will prevent errors from occurring.

Advanced Focus Academics Post Survey (v 8.2)

Name: ___________________________ Class #:_____________ Crew position:____________

The purpose of this survey is to evaluate the training for research purposes only.
This will not be a reflection of your individual performance.

Circle the best answer for each question.

1. Which of the following best describes a threat?
   a. An event that requires the crew's attention and response if mission requirements are to be met and/or safety maintained.
   b. An event that is considered an error which impacts the success and/or the safety of the mission.
   c. An event that is considered an error, caused by a crewmember, that must be identified and corrected to prevent an undesirable outcome.
   d. An event that causes one of the crewmembers to make an error.

2. Which of the following is <u>not</u> considered to be a defense against threats?
   a. Alarms.
   b. Effective Squadron Operating Procedures.
   c. Very experienced crewmembers.
   d. Effective use of CRM skills.

3. Which of the following statements is true?
   a. Anticipating threats prior to the mission will prevent errors from occurring.
   b. Experienced crews can expect to make fewer errors than inexperienced crews.
   c. It is possible to have an error-free flight.
   d. Error can happen to anyone at any time.

4. The pilot put the gear down and started the gear bump down check. The SO was unfamiliar with the check, in fact, the pilot was the only crewmember experienced with this procedure. Therefore, the IP, UIP and ISO also became involved with listening to and observing how to perform this task. As the crewmembers were accomplishing the bump down procedure, they did not notice the altitude warnings on the HUD as the aircraft descended below the minimum safe altitude. Of the actions listed, which one would have the best chance of helping the crew avoid the aircraft descending below MSA?
   a. The pilot should ask the UIP or IP to back him up on monitoring the aircraft flight data as he is instructing/accomplishing the bump down procedure.
   b. The pilot should tell the SO to remove unnecessary information from the display.
   c. The pilot should recognize that the aircraft did not start leveling off at the MSA, and announce the information to the crew as soon as possible to prevent the aircraft from descending further.
   d. The pilot should maintain his focus on flying the aircraft and let the ISO train the SO.

5. Your crew is flying a mission above FL 230 and is redirected to a known combat zone to gather imagery. You notice several thunderstorms as you approach the area of interest. The thunderstorms are situated such that you might be able to get between them and get the required imagery. You estimate the gap between the thunderstorms to be approximately 12-18 nm. Which of the following would be the best course of action for the pilot to select to maintain safety and execute the mission?

    a. Direct the SO to determine the exact distance between thunderstorms, then decide whether they can conduct the combat mission or not.
    b. Decide whether there is sufficient space between thunderstorms to conduct the combat mission, then clearly direct how to accomplish the mission.
    c. Solicit information on alternative courses of action from the crew, then make and announce the decision, and brief and execute the plan.
    d. Determine the course of action, then tell the crew how to execute the mission.

6. The aircraft is heading toward a mountain that appears higher than your current altitude. The crew consists of an experienced MQ-1/former fighter pilot and a SO fresh out of FTU. The pilot commands a right turn to avoid the mountain and does not disengage the airspeed and heading hold functions. Which of the following would be the best response from the SO?

    a. SO should say, “Pilot the aircraft looks like it is co-altitude with the mountain.”
    b. SO should say, “Pilot, that ridgeline looks pretty high to me.”
    c. SO should say, “Pilot, disengage airspeed and heading hold.”
    d. SO should say, “Pilot, did you intend to leave the airspeed and heading hold engaged?”

7. All of the crewmembers were involved with the bump down procedure. As the crewmembers were accomplishing the bump down procedure, they did not notice the altitude warnings on the HUD as the aircraft descended below the minimum safe altitude. Which SA warning signs are most likely present and how should crewmembers respond?

    a. UIP recognizes ambiguity. “IP, I don’t see the proper readouts on the HUD.”
    b. IP recognizes fixation. “Pilot, we have spent too much time on this procedure; check our aircraft flight path.”
    c. ISO recognizes confusion. “IP, my SO and I are still not clear on the bump down procedure.”
    d. Pilot recognizes information overload. “SO, clean off some of the data on the sensor display.”

APPENDIX C

Birds of Prey CRM

Interactive Case Histories (ICH)

Beta-Test Evaluation – Students
Creech AFB

(January 2009)

PARTICIPANT INFORMATION

Current Designation:

□ Pilot □ Sensor □ Instructor □ Other

Total years with this designation?

□ Less than 6 months □ 6 months – 1 year □ 1 – 3 years □ 3 – 5 years □ 5+ years

Total years military flying experience

□ Same as above □ 1 – 3 years □ 3 – 5 years □ 5+ years

Date: _____________ Time: ______________ Class No.:____________

Crew Training International

Anacapa Sciences, Inc.

Birds of Prey CRM -- ICH Beta-Test Evaluation

OVERALL REACTIONS

1) What was your reaction to the overall “look and feel” of the ICH?

□ don’t have enough information to answer

□ very negative □ negative □ neutral □ positive □ very positive

2) How long did it take you to work though the entire case history, including reading the tutorial?

_____________ (minutes)

3) Based on your experience so far, OVERALL, how hard or easy is it to use the ICH?

□ don’t have enough information to answer

□ very hard □ hard □ not easy or hard □ easy □ very easy

4) How *useful* was the information provided in the ICH?

□ don’t have enough information to answer

□ not useful □ moderately useful □ very useful

5) How likely is it that you would recommend ICH to other aviators?

□ don’t have enough information to answer

□ very unlikely □ unlikely □ just not sure □ likely □ very likely

CASE HISTORY CHECKLIST

6) The case history checklist appears on the left-hand side of each ICH case history. It is intended to provide the recommended steps that should be used to review a case history. How easy or hard was it to use the case history checklist?

□ don't have enough information to answer

□ very hard □ hard □ not easy or hard □ easy □ very easy

7) As you check off steps in the checklist, new information items are added to the case history text shown in the right-hand portion of the screen. Was it hard or easy to understand this "progressive addition" of information to the case history text?

□ don't have enough information to answer

□ very hard □ hard □ not easy or hard □ easy □ very easy

CHECK LIST STEP #1: ICH TUTORIAL / HELP

8) How hard or easy is it to understand the information presented by the ICH Tutorial / Help?

□ don't have enough information to answer

□ very hard □ hard □ not easy or hard □ easy □ very easy

Is there information that is missing from the Tutorial / Help that might be added to make it better? ____

____

Is there information that is currently included in the Tutorial / Help that might be removed to make
it better? ____

____

CHECK LIST STEPS #2 and #3: KNOWLEDGE ITEMS (KI) AND SUPPORTING INFORMATION (SI)

9) Knowledge Items (KI) and Supporting Information (SI) icons and links appear in the body of the case history during steps two and three in the case history checklist? Knowledge Items are intended to provide additional information about things like procedures and definitions. Supporting Information items are intended to provide supplementary details that will help you understand the events and actions that are taking place in the case history. How easy or hard was it to find and use the Knowledge Item and Supporting Information icons and links?

□ don't have enough information to answer

□ very hard □ hard □ not easy or hard □ easy □ very easy

10) How useful was the additional information presented by the Knowledge Items (KI)?

□ don't have enough information to answer

□ not useful □ moderately useful □ very useful

11) How useful was the additional information presented by the Supporting Information items (SI)?

□ don't have enough information to answer

□ not useful □ moderately useful □ very useful

12) Were there other knowledge items that could have been added to the case history to help your understanding of what happened in the mishap? ______________________________

____

____

Were there other types of supplemental information that could have provided in the case history that would have helped your understanding of what happened in the mishap? ___

____

____

CHECK LIST STEP #4: GOOD AND POOR CRM BEHAVIOR ASSESSMENT

13) The fourth step in the ICH checklist asks you to play the role of an instructor for the crew in the case history by grading four CRM behaviors. You indicate whether you agree, are neutral, or disagree with whether the behavior broke down during the flight? How hard or easy was it to understand the instructions for completing this "gradesheet?"

☐ don't have enough information to answer

☐ very hard ☐ hard ☐ not easy or hard ☐ easy ☐ very easy

14) How hard or easy was it to understand the CRM behaviors presented for you to grade?

☐ don't have enough information to answer

☐ very hard ☐ hard ☐ not easy or hard ☐ easy ☐ very easy

15) While you are completing this "gradesheet" you can go back to the case history to review information and details that are presented there. How hard or easy was it to move back and forth between the case history and the gradesheet pop-up window?

☐ don't have enough information to answer

☐ very hard ☐ hard ☐ not easy or hard ☐ easy ☐ very easy

CHECK LIST STEP #5: GOOD AND POOR CRM LINKS

16) As part of step five in the ICH checklist, you are asked to review a series of color coded links that appear within the body of the case history. These links highlight good (green) and poor (red) CRM behaviors on the part of the crew in the case history. How easy or hard was it to find and access the good and poor CRM links?

☐ don't have enough information to answer

☐ very hard ☐ hard ☐ not easy or hard ☐ easy ☐ very easy

17) When you click on a good or poor CRM link, a box slides open to provide an explanation of the crew's actions and responses with regard to this behavior. How useful was the information provided in the explanations of the good and poor CRM behavior links?

☐ don't have enough information to answer

☐ not useful ☐ moderately useful ☐ very useful

CHECK LIST STEP #6: WRAP-UP

18) Step six in the ICH checklist asks you to review a wrap-up of the situation, actions, and responses described in the case history. How useful was the information provided in the Wrap-up window?

☐ don't have enough information to answer

☐ not useful ☐ moderately useful ☐ very useful

CRM SKILLS AND BEHAVIORS

19) A listing of CRM Skills and Behaviors along with their associated definitions and explanations is available from a link at the bottom of the ICH checklist? How useful were these definitions in understanding CRM Skills and Behaviors?

