

Technical Report 1353

Delivering Training Assessments in a Soldier-Centered Learning Environment: Year Two

Michael Lodato

Jessie Hyland

Rebecca Mulvaney

ICF International

Randall Spain

U.S. Army Research Institute



December 2015

**United States Army Research Institute
for the Behavioral and Social Sciences**

Approved for public release; distribution is unlimited.

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved for distribution:

**MICHELLE SAMS, Ph.D.
Director**

Research accomplished under contract
for the Department of the Army

ICF International

Technical review by

Christopher Vowels, U.S. Army Research Institute
Steven Burnett, U.S. Army Research Institute

NOTICES

DISTRIBUTION: This Technical Report has been submitted to the Defense Information Technical Center (DTIC). Address correspondence concerning reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: DAPE-ARI-ZXM, 6000 6th Street (Bldg. 1464 / Mail Stop: 5610), Ft. Belvoir, Virginia 22060-5610.

FINAL DISPOSITION: Destroy this Technical Report when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) December 2015			2. REPORT TYPE Final			3. DATES COVERED (from. . . to) December 2012 - January 2014		
4. TITLE AND SUBTITLE Delivering Training Assessments in a Soldier-Centered Learning Environment: Year Two						5a. CONTRACT OR GRANT NUMBER W5J9CQ-11-D-0002		
						5b. PROGRAM ELEMENT NUMBER 622785		
6. AUTHOR(S) Michael Lodato, Jessie Hyland, Rebecca Mulvaney; Randall Spain						5c. PROJECT NUMBER A790		
						5d. TASK NUMBER 0005		
						5e. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) ICF International 9300 Lee Highway Fairfax, VA 22030						8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 6000 6 TH Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610						10. MONITOR ACRONYM ARI		
						11. MONITOR REPORT NUMBER Technical Report 1353		
12. DISTRIBUTION/AVAILABILITY STATEMENT; Distribution Statement A: Approved for public release; distribution is unlimited.								
13. SUPPLEMENTARY NOTES Contractor Officer's Representative: Dr. Steven Burnett								
14. ABSTRACT (<i>Maximum 200 words</i>): This research investigated the effectiveness of learner-centered technology-based training prototypes that were developed to test training concepts outlined in the Army Learning Model (ALM). The prototypes integrated training and assessments across mobile, virtual, and game-based training platforms. Approximately 120 Soldiers from Ft. Gordon, GA completed training on different versions of the prototypes to answer questions about the overall effectiveness of the prototype training, the value of using adaptive assessments, and the benefits of including interim assessments in the training. Results of the first experiment showed Soldiers who received the prototype training scored higher on measures of learning, transfer, and overall satisfaction than Soldiers who received traditional lecture-based instruction. Results of the second experiment showed adaptive assessments predicted training transfer better than non-adaptive assessments. Results of the final experiment showed students who received interim assessments scored better on measures of training transfer than students who did not receive these assessments. The practical and theoretical implications of these results are discussed along with lessons learned and implications for using technology-based training applications in support of the ALM.								
15. SUBJECT TERMS Training effectiveness, Training- Collaborative, Training Transfer, Computer Adaptive Testing								
SECURITY CLASSIFICATION OF						19. LIMITATION OF ABSTRACT Unlimited Unclassified	20. NUMBER OF PAGES 89	21. RESPONSIBLE PERSON Steven Burnett 334-155-9109
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified						

Technical Report 1353

Delivering Training Assessments in a Soldier-Centered Learning Environment: Tear Two

**Michael Lodato
Jessie Huang
Rebecca Mulvaney
ICF International**

**Randall Spain
U. S. Army Research Institute**

**Fort Benning Research Unit
Scott E. Graham, Chief**

December 2015

Approved for public release; distribution is unlimited.

DELIVERING TRAINING ASSESSMENTS IN A SOLDIER-CENTERED LEARNING ENVIRONMENT: YEAR TWO

EXECUTIVE SUMMARY

Research Requirement:

The U.S. Army Learning Model (ALM) sets a vision of Army learning that includes a more flexible, tailored, learner-centered system of learning (U.S. Army Training and Doctrine Command, 2011). This includes reducing instructor-led lectures and replacing them with a blended approach that incorporates virtual and constructive simulations, intelligent tutoring, gaming technology, and other technology-delivered instruction. A critical part of this vision is the inclusion of assessments in these technologies to tailor the training experience and ensure learning has occurred to a standard. However, questions regarding how best to design, develop, and otherwise integrate training and assessments within these platforms to maximize soldier training remain unanswered. In response to these needs, the U.S. Army Research Institute for the Behavioral and Social Science (ARI) developed prototype training applications and assessments to serve as a test-bed for conducting research on assessment strategies in maturing learning technologies. Included were mobile, virtual classrooms, and collaborative game-based technologies (Brusso et al., 2014). This report builds on that research by describing a set of three experiments that examined the effectiveness of the prototype training applications and assessments in an Army schoolhouse.