☐ don't have enough information to answer

☐ not useful ☐ moderately useful ☐ very useful

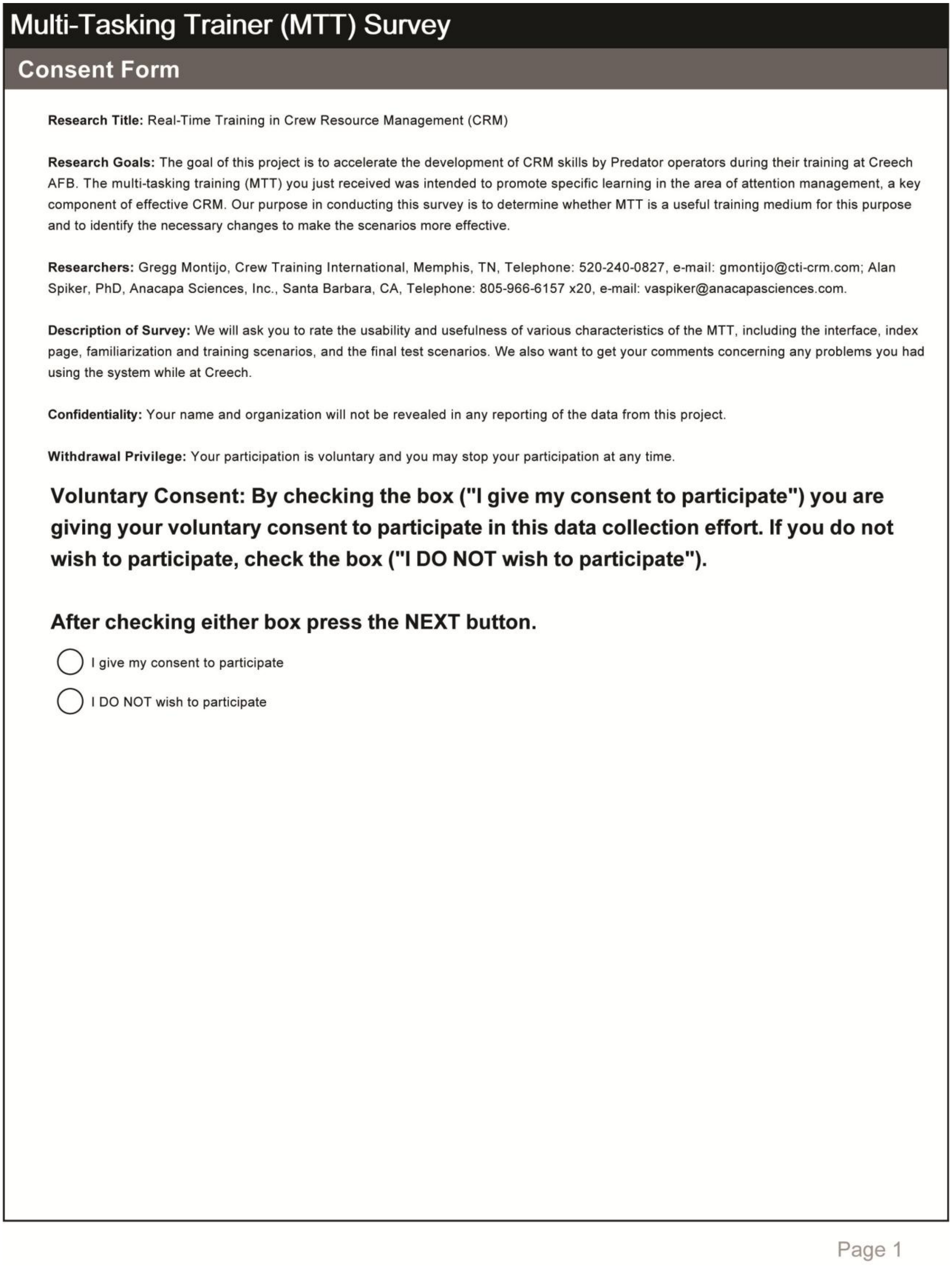
20) The CRM Skills and Behaviors are intended to identify important CRM-related skills and behaviors. Did you find this list of Skills and Behaviors to be a useful way of describing CRM skills or behaviors?

☐ don't have enough information to answer

☐ not useful ☐ moderately useful ☐ very useful

OTHER COMMENTS ABOUT ICH

APPENDIX D



Multi-Tasking Trainer (MTT) Survey

Consent Form

Research Title: Real-Time Training in Crew Resource Management (CRM)

Research Goals: The goal of this project is to accelerate the development of CRM skills by Predator operators during their training at Creech AFB. The multi-tasking training (MTT) you just received was intended to promote specific learning in the area of attention management, a key component of effective CRM. Our purpose in conducting this survey is to determine whether MTT is a useful training medium for this purpose and to identify the necessary changes to make the scenarios more effective.

Researchers: Gregg Montijo, Crew Training International, Memphis, TN, Telephone: 520-240-0827, e-mail: gmontijo@cti-crm.com; Alan Spiker, PhD, Anacapa Sciences, Inc., Santa Barbara, CA, Telephone: 805-966-6157 x20, e-mail: vaspiker@anacapasciences.com.

Description of Survey: We will ask you to rate the usability and usefulness of various characteristics of the MTT, including the interface, index page, familiarization and training scenarios, and the final test scenarios. We also want to get your comments concerning any problems you had using the system while at Creech.

Confidentiality: Your name and organization will not be revealed in any reporting of the data from this project.

Withdrawal Privilege: Your participation is voluntary and you may stop your participation at any time.

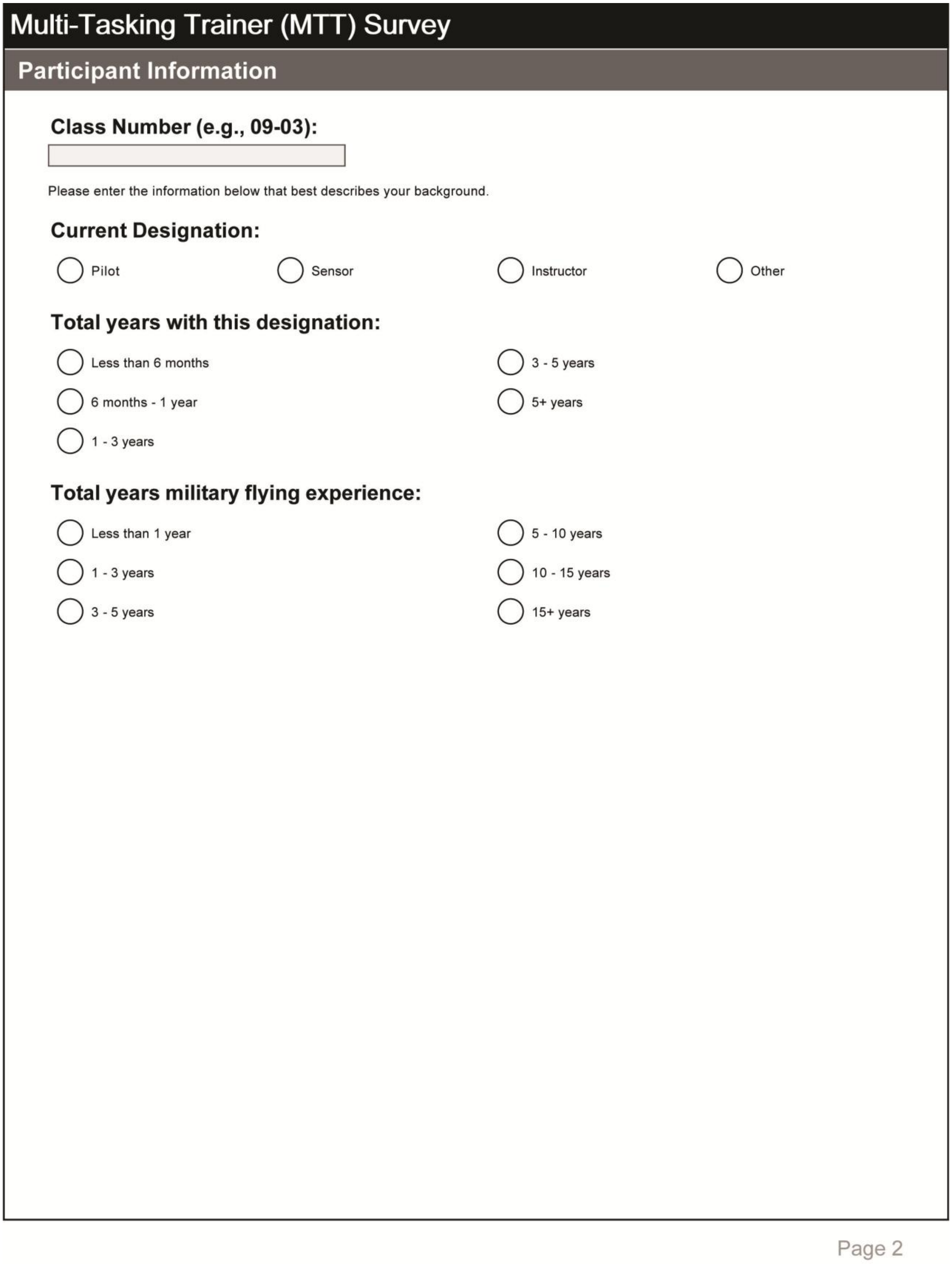
Voluntary Consent: By checking the box ("I give my consent to participate") you are giving your voluntary consent to participate in this data collection effort. If you do not wish to participate, check the box ("I DO NOT wish to participate").

After checking either box press the NEXT button.

○ I give my consent to participate

○ I DO NOT wish to participate

Page 1



Multi-Tasking Trainer (MTT) Survey

Participant Information

Class Number (e.g., 09-03):

Please enter the information below that best describes your background.

Current Designation:

○ Pilot ○ Sensor ○ Instructor ○ Other

Total years with this designation:

○ Less than 6 months

○ 6 months - 1 year

○ 1 - 3 years

○ 3 - 5 years

○ 5+ years

Total years military flying experience:

○ Less than 1 year

○ 1 - 3 years

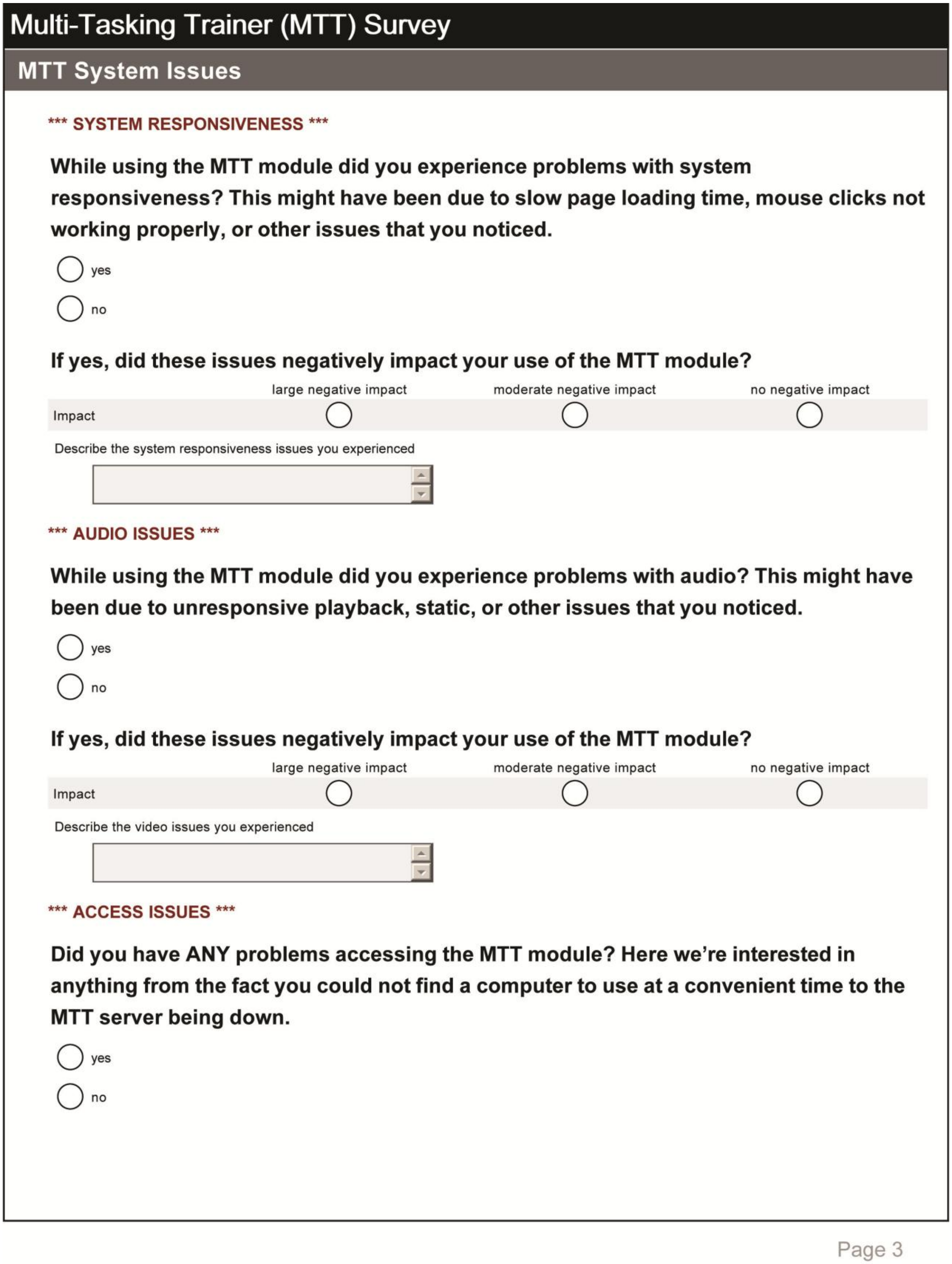
○ 3 - 5 years

○ 5 - 10 years

○ 10 - 15 years

○ 15+ years

Page 2



Multi-Tasking Trainer (MTT) Survey

MTT System Issues

*** SYSTEM RESPONSIVENESS ***

While using the MTT module did you experience problems with system responsiveness? This might have been due to slow page loading time, mouse clicks not working properly, or other issues that you noticed.