The goal of Experiment 1 was to examine the effectiveness of the prototype training with respect to the classroom-based training it is intended to replace. Within this overall research question we were interested in answering the following questions: 1a) Are reactions to the training prototypes at least as positive as reactions to the traditional classroom training; 1b) Do learners taking the prototype training acquire at least the same level of knowledge as learners taking the classroom-based training; 1c) Do learners who have taken the prototype training perform as well or better than learners who have taken the classroom-based training when using a real radio? The goal of Experiment 2 was to examine the potential benefits of using computer-adaptive tests (CATs) in an Army training context. Specifically, we examined whether adaptive assessments were more precise and better predictors of performance than non-adaptive assessments. The goal of Experiment 3 was to determine whether or not interim assessment during the mobile training and the virtual classroom resulted in higher levels of learning.

Procedure:

Using a quasi-experimental design, approximately 120 Soldiers from Ft. Gordon, GA completed different versions of the prototype training in order to answer questions about its effectiveness on learning outcomes, the value of using adaptive assessments, and the benefits of including interim assessments in the training. Four different types of measures were developed and delivered to Soldiers during the experiment: a demographic question, multiple choice pre- and post-test, a work sample task administered at the end of training, and reaction measures administered at the conclusion of each of the trainings in the prototype condition and the control condition.

Findings:

Results from the first experiment demonstrated significant benefits of the prototype training in several critical areas. First, participants in both the classroom training and in the prototype training demonstrated an increase in knowledge about the training content as a result of training. This change in knowledge was not significantly different between groups. Second, participants in the prototype condition performed better than their counterparts in the control condition on the work sample task. Specifically, participants in the prototype training condition completed the work sample task more accurately and quickly than those in the control condition. Finally, participants in the prototype condition reported more favorable ratings of training satisfaction and training utility than participants in the control condition. Results of Experiment 2 showed the CAT to be more precise than the non-adaptive test in measuring student ability. In addition, CAT scores had a stronger relationship ($r = .44$) with scores on the work-sample task than the non-adaptive assessment ($r = .29$). Results of Experiment 3 demonstrated mixed evidence that interim assessments contribute to learning. Soldiers who received interim assessments scored higher on the work sample test compared to Soldiers who did not receive these assessments. However, there were no differences between the two conditions with regard to post-test training scores.

Utilization and Dissemination of Findings:

The results of this research suggest that, when designed using principles of instructional design and the learning sciences, ALM-centric training offers an effective means for training Soldiers. Future research should continue to examine how training technologies can be used to support the educational principles outlined in the ALM. Training developers and leadership at TRADOC may benefit from the results of this study as they provide empirical evidence on the effectiveness of using training technologies to support the ALM. More specifically, researchers should empirically examine if macro-based adaptive training approaches, such as using pre-tests to tailor the training experience to the needs of individual learners, provides a significant benefit in terms of training satisfaction, learning, and performance over non-adaptive training strategies.

DELIVERING TRAINING ASSESSMENTS IN A SOLDIER-CENTERED LEARNING ENVIRONMENT: YEAR TWO

CONTENTS

	Page
INTRODUCTION	1
Overview of Training and Assessment Prototype System	2
EXPERIMENT 1: DISTRIBUTED LEARNING VERSUS CLASSROOM-BASED TRAINING	4
Method.....	6
Participants	6
Materials.....	6
Experimental Design	7
Procedure.....	8
Results	11
Equivalence of Groups	11
Research Question 1: Is the prototype training at least as effective as the traditional classroom training?.....	14
Discussion	18
EXPERIMENT 2: POTENTIAL BENEFITS OF USING CATS IN AN ARMY TRAINING CONTEXT	19
Method.....	20
Participants	20
Materials.....	20
Experimental Design and Procedure	20
Results	21
Equivalence of Groups	21
Overall Research Question 2: Does a CAT offer value over and above non-adaptive testing in a training context?	23
Discussion	25
EXPERIMENT 3: THE EFFECTS OF PERIODIC TESTING DURING TRAINING.....	27
Method.....	28
Participants	28
Materials.....	28
Experimental Design	28
Results	29
Equivalence of Groups	29
Research Question 3: Do interim assessments lead to better learning outcomes?	32
Discussion	36