○ yes

○ no

If yes, did these issues negatively impact your use of the MTT module?

	large negative impact	moderate negative impact	no negative impact
Impact	○	○	○

Describe the system responsiveness issues you experienced

*** AUDIO ISSUES ***

While using the MTT module did you experience problems with audio? This might have been due to unresponsive playback, static, or other issues that you noticed.

○ yes

○ no

If yes, did these issues negatively impact your use of the MTT module?

	large negative impact	moderate negative impact	no negative impact
Impact	○	○	○

Describe the video issues you experienced

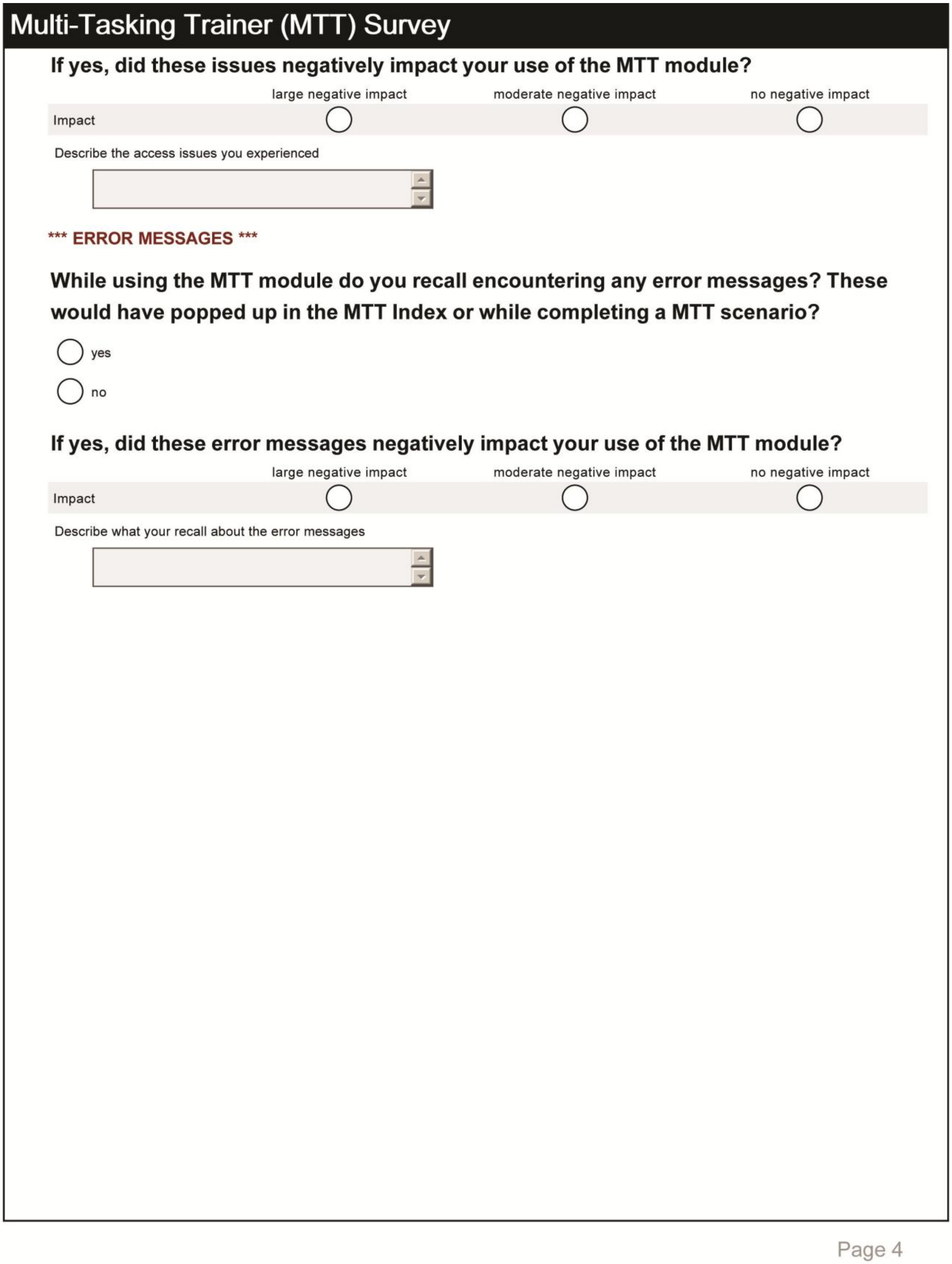
*** ACCESS ISSUES ***

Did you have ANY problems accessing the MTT module? Here we're interested in anything from the fact you could not find a computer to use at a convenient time to the MTT server being down.

○ yes

○ no

Page 3



Multi-Tasking Trainer (MTT) Survey

If yes, did these issues negatively impact your use of the MTT module?

	large negative impact	moderate negative impact	no negative impact
Impact	○	○	○

Describe the access issues you experienced

***** ERROR MESSAGES *****

While using the MTT module do you recall encountering any error messages? These would have popped up in the MTT Index or while completing a MTT scenario?

○ yes

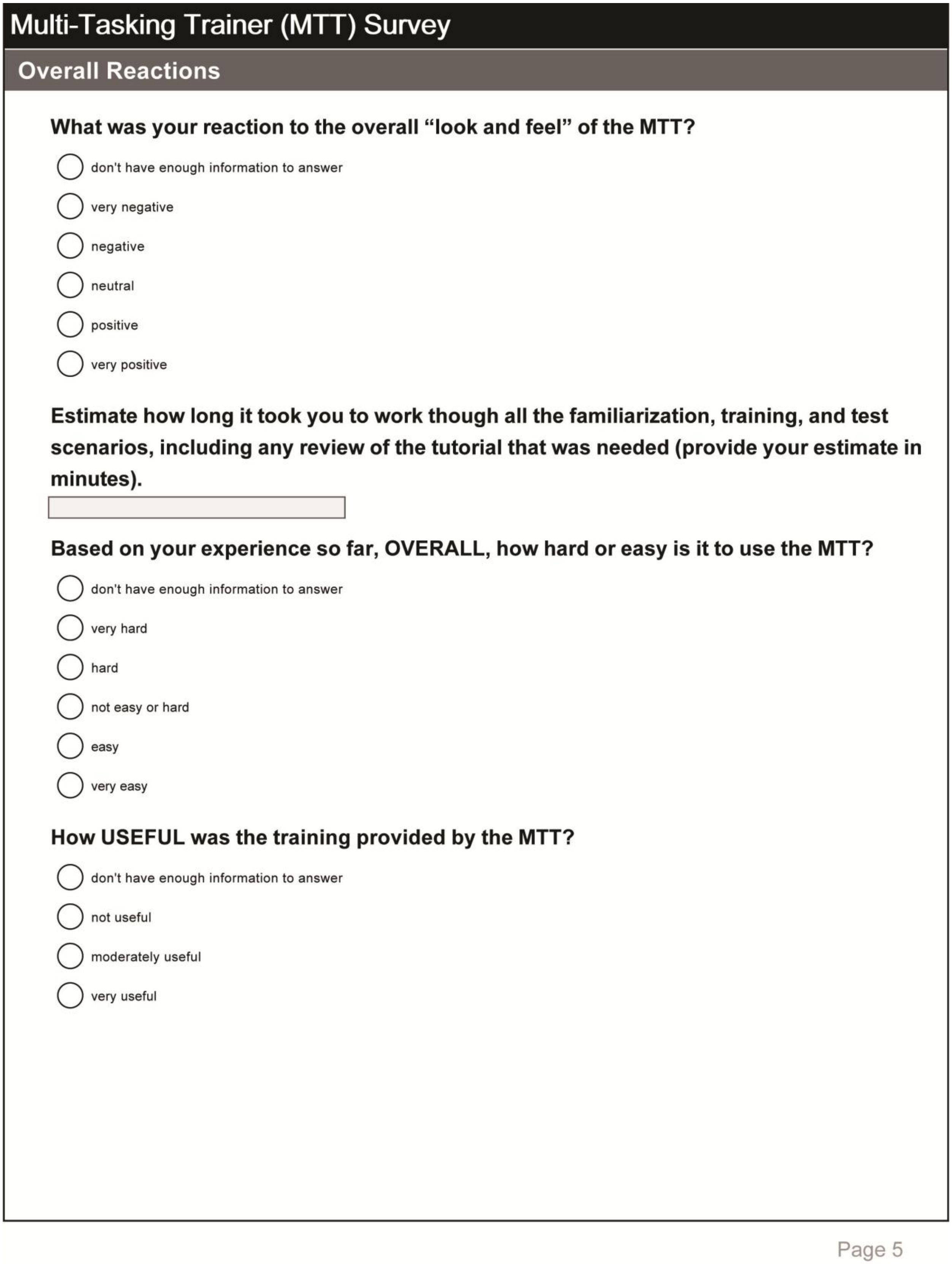
○ no

If yes, did these error messages negatively impact your use of the MTT module?

	large negative impact	moderate negative impact	no negative impact
Impact	○	○	○

Describe what your recall about the error messages

Page 4



Multi-Tasking Trainer (MTT) Survey

Overall Reactions

What was your reaction to the overall "look and feel" of the MTT?

○ don't have enough information to answer

○ very negative

○ negative

○ neutral

○ positive

○ very positive

Estimate how long it took you to work though all the familiarization, training, and test scenarios, including any review of the tutorial that was needed (provide your estimate in minutes).

Based on your experience so far, OVERALL, how hard or easy is it to use the MTT?

○ don't have enough information to answer

○ very hard

○ hard

○ not easy or hard

○ easy

○ very easy

How USEFUL was the training provided by the MTT?

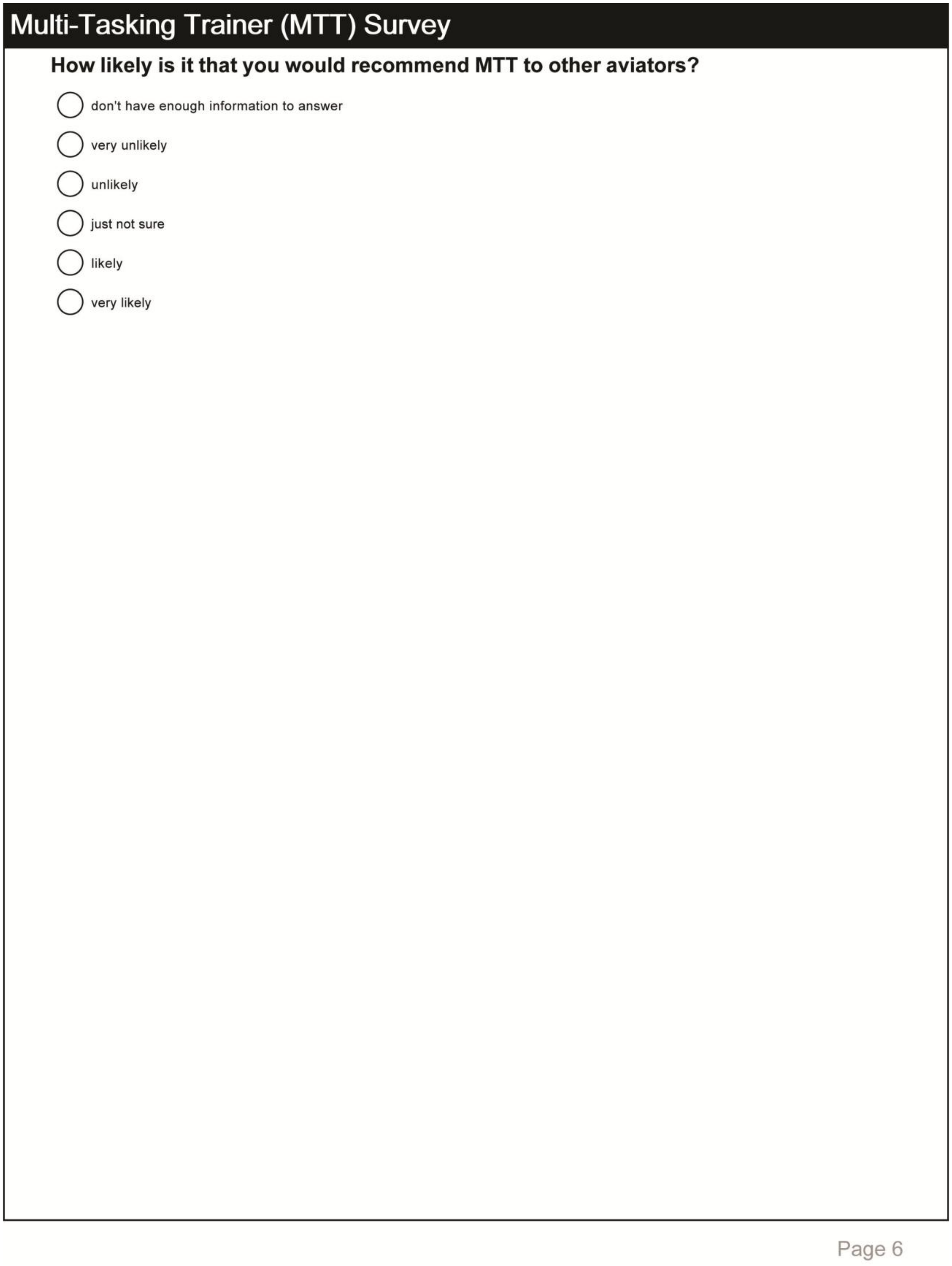
○ don't have enough information to answer

○ not useful

○ moderately useful

○ very useful

Page 5



Multi-Tasking Trainer (MTT) Survey

How likely is it that you would recommend MTT to other aviators?

○ don't have enough information to answer

○ very unlikely

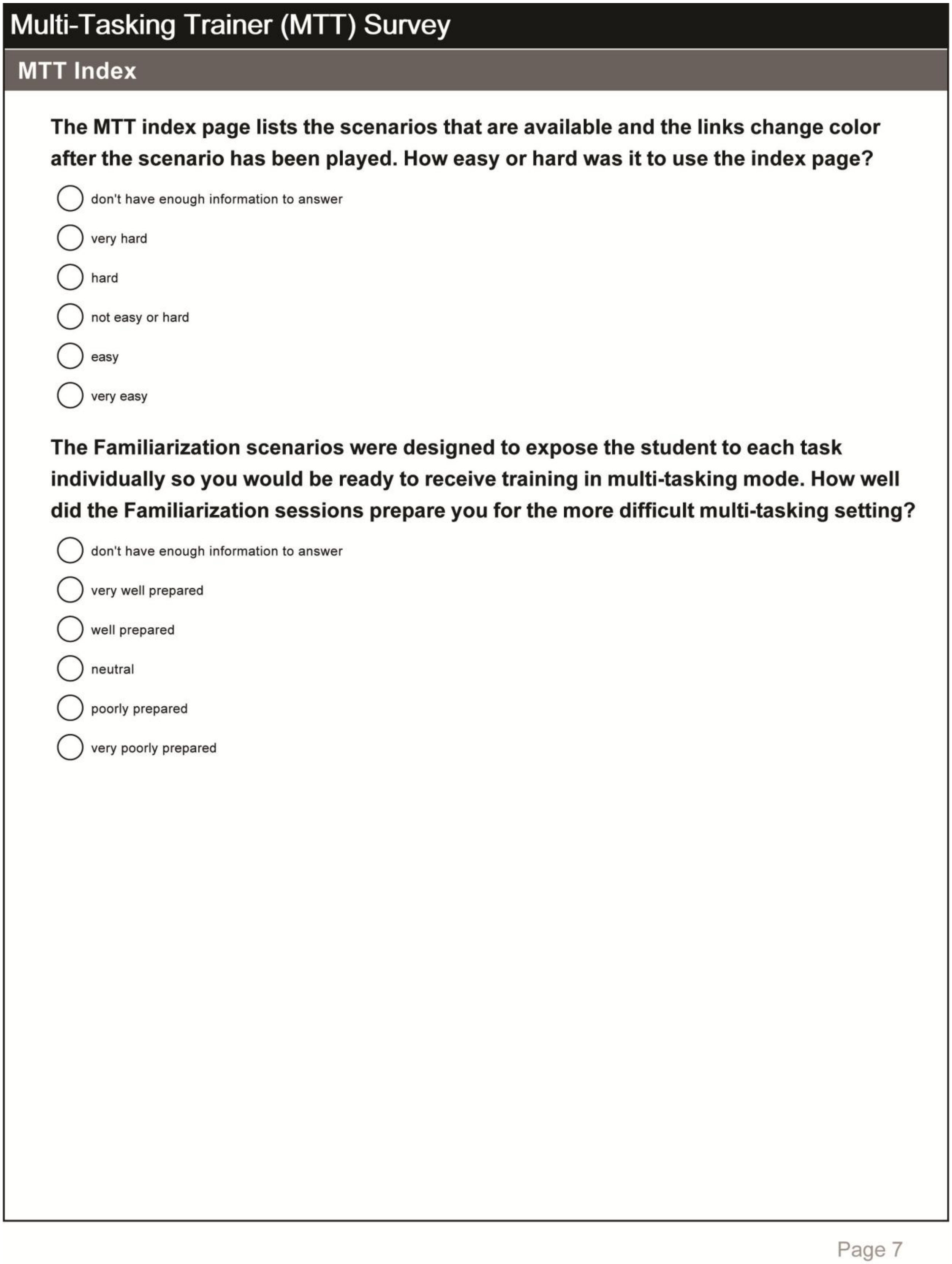
○ unlikely

○ just not sure

○ likely

○ very likely

Page 6



Multi-Tasking Trainer (MTT) Survey

MTT Index

The MTT index page lists the scenarios that are available and the links change color after the scenario has been played. How easy or hard was it to use the index page?

○ don't have enough information to answer

○ very hard

○ hard

○ not easy or hard

○ easy

○ very easy

The Familiarization scenarios were designed to expose the student to each task individually so you would be ready to receive training in multi-tasking mode. How well did the Familiarization sessions prepare you for the more difficult multi-tasking setting?

○ don't have enough information to answer

○ very well prepared

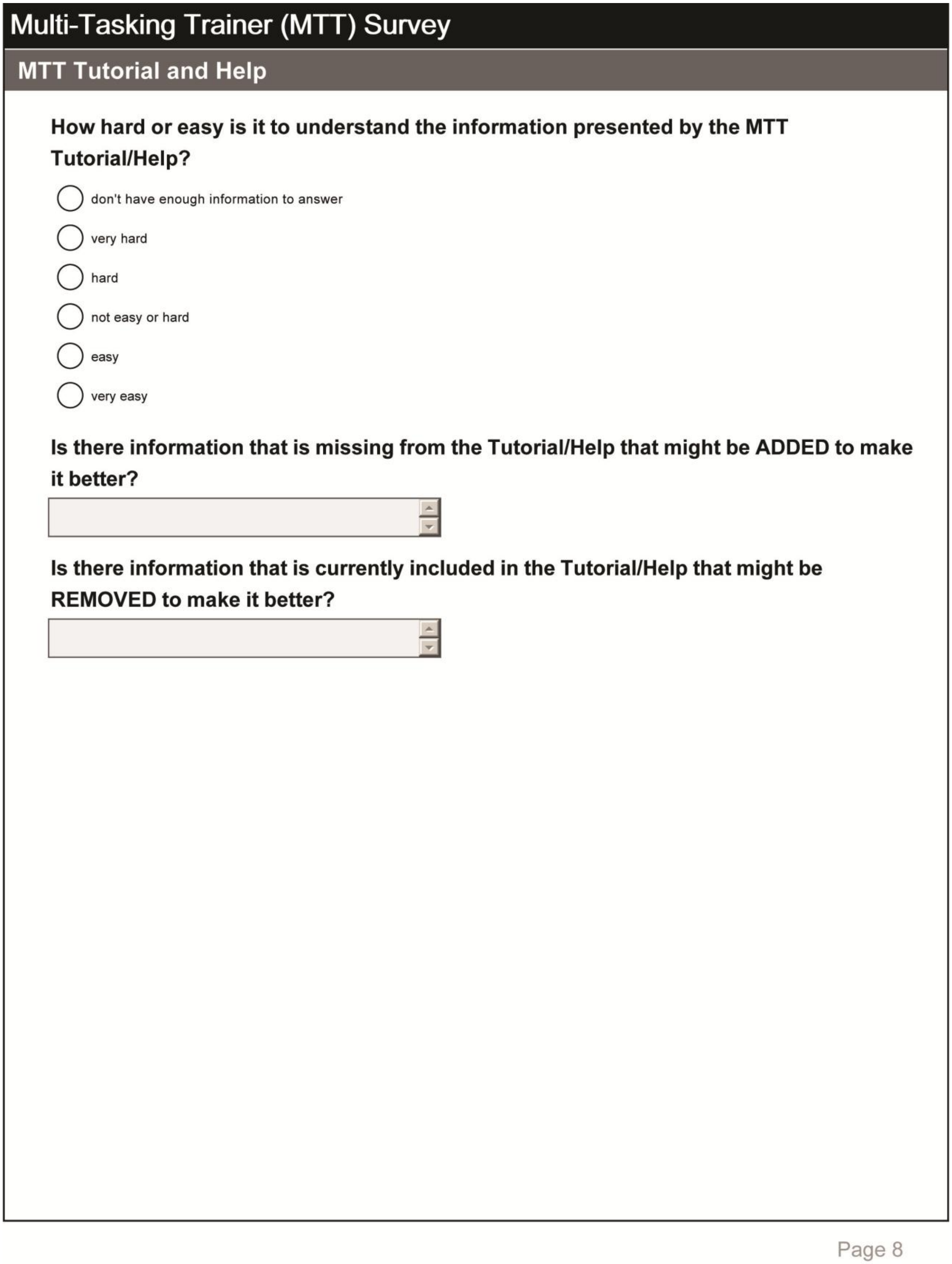
○ well prepared

○ neutral

○ poorly prepared

○ very poorly prepared

Page 7



Multi-Tasking Trainer (MTT) Survey

MTT Tutorial and Help

How hard or easy is it to understand the information presented by the MTT Tutorial/Help?