CONTENTS (continued)

	Page
GENERAL DISCUSSION	36
Review of overall findings	36
Lessons Learned	37
Implications for Army Training	38
Conclusion	39
REFERENCES	40

APPENDICES

APPENDIX A: DEMOGRAPHICS QUESTIONNAIRE	A-1
APPENDIX B: PRE-TEST	B-1
APPENDIX C: POST-TEST	C-1
APPENDIX D: WORK SAMPLE TASK	D-1
APPENDIX E: REACTIONS MEASURES	E-1
APPENDIX F: PRIVACY ACT STATEMENT	F-1
APPENDIX G: INFORMED CONSENT	G-1
APPENDIX H: LIST OF ACRONYMS	H-1

TABLES

TABLE 1. MEANS, STANDARD DEVIATIONS AND T-VALUES ON EQUIVALENCE MEASURES	12
TABLE 2. SUMMARY OF LOGISTIC REGRESSION ANALYSIS FOR DEMOGRAPHIC VARIABLES PREDICTING GROUP ASSIGNMENT	13
TABLE 3. MEANS, STANDARD DEVIATIONS, AND TREATMENT EFFECTS ON PERCEIVED LEARNING, PERCEIVED LEARNING TRANSFER, AND SATISFACTION	15
TABLE 4. MEANS, STANDARD DEVIATIONS AND TREATMENT EFFECTS ON PRE-/POST-TEST SCORES AND DELTA	16

	Page
TABLE 5. MEANS, STANDARD DEVIATIONS AND TREATMENT EFFECTS ON WORK SAMPLE TASK	17
TABLE 6. MEANS, STANDARD DEVIATIONS AND T-VALUES ON EQUIVALENCE MEASURES	21
TABLE 7. SUMMARY OF LOGISTIC REGRESSION ANALYSIS FOR DEMOGRAPHIC VARIABLES PREDICTING TREATMENT CONDITION ASSIGNMENT	22
TABLE 8. MEANS, STANDARD DEVIATIONS AND T-VALUES ON PARTICIPANT SCORES AND SEM.....	24
TABLE 9. CORRELATIONS BETWEEN ABILITY ESTIMATES AND PRE AND POST-TESTS AND WORK SAMPLE TASK	24
TABLE 10. MEANS, STANDARD DEVIATIONS AND T-VALUES ON EQUIVALENCE MEASURES.....	30
TABLE 11. SUMMARY OF LOGISTIC REGRESSION ANALYSIS FOR DEMOGRAPHIC VARIABLES PREDICTING TREATMENT CONDITION ASSIGNMENT	31
TABLE 12. MEANS, STANDARD DEVIATIONS AND TREATMENT EFFECTS ON PRE-/POST-TEST SCORES AND DELTA	32
TABLE 13. MEANS, STANDARD DEVIATIONS AND TREATMENT EFFECTS ON WORK SAMPLE TASK.....	33
TABLE 14. MEANS, STANDARD DEVIATIONS, AND TREATMENT EFFECTS ON PERCEIVED LEARNING, PERCEIVED LEARNING TRANSFER, AND SATISFACTION.....	35

DELIVERING TRAINING ASSESSMENTS IN A SOLDIER-CENTERED LEARNING ENVIRONMENT: YEAR TWO

INTRODUCTION

The U.S. Army Research Institute for the Behavioral and Social Science (ARI) is currently supporting the Army's advancement of the U.S. Army Learning Model (ALM) (TRADOC, 2011). The ALM sets a vision of Army learning that includes a more flexible, tailored, learner-centered system of learning. The purpose of the current program of research was to create technology-based training prototypes that reflected the tenets of the ALM and to conduct research using these prototypes that can guide effective implementation across all institutional domains.