○ don't have enough information to answer

○ very hard

○ hard

○ not easy or hard

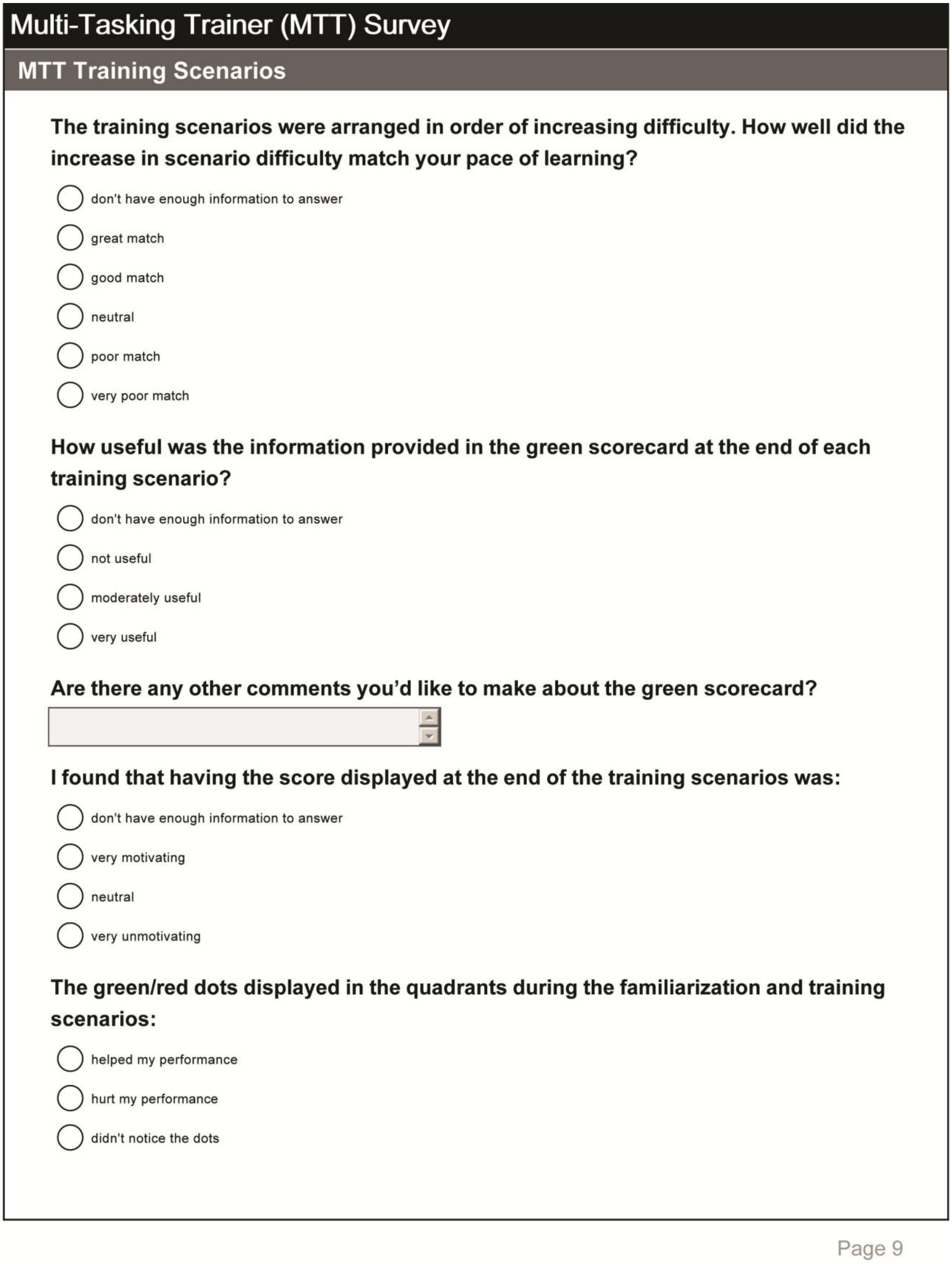
○ easy

○ very easy

Is there information that is missing from the Tutorial/Help that might be ADDED to make it better?

Is there information that is currently included in the Tutorial/Help that might be REMOVED to make it better?

Page 8



Multi-Tasking Trainer (MTT) Survey

MTT Training Scenarios

The training scenarios were arranged in order of increasing difficulty. How well did the increase in scenario difficulty match your pace of learning?

○ don't have enough information to answer

○ great match

○ good match

○ neutral

○ poor match

○ very poor match

How useful was the information provided in the green scorecard at the end of each training scenario?

○ don't have enough information to answer

○ not useful

○ moderately useful

○ very useful

Are there any other comments you'd like to make about the green scorecard?

I found that having the score displayed at the end of the training scenarios was:

○ don't have enough information to answer

○ very motivating

○ neutral

○ very unmotivating

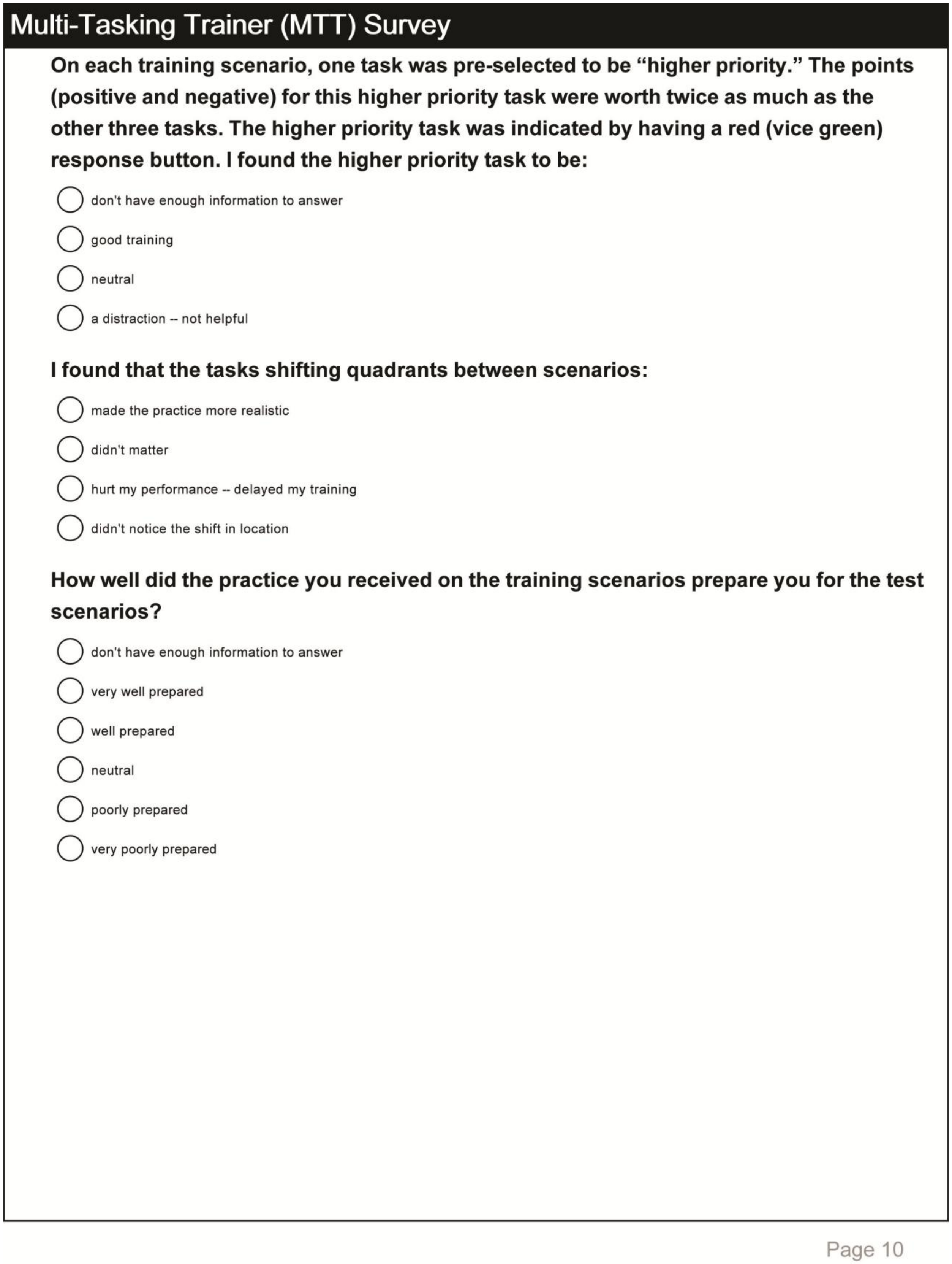
The green/red dots displayed in the quadrants during the familiarization and training scenarios:

○ helped my performance

○ hurt my performance

○ didn't notice the dots

Page 9



Multi-Tasking Trainer (MTT) Survey

On each training scenario, one task was pre-selected to be "higher priority." The points (positive and negative) for this higher priority task were worth twice as much as the other three tasks. The higher priority task was indicated by having a red (vice green) response button. I found the higher priority task to be:

○ don't have enough information to answer

○ good training

○ neutral

○ a distraction -- not helpful

I found that the tasks shifting quadrants between scenarios:

○ made the practice more realistic

○ didn't matter

○ hurt my performance -- delayed my training

○ didn't notice the shift in location

How well did the practice you received on the training scenarios prepare you for the test scenarios?

○ don't have enough information to answer

○ very well prepared

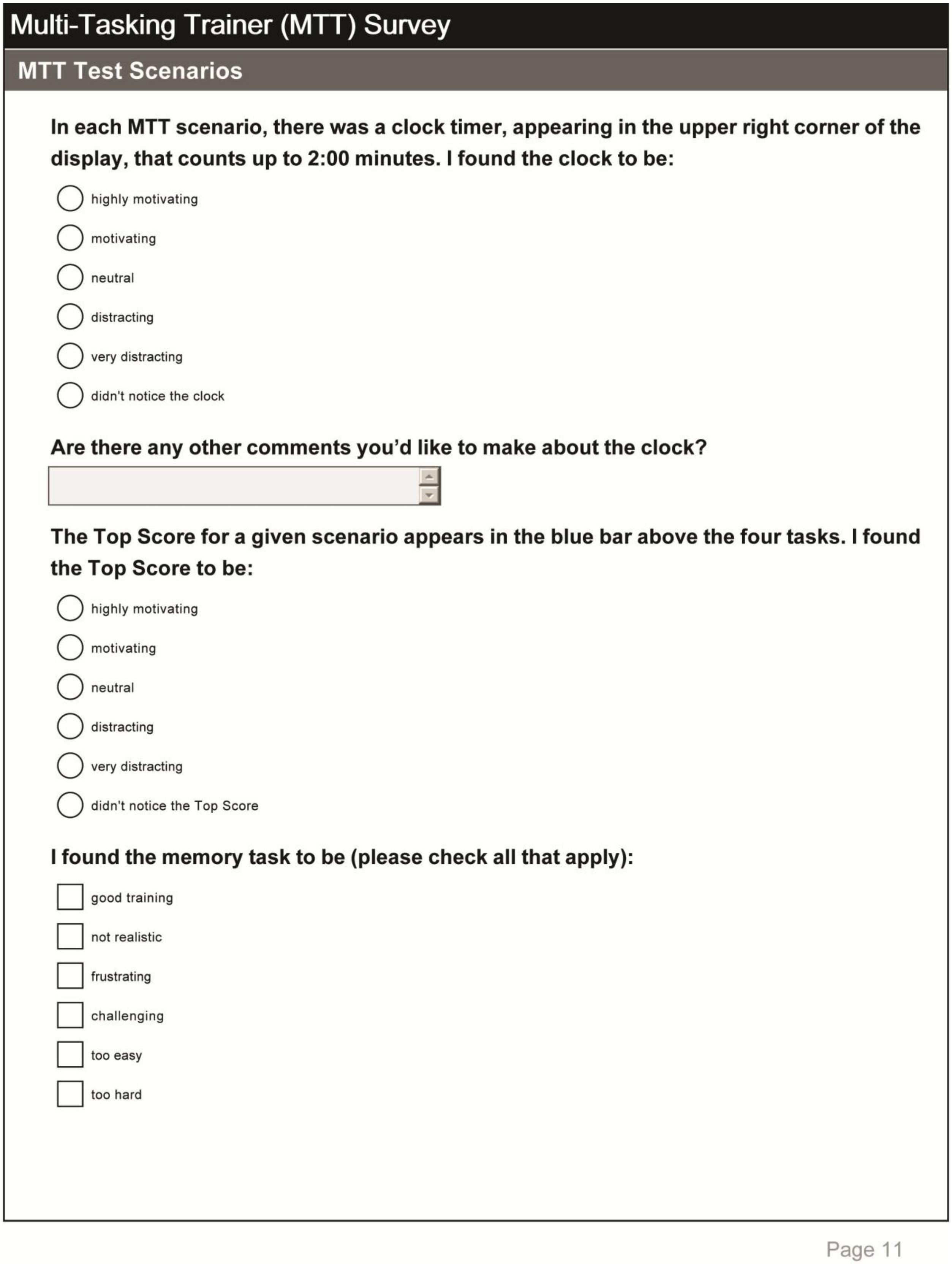
○ well prepared

○ neutral

○ poorly prepared

○ very poorly prepared

Page 10



Multi-Tasking Trainer (MTT) Survey

MTT Test Scenarios

In each MTT scenario, there was a clock timer, appearing in the upper right corner of the display, that counts up to 2:00 minutes. I found the clock to be:

○ highly motivating

○ motivating

○ neutral

○ distracting

○ very distracting

○ didn't notice the clock

Are there any other comments you'd like to make about the clock?

The Top Score for a given scenario appears in the blue bar above the four tasks. I found the Top Score to be:

○ highly motivating

○ motivating

○ neutral

○ distracting

○ very distracting

○ didn't notice the Top Score

I found the memory task to be (please check all that apply):

□ good training

□ not realistic

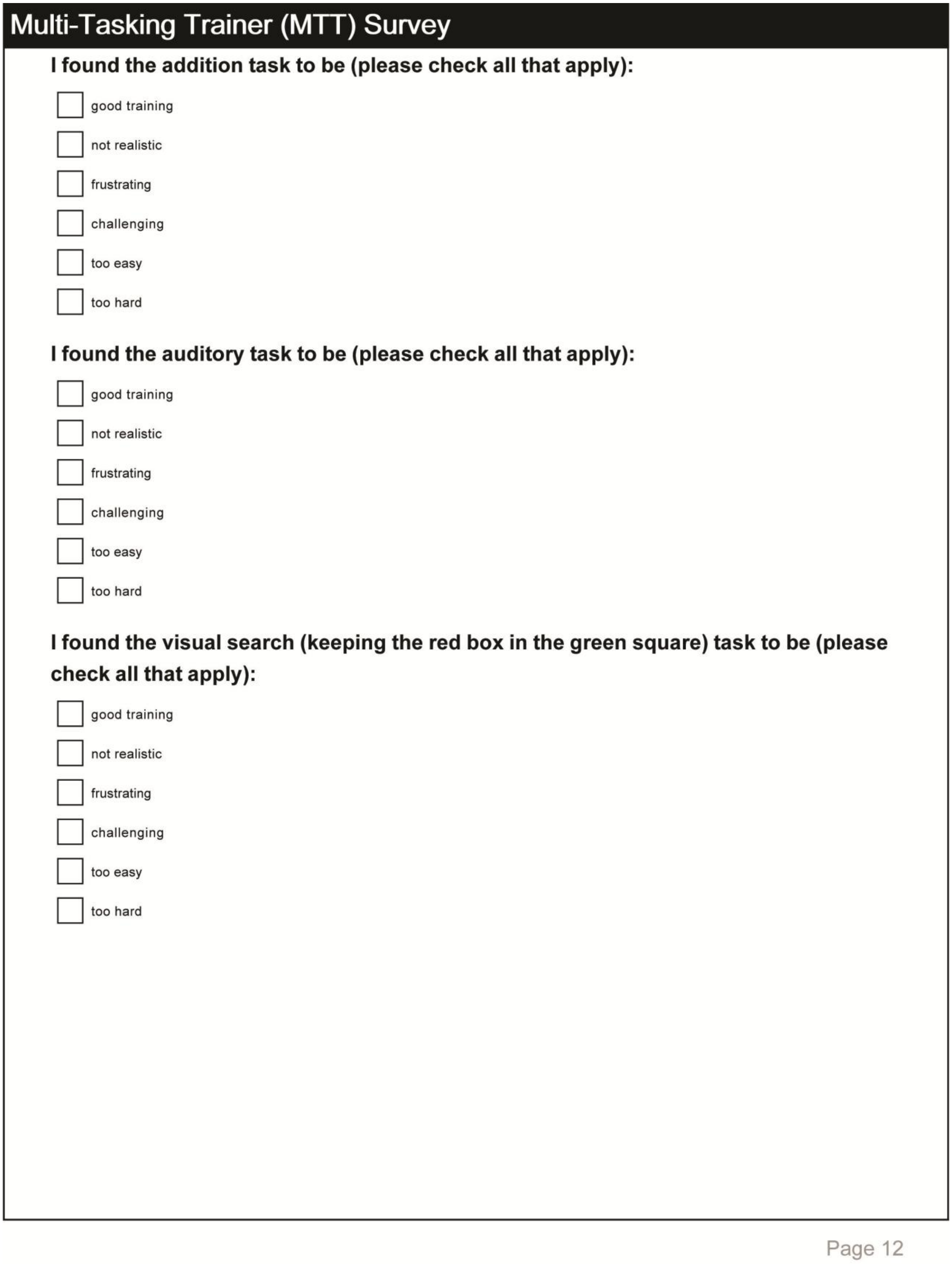
□ frustrating

□ challenging

□ too easy

□ too hard

Page 11



Multi-Tasking Trainer (MTT) Survey

I found the addition task to be (please check all that apply):

□ good training

□ not realistic

□ frustrating

□ challenging

□ too easy

□ too hard

I found the auditory task to be (please check all that apply):

□ good training

□ not realistic

□ frustrating

□ challenging

□ too easy

□ too hard

I found the visual search (keeping the red box in the green square) task to be (please check all that apply):