The ALM specifically calls for reducing instructor-led lectures and replacing them with a blended approach that incorporates virtual and constructive simulations, intelligent tutoring, gaming technology, and other technology-delivered instruction. These technologies have potential to facilitate deeper levels of learning and greater engagement than simple lectures or early versions of computer-based training (Cannon-Bowers & Bowers, 2010). For example, these technologies can give the learner control over the learning process. This control can enhance self-regulation, which is thought to be an important component of mastering information or skills (Smith, Ford, & Kozlowski, 1997). In addition, game-based training technologies, by definition, are designed to engage learners by maximizing interactivity, goals, competition, and feedback – often with a secondary goal of entertaining the learner (Cannon-Bowers & Bowers, 2010; Vogel, Greenwood-Ericksen, Cannon-Bowers, & Bowers, 2006). These technologies offer the capability to couch learning in more realistic settings. Virtual environments can allow learners to practice skills in a simulated setting before applying them in live situations. Mobile learning technologies enable access to training and information in real time, potentially allowing learners to access information in a live setting (Holden & Sykes, 2011; Norris & Soloway, 2004; Roschelle & Pea, 2002). Researchers are only beginning to investigate optimal use and effectiveness of these types of technologies for these purposes.

The ALM also stresses the importance of using valid and reliable assessment to ensure learning has occurred to a standard and as a key strategy to tailor training. The ALM specifically recommends the use of pre/post-tests to gauge learning, while still incorporating assessments throughout the training. A variety of assessment types beyond standard multiple-choice tests may be used to assess learning. For example, computer-adaptive tests (CATs) can be used to achieve better test precision with fewer items than non-adaptive multiple-choice tests. CATs have also been successfully demonstrated using mobile technologies (Triantafillou, Georgiadou, & Economides, 2008). In addition, simulated tasks (or “work samples”) offer a high fidelity option for measuring performance. This type of assessment requires trainees to perform tasks or work activities that mirror the tasks they would have to perform on the job and allow for the measurement of the application of skills rather than the acquisition of knowledge (Callinan & Robertson, 2000).

The vision described by the ALM seems promising; however, the Army has requested research to guide the utilization and integration of assessments as well as various learning technologies to achieve successful implementation of ALM within the institutional training domain. Questions regarding the effectiveness of using maturing technologies in a blended learning classroom need to be answered along with questions regarding the effectiveness and utility of different assessment techniques need to be addressed if the Army's vision for a Soldiers-centered learning environment is to succeed.

To meet this need, a two year research program was initiated to develop prototype training applications and assessments that align with the goals of the ALM. In the first year of the research program, prototype training materials and assessments were developed capable of testing ALM concepts in an integrated, technology-enabled environment (Brusso et al., 2014). The second year of the research program focused on conducting experimental studies with the prototype training and assessments to provide empirical evidence regarding the effectiveness of different training and assessment strategies highlighted in the ALM. The purpose of this report is to describe this research along with lessons learned and implications for using the prototype concepts to support future Army training. In the following sections, we briefly review the prototype training system developed during the first year of this research program. Then we report the background, methodology, and results for each of the three experiments. Finally, we conclude with a discussion of our results and lessons learned and implications for developing training that aligns with the goals of the ALM.

Overview of Training and Assessment Prototype System

In the first year of the research program, the research team identified pre-existing didactic-based training and transformed it into technology-based instruction that integrated training and assessments across mobile, virtual, and game-based technologies (see Brusso et al., 2014). The purpose of integrating the training across the three training devices was to leverage the unique affordance of each device and provide a proof-of-concept of how training and assessments could be delivered across multiple training platforms. The selected training content for the prototype concerned the use and operation of the Army Navy/ Portable Radio Communication (AN/PRC)-148 Joint Enhanced Multiband Inter/Intra Team (JEM) Radio. Training on the radio traditionally takes place over two days at the Signal Regimental Non-Commissioned Officer (NCO) Academy at Ft. Gordon, GA and includes approximately 90 PowerPoint slides, several hands-on activities, and an open-book test.

In transforming the classroom-based JEM training to mobile, virtual, and collaborative training, the research team proposed a highly integrated approach. Our general concept was to sequence and integrate the training across the different training platforms and to integrate assessments to track learning within and between modules. To achieve this we divided the training into three different phases that linked learning objectives to specific training platforms.

The first phase of training was delivered on a tablet (iPad) and focused on training Soldiers on the basic components, features, and functionality of the JEM radio. The training included a combination of didactic learning content and hands-on activities with a virtual radio that allowed users to manipulate simple features of the radio using the tablet's touchscreen interface. Assessments were presented throughout the training in formats ranging from

traditional multiple-choice, true / false, and matching questions, to exercises using a virtual JEM radio. The training also included a CAT that students completed on the tablet before and after the training. The purpose of the pre-test was to assess students' baseline knowledge of the radio, whereas the purpose of the post-test was to measure learning progress over the course of training (Brusso et al., 2014).