□ good training

□ not realistic

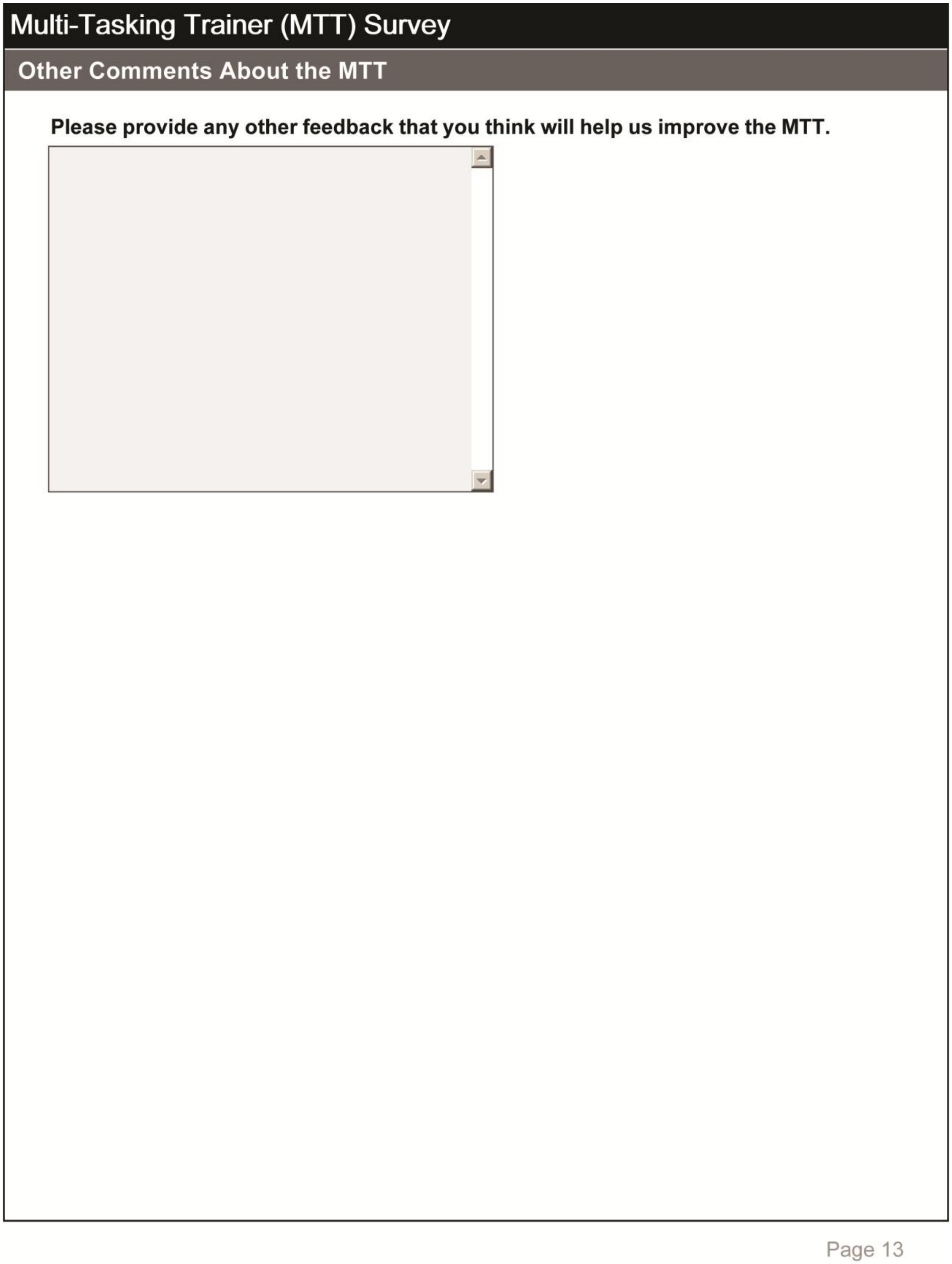
□ frustrating

□ challenging

□ too easy

□ too hard

Page 12

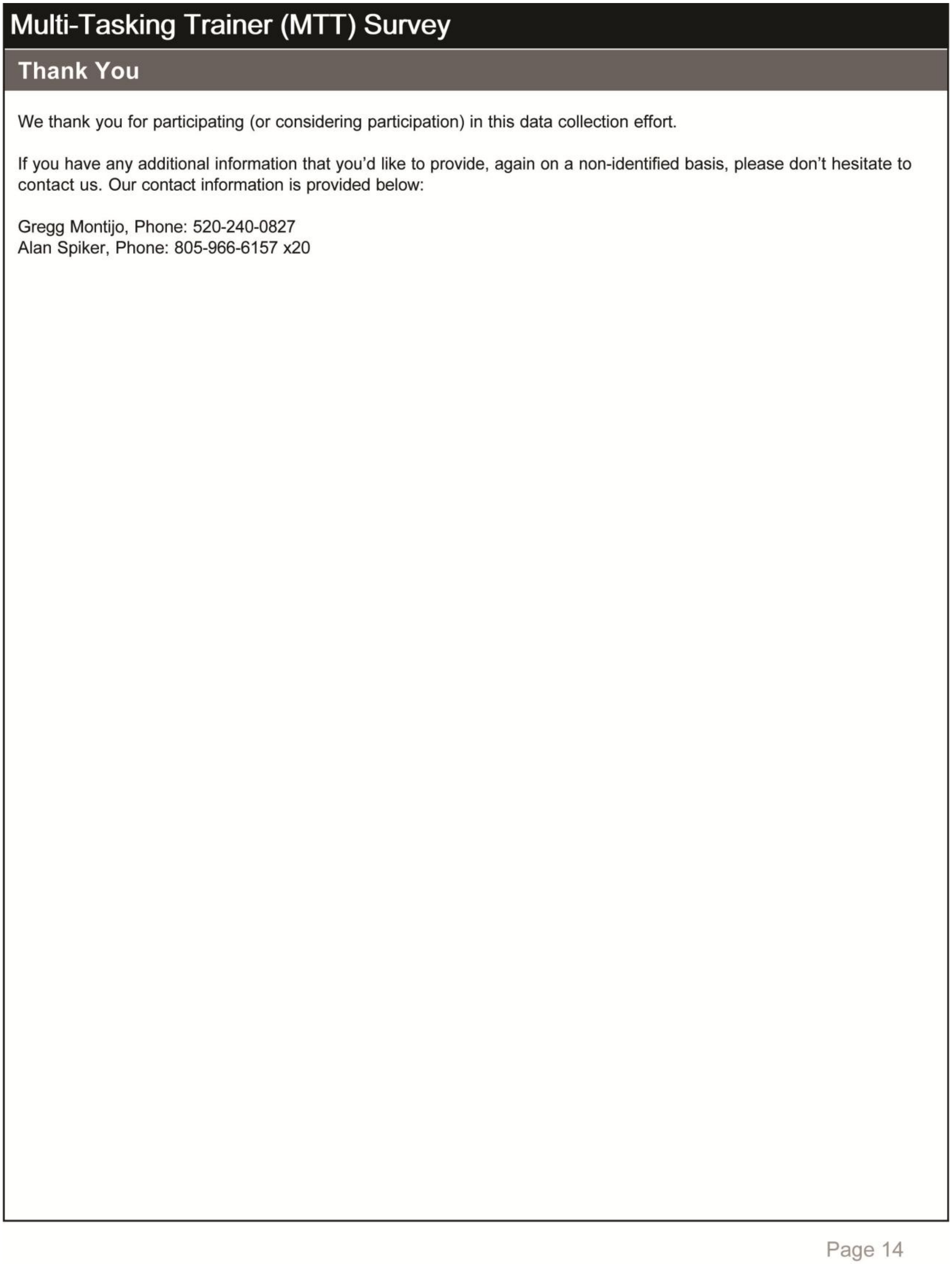


Multi-Tasking Trainer (MTT) Survey

Other Comments About the MTT

Please provide any other feedback that you think will help us improve the MTT.

Page 13



Multi-Tasking Trainer (MTT) Survey

Thank You

We thank you for participating (or considering participation) in this data collection effort.

If you have any additional information that you'd like to provide, again on a non-identified basis, please don't hesitate to contact us. Our contact information is provided below:

Gregg Montijo, Phone: 520-240-0827
Alan Spiker, Phone: 805-966-6157 x20

Page 14

APPENDIX E

Human Factors (HF) Skills RATING FORM UAV PREDATOR

Class Number	Student	Training Date	Instructor/Observer	Crew Task(s) Observed

HF Skill Rating Scale

POOR	MARGINAL	STANDARD	VERY GOOD	EXCEPTIONAL	
0 – performance indicates a lack of ability or knowledge	1 – performance is safe, but proficiency is limited; Makes errors of commission or omission	2 – performance is essentially correct; recognizes and corrects errors	3 – performance is correct, efficient, skillful, and without hesitation	4 – performance reflects an unusually high degree of ability	□

Mark □ with a + if the pilot's HF behavior was a strength and – if it was a weakness in making your rating.

Avoids Channelized Attention: attention is not focused too long on items or tasks at the expense of more pressing information	**0 1 2 3 4**

□ Effective cross-check; includes all relevant displays/information in scan (e.g., SO alerts pilot to lost-link condition)
□ Cross-check does not stagnate; doesn't dwell too long on one display (e.g., doesn't allow "eye magnets" to monopolize attention)
□ Switches attention as the situation/priority changes (e.g., abandons ISR tasking to handle a lost link situation)
□ Adjusts well to different cockpit environments with little/no negative transfer (e.g., MGCS vs. FFGCS)
□ Doesn't allow radios/chat to divert attention from higher priorities (e.g., correctly prioritizes answering chat/radios vs. other tasks)
□ Able to shift attention without external cues (e.g., immediately notices airspeed in yellow/red range and challenges deviation)

Record any specific actions that you found notable:

Task Prioritization: performs high priority tasks before lower priority tasks	**0 1 2 3 4**

□ Able to tell which task has higher priority (e.g., concentrates on landing pattern vs. troubleshooting a noncritical system)
□ Able to handle interruptions; doesn't disrupt ongoing tasks (e.g., returns to checklist in progress at the right step)
□ Returns to an interrupted task; completes checklist (e.g., returns to assigned ISR tasking after completing a weather scan)
□ Able to suspend lower priority tasks in favor of higher ones (e.g., suspends a checklist to handle an immediate priority)
□ Able to accomplish tasks concurrently (e.g., runs before-landing checklist while checking over aircraft with ball camera)
□ Adheres to aviate-navigate-communicate (e.g., maintains aircraft control in all situations)

Record any specific actions that you found notable:

Select Course of Action (COA): exercises sound decision-making by selecting a choice that leads to a good outcome	**0 1 2 3 4**

□ Considers all applicable options (e.g., abandons previous course of action that is no longer viable)
□ Able to distinguish assumptions from facts (e.g., actively seeks additional data/info to verify assumptions)
□ Avoids making hasty decisions; uses all available information (e.g., requests help from MCC, CAOC, SOF, etc.)
□ Doesn't take too long to make a decision; is not "overcome by events" (e.g., makes decisions while options still exist)
□ Identifies potential risks from a given decision (e.g., distinguishes between acceptable risk vs. unnecessary risk)
□ Recognizes need for a follow-on decision (e.g., selects another checklist/course of action if initial corrective action doesn't work)

Record any specific actions that you found notable:

SEE OTHER SIDE FOR REST OF GRADING SHEET

Crew Coordination: crew works together as a team to accomplish required task or mission — 0 1 2 3 4

□ Crewmembers divide tasks and responsibilities efficiently (e.g., able to flex to other crewmember's tasks inc/dec)
□ Crewmember performs team tasks as well as individual tasks (e.g., defaults priority back to individual tasks when required)
□ Crewmember anticipates information needs of the other (e.g., ready with next checklist prior to it being requested)

□ Crewmember provides timely data to other crewmember (e.g., SO immediately informs pilot of lost-link condition)

□ Crewmember cross-checks other crewmember's performance (e.g., ensures response in "challenge and response" is correct)

□ Crewmembers maintain "Shared Mental Model" as priorities change (e.g., crewmembers brief each other on changes to the original plan)

□ Crewmembers effectively convey "Shared Mental Model" to oncoming crew (e.g., at change-over)

Record any specific actions that you found notable:

Degree of Instructor Influence: degree of assistance rendered so that crew could accomplish the observed crew tasks
(Please use scale anchors below to make your rating.) — 0 1 2 3 4

0 – Extensive assistance is required for the crew to successfully accomplish the observed crew tasks	1 – Substantial assistance is required for the crew to successfully accomplish parts of the observed crew tasks	2 – Limited assistance is provided for the crew to successfully accomplish the observed crew tasks	3 – Coaching (for technique only) is provided for the crew to successfully accomplish the observed crew tasks	4 – No assistance; errors are corrected by the crew and/or crewmember when discovered