The second phase of training was delivered in a “virtual schoolhouse” style interface on a laptop. The training material covered in this phase of training addressed more complex skills such as advanced programming and troubleshooting procedures. This training was instructor-led and allowed for the instructor to select and interact with training content that was delivered synchronously to students' workstations. While in the classroom, students could communicate through text-based chat and voice communication with their instructors and peers. This phase of training also included intermittent assessments that students completed individually and a CAT that students completed at the conclusion of this phase of training (Brusso et al., 2014).

The third phase of training was delivered in a game-based training environment. The training extended features of the first and second phases and included realistic scenario-driven vignettes that required students to work in teams to collaboratively resolve issues related to manipulating and troubleshooting a virtual JEM radio. The training scenario included both group and individual level assessments. Groups were assessed as a whole during the vignettes where the overall score was determined by the number of successful radio actions completed by each team member. Thus team performance was an aggregate score of group performance. In addition, players completed individual knowledge-based assessments as they progressed through the collaborative scenario. These items were scored at the individual level (Brusso et al., 2014).

Performance data and training scores from the CAT, the interim assessments, and the collaborative scenario were collected and stored in a prototype learning management system. Figure 1 provides an overview of the prototype concept.

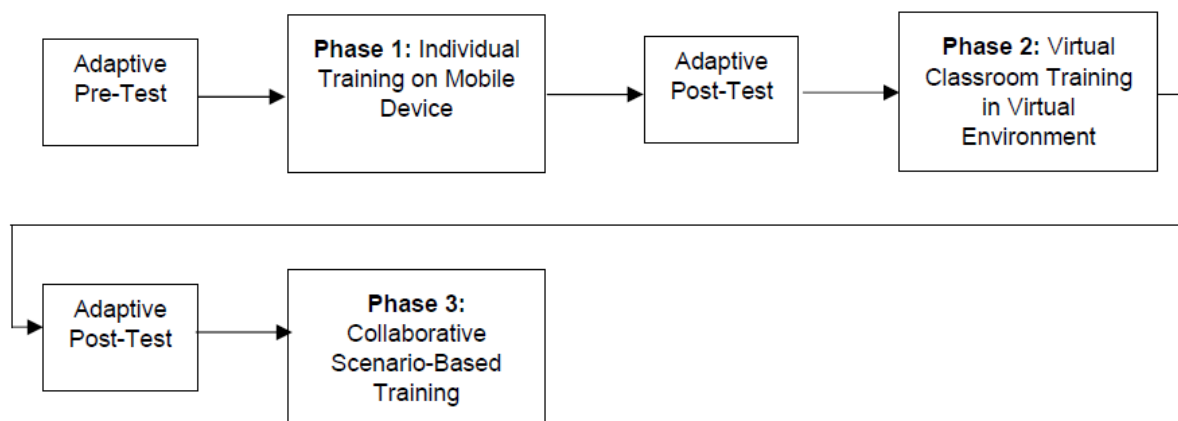


Figure 1: Prototype training concept. The adaptive pre-test and first adaptive post-test were delivered on a tablet; the second adaptive post-test was delivered on a computer. Phases 1 and 2 included checks on learning, and Phase 3 included individual and group assessment.

The prototype training was intended to serve as a test-bed for conducting research on training and assessment strategies in support of the ALM. Specifically, it was intended to address the following questions:

- What is the effectiveness of distributed learning technologies versus classroom-based instruction on training outcomes such as satisfaction, learning, and performance,
- What are the potential benefits of using CATs in an Army training context, and
- What is the value of including “check on learning” items in distributed learning activities?

In the sections that follow, we provide a description of three experiments designed to answer these research questions.

EXPERIMENT 1: DISTRIBUTED LEARNING VERSUS CLASSROOM-BASED TRAINING

While distributed learning is not a new concept, it has continued to gain attention in the training literature as internet service, laptops, and mobile computing devices have become more prevalent. The continuum of distributed learning practices has increased in levels of flexibility, interactivity, and accessibility to instructional materials and content (Taylor, 2001). In recent years, at least three meta-analyses have been conducted to compare distributed learning and traditional classroom methods with respect to a wide range of learning outcomes (Bernard et al., 2004; Sitzmann, Kraiger, Stewart, & Wisher, 2006; Zhao, Lei, Yan, Lai, & Tan, 2005). In all cases, the differences in learning outcomes for distributed learning versus live classroom-based learning were not statistically significant. Means, Toyama, Murphy, Bakia, and Jones (2009) expanded on these findings by using meta-analysis to compare “blended” learning (i.e., both on-line and live classroom components) to face-to-face instruction. Their findings demonstrated that on-line/blended learning had a greater mean effect size over live classroom education (+0.24,¹ $p < .001$).

In comparing distributed learning to classroom-based learning, Zhao et al. (2005) also examined contributing factors related to the degree of heterogeneity among the studies in their review. One goal was to determine why some distance learning studies yielded far more positive outcomes than others. Zhao et al., (2005) found that several factors contributed to this. For example, they found that difference in content area was a significant predictor for differences. Programs in business, computer science, and medical science, in particular, showed greater effect sizes for distributed learning than traditional classroom training. That is, these content areas were more effectively taught using distance learning. Military training and mathematics also showed significantly greater effect sizes, although these studies typically included very small sample sizes (Zhao et al., 2005). They also found that type of interaction (i.e., synchronous,²

¹ Typically, an effect size of approximately 0.20 is considered small, 0.50 is considered a medium effect, and approximately 0.80 is considered a larger effect (Cohen, 1992).

² Synchronous distance learning refers to learning that occurs when the instructor and learners interact in different places but during the same time.

asynchronous³) contributed to effect size differences between distributed learning and face-to-face learning. Specifically, those distributed learning efforts employing both synchronous and asynchronous components showed a larger effect size (+0.22,⁴ $p < .01$).

In addition to learning and performance outcomes, researchers tend to agree that engagement in distributed learning, especially videogame-based training environments or virtual worlds (such as the multi-user virtual environment [MUVE] *Second Life*) can lead to positive affective outcomes, such as increases in trainee motivation and engagement (Chang, Gütl, Kopeinik, & Williams, 2009; Mautone, Spiker, Karp, & Conkey, 2010; Topolski et al., 2010). These positive reactions can play an important role in subsequent training transfer (e.g. Colquitt, LePine & Noe, 2000).

The prototype training developed for this effort combines asynchronous mobile learning with synchronous training in a virtual environment. Although there is considerable evidence that these forms of training should be at least as effective as classroom-based instruction, little research has been conducted in a military setting.

The main goal of Experiment 1 was to examine the effectiveness of the prototype training in terms of learning, performance, and overall satisfaction – namely we were interested in determining if the prototype training was at least as effective as the traditional classroom-based training for teaching Soldiers the fundamental concepts and procedures for operating the JEM radio. Within this overall research question we were interested in answering the following questions:

- 1a) Are reactions to the training prototypes at least as positive as reactions to the traditional classroom training?
- 1b) Do learners taking the prototype training acquire at least the same level of knowledge as learners taking the classroom-based training?
- 1c) Do learners who have taken the prototype training perform as well or better than learners who have taken the classroom-based training when using a real radio?

Based on previous research, we expected reactions to the training prototypes to be at least as positive as reactions to the traditional classroom training (e.g., Chang et al., 2009; Mautone et al., 2010; Topolski et al., 2010); we expected participants in the prototype training condition to score the same as or above classroom-based participants on knowledge acquisition (e.g., Bernard et al., 2004; Sitzmann et al., 2006; Zhao et al., 2005); and we expected learners who received the prototype training to perform as well or better on a hands-on work sample task than learners who received the classroom-based training (e.g., Bernard et al., 2004; Sitzmann et al., 2006; Zhao et al., 2005).

³ Asynchronous distance learning occurs when the instructor and the learners interact in different places and during different times.

⁴ Typically, an effect size of approximately 0.20 is considered small, 0.50 is considered a medium effect, and approximately 0.80 is considered a larger effect (Cohen, 1992).