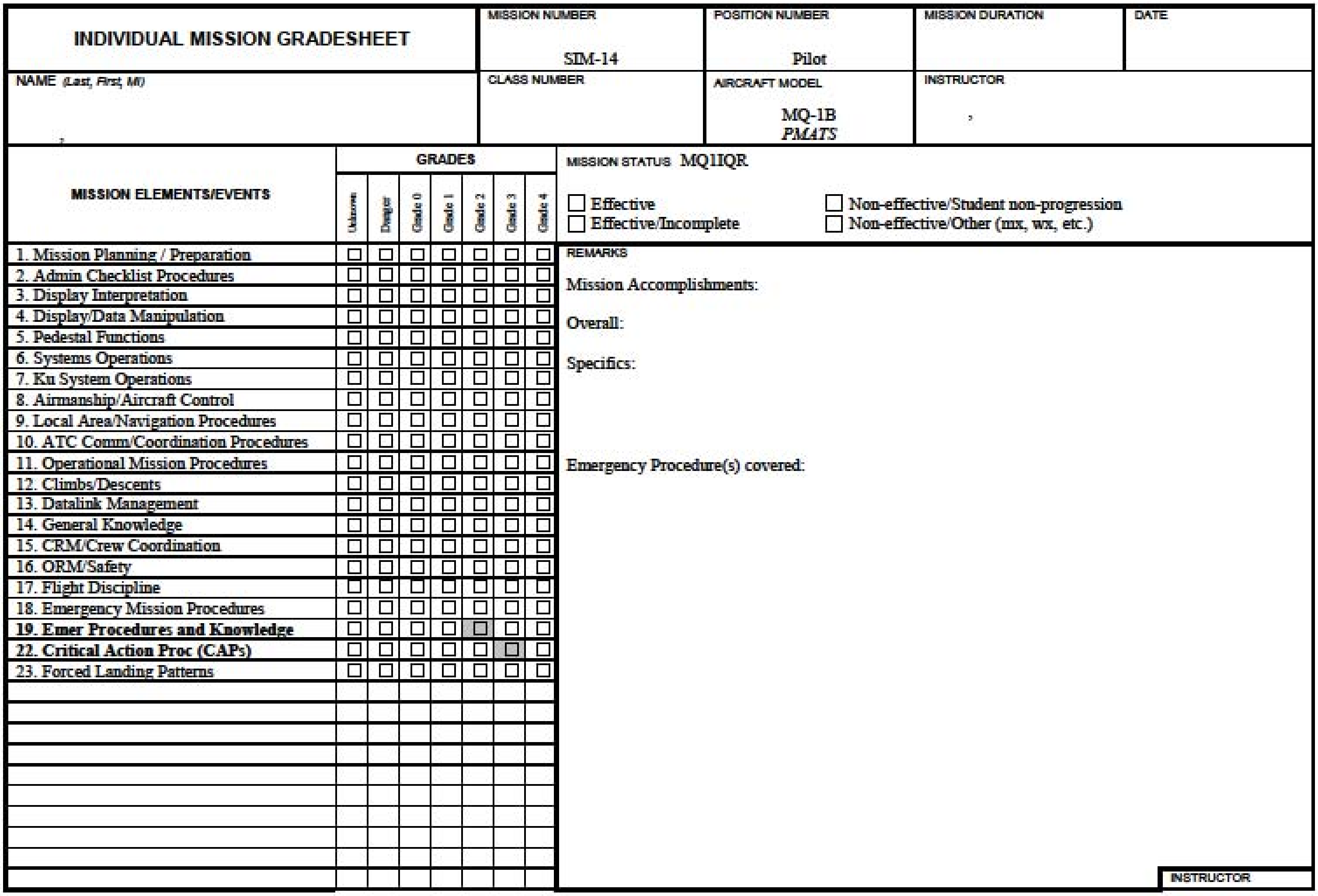
Please record any comments:

Crew CRM Performance: able to execute all aspects of CRM as a *crew* to achieve mission success
(Please use scale anchors below to make your rating; provide only a single rating for the crew.) — 0 1 2 3 4

0 – Observed CRM performance is significantly below expectations for this level of training	1 – Observed CRM performance is less than desired for effective tactical mission success	2 – The demonstrated behavior promotes and maintains mission operations effectiveness	3 – Observed CRM performance is significantly above expectations for this level of training	4 – Observed CRM performance represents a high level of skill in the application of specific behaviors

Please record any comments:

APPENDIX F



INDIVIDUAL MISSION GRADESHEET

MISSION NUMBER: SIM-14

POSITION NUMBER: Pilot

MISSION DURATION:

DATE:

NAME *(Last, First, MI)*:

CLASS NUMBER:

AIRCRAFT MODEL: MQ-1B *PMATS*

INSTRUCTOR:

MISSION ELEMENTS/EVENTS	Unknown	Danger	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
1. Mission Planning / Preparation	□	□	□	□	□	□	□
2. Admin Checklist Procedures	□	□	□	□	□	□	□
3. Display Interpretation	□	□	□	□	□	□	□
4. Display/Data Manipulation	□	□	□	□	□	□	□
5. Pedestal Functions	□	□	□	□	□	□	□
6. Systems Operations	□	□	□	□	□	□	□
7. Ku System Operations	□	□	□	□	□	□	□
8. Airmanship/Aircraft Control	□	□	□	□	□	□	□
9. Local Area/Navigation Procedures	□	□	□	□	□	□	□
10. ATC Comm/Coordination Procedures	□	□	□	□	□	□	□
11. Operational Mission Procedures	□	□	□	□	□	□	□
12. Climbs/Descents	□	□	□	□	□	□	□
13. Datalink Management	□	□	□	□	□	□	□
14. General Knowledge	□	□	□	□	□	□	□
15. CRM/Crew Coordination	□	□	□	□	□	□	□
16. ORM/Safety	□	□	□	□	□	□	□
17. Flight Discipline	□	□	□	□	□	□	□
18. Emergency Mission Procedures	□	□	□	□	□	□	□
19. Emer Procedures and Knowledge	□	□	□	□	▨	□	□
22. Critical Action Proc (CAPs)	□	□	□	□	□	▨	□
23. Forced Landing Patterns	□	□	□	□	□	□	□

MISSION STATUS MQ1IQR

□ Effective □ Non-effective/Student non-progression

□ Effective/Incomplete □ Non-effective/Other (mx, wx, etc.)

REMARKS

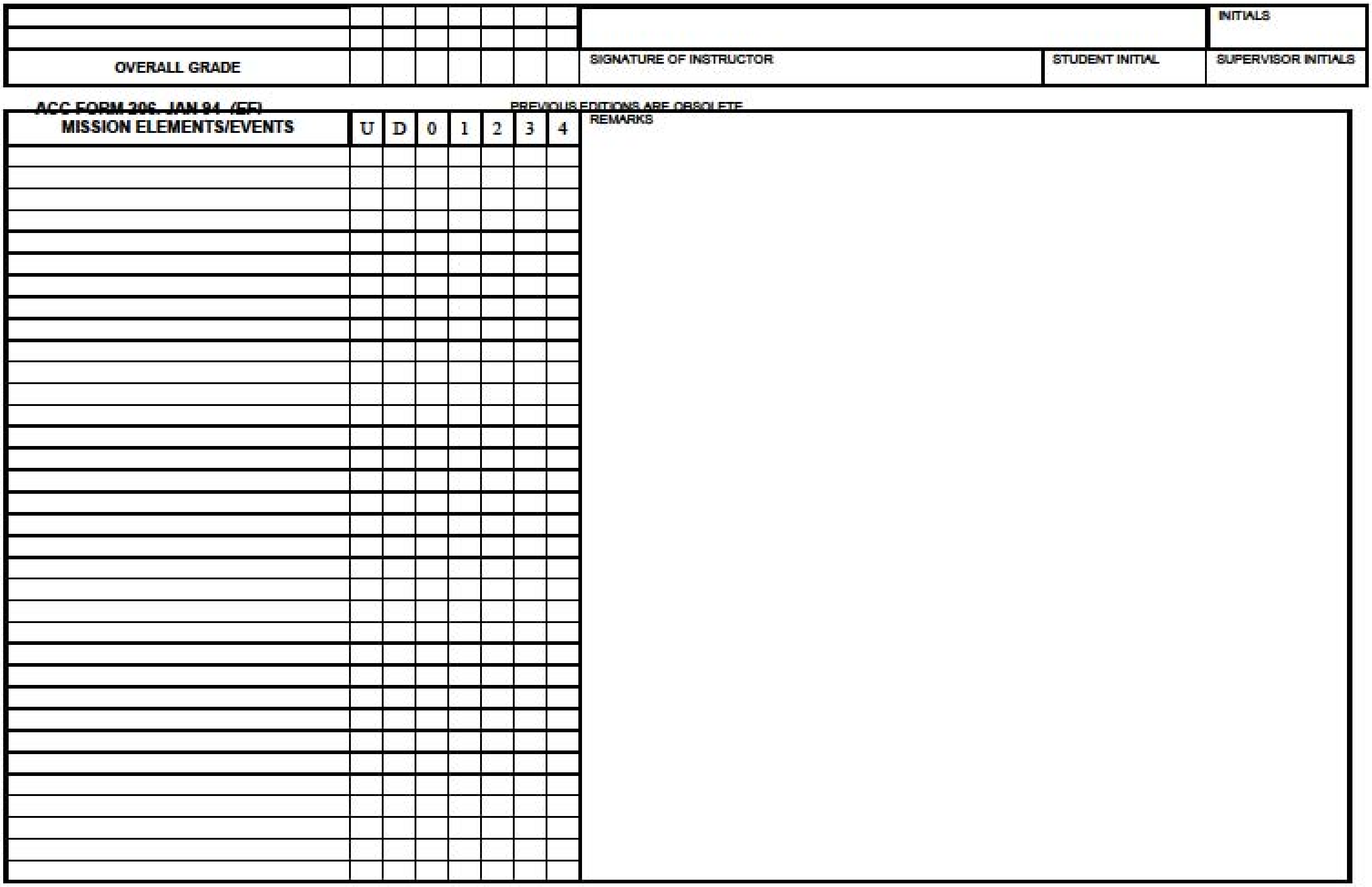
Mission Accomplishments:

Overall:

Specifics:

Emergency Procedure(s) covered:

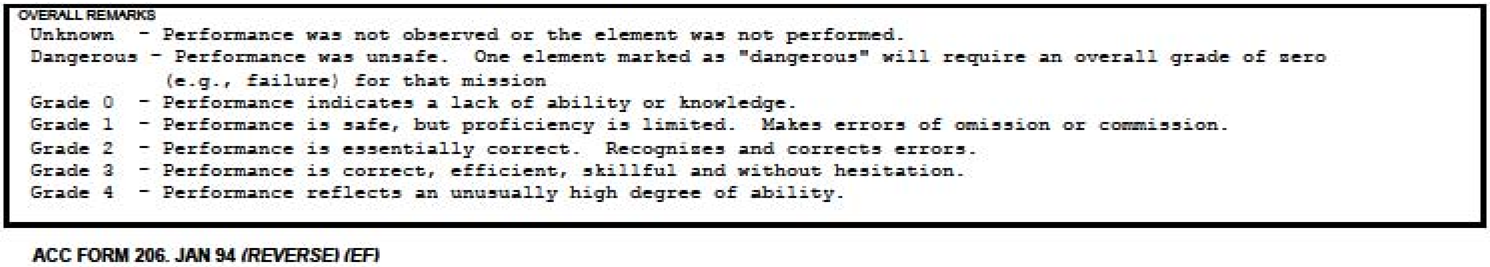
INSTRUCTOR



										INITIALS
OVERALL GRADE								SIGNATURE OF INSTRUCTOR	STUDENT INITIAL	SUPERVISOR INITIALS

ACC FORM 206, JAN 94 (EF) PREVIOUS EDITIONS ARE OBSOLETE

MISSION ELEMENTS/EVENTS	U	D	0	1	2	3	4	REMARKS



OVERALL REMARKS

Unknown - Performance was not observed or the element was not performed.

Dangerous - Performance was unsafe. One element marked as "dangerous" will require an overall grade of zero (e.g., failure) for that mission

Grade 0 - Performance indicates a lack of ability or knowledge.

Grade 1 - Performance is safe, but proficiency is limited. Makes errors of omission or commission.

Grade 2 - Performance is essentially correct. Recognizes and corrects errors.

Grade 3 - Performance is correct, efficient, skillful and without hesitation.

Grade 4 - Performance reflects an unusually high degree of ability.

ACC FORM 206. JAN 94 *(REVERSE) (EF)*