Method

Participants

A total of 65 Soldiers from Fort Gordon, GA participated in the experiment. Participants were recruited from the Regimental Non-Commission Officer Academy's (NCOA) Advanced Leadership Course (ALC) and Senior Leadership Course (SLC). ALC is branch-specific course for Soldiers selected for promotion to staff sergeant that focuses on preparing leadership and technical skills required to effectively lead squad/platoon size units. SLC is also a branch-specific course that provides an opportunity for Soldiers selected for promotion to sergeant first class to acquire the leader, technical, and tactical skills, knowledge, and experience needed to lead platoon/company size units. Participants had an average tenure of 10 years in the military and had an average of 3 deployments. Participants were all from the military occupational specialties (MOS) of 25U (Signal Support Specialists) with ranks from E5 through E7 and an average of 29 months in rank. The control condition (i.e., traditional classroom training) included 32 Soldiers, 23 enrolled in ALC, and 9 enrolled in the SLC. The treatment condition (i.e., prototype training) included 33 Soldiers, 20 from ALC and 13 from SLC. Data for the control and treatment conditions were collected separately during two different course sections that were several weeks apart. The first course section was used to collect data for the control condition and the second for the treatment condition.

Materials

Four different types of measures were developed and delivered to Soldiers during the experiment: a demographic question, a multiple choice pre- and post-test, a work sample task administered at the end of training, and reaction measures administered at the conclusion of each of the training modules in the prototype condition and at the end of the classroom-based training in control condition.

Demographic questionnaire (Appendix A). The demographic questionnaire was designed to collect background information on each Soldier including time in military (years), rank, time in rank (months), military occupational specialty (MOS), and number of deployments. The questionnaire also included a number of items to assess Soldiers' prior experience with the JEM and similar handheld radios (i.e., the AN/PRC-148 Multiband Inter/Intra Team Radio; MBITR) including questions on participants experiencing using the JEM radio while deployed, perceived expertise with the JEM radio, experience loading keys, and confidence performing various tasks on the JEM radio. The format of these questions varied, some were open-ended, some included Likert-type response scales, and others included yes/no responses.

Pre-test and post-test (Appendices B and C). To assess the baseline level of JEM-related knowledge, we developed a pre-test that was administered prior to beginning training. This test consisted of 14 multiple-choice questions and 11 open-ended questions about the JEM radio. The purpose of including the open-ended items was to further assess baseline knowledge to ensure all groups were entering the experiment with roughly the same level of knowledge and skill using the JEM. The post-test was a parallel form of the pre-test with 14 multiple-choice questions that are different from those included in the pre-test. Items for both tests were drawn

from the CAT item bank and the same content balancing strategy (i.e., each test consists of similar number of items from each content area). A description of the CAT item development process is provided in Brusso et al. (2014). Items with comparable levels of difficulty, based on pilot test data, were selected to ensure the pre- and post-tests were parallel forms. No open-ended questions were included in the post-test.

Work sample task (Appendix D). The work sample task was a structured, timed, hands-on exercise that required participants to perform various tasks using an actual JEM radio. The measure consisted of five tasks based on the learning objectives covered in the training content: preventive maintenance checks and services (PMCS), keyfill, channel programming, cloning, and zeroizing the radio. Each task consisted of setup instructions for the test administrator, equipment requirements, instructions for the test-taker, a scoring rubric and a time limit. Scoring of the work sample task included a behavioral checklist where Soldiers earned points for successfully completing procedures for each task (each task consisted of multiple procedures). In addition, several tasks included follow-up questions that were asked verbally by the test administrator. Each follow-up question was worth one point. These questions, and potential answers, were included in the scoring rubric. If Soldiers failed to complete a task within the given time limit, the test administrator awarded zero points for the remaining procedures under the task. In total, Soldiers could earn up to 26 points in completing the work sample task.

Reaction measures (Appendix E). Reaction questionnaires were developed to capture participants' opinions of the prototype and classroom-based training. The reaction questionnaires included a combination of items, including questions specific to the training prototypes and general questions that were asked upon completion of the entire training event. Items addressed several different dimensions including usability, comfort/familiarity with technology, content/ISD, engagement/attention during training, technical issues, interacting (virtual classroom only), instructor (virtual classroom only), perceived learning, perceived learning transfer, satisfaction, and open-ended user comments. The item response format for these questions varied from simple yes/no responses, to 5-point Likert scales, to open-ended questions.

Equipment. Two laptops were used to maintain the server and local network and 20 additional laptops were used for administering the virtual classroom and collaborative scenario. Ten iPads running iOS 6 were used for administering the mobile training. Six JEM radios were utilized for administering the work sample task at the end of each condition.

Experimental Design

The current experiment used a quasi-experimental non-equivalent control-group treatment-group design (Campbell & Stanley, 1963). This type of design differs from a true experimental design in that participants are not randomly assigned to conditions. Rather, groups that constitute a naturally assembled collective, such as classes, are randomly assigned to a condition. In this experiment, classes were randomly assigned to either the control condition or treatment condition. We confirmed experimental equivalence by examining demographic information and pre-test scores.

Soldiers in the control condition participated in the traditional classroom training that included a half day, in-person didactic presentation, followed by a half day practical exercise

portion where Soldiers performed various tasks using an actual JEM radio. Soldiers in the treatment condition completed the prototype mobile, virtual classroom, and collaborative scenario trainings instead of the classroom training. Each condition spent an equivalent amount of time in training. Participants in the experimental condition spent approximately an hour and a half to three hours in the first phase of training (i.e., tablet), three hours in virtual classroom training, and two hours in collaborative training, for a total of eight hours of training time, on average. Participants in the control condition also spent approximately four hours in the lecture portion of training and two to three hours in the hands-on portion of training. The purpose of this design was to examine whether the prototype training was at least as effective as traditional classroom-based training, as measured by performance on a hands-on work sample task administered at the conclusion of each condition.

Procedure

Data collection for this experiment was carried out over two separate trips to Fort Gordon, GA. During the first trip, we administered the control condition to a group of Soldiers from ALC and SLC. Several weeks later, when new classes were in session, we returned to Fort Gordon and administered the experimental condition. Both conditions began with an introduction to the research given by a member of the research team. During this introduction participants reviewed the privacy act statement (Appendix F) and read and signed the informed consent (Appendix G). Soldiers then completed the demographic questionnaire and the paper-based pre-test before beginning their respective training conditions.

Control condition. Soldiers in the control condition participated in the traditional classroom-based JEM training offered at Fort Gordon. This training included a half-day of lecture-based instruction in which the instructor taught from a set of PowerPoint Slides and asked periodic check-on-learning questions. After a lunch break, the Soldiers returned for the practical exercise portion of the training, where they all met in a lab and had the opportunity to practice using the radios under the supervision of the same instructor. At the conclusion of the day, the Soldiers completed the reaction measure. The final post-test and work sample task were then administered to each participant individually over the course of the next two days. Each participant was met by a researcher, who first handed out the post-test to be completed and then administered the work sample task and evaluated the participant's performance. The post-test consisted of 14 item multiple-choice questions regarding the functionality of the JEM radio. Directly following this measure, participants completed the work sample task. Each work sample task session lasted approximately 45 minutes. Upon completion participants were debriefed on the research and their participation concluded. Figure 2 provides a summary of the schedule of events for the control condition.

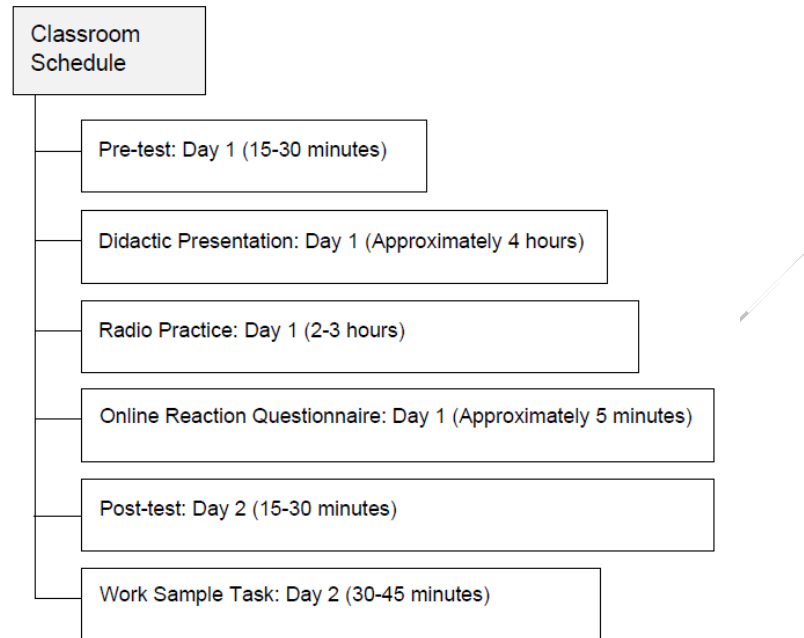


Figure 2: Traditional classroom-based training schedule.

Treatment condition. Students in the treatment condition participated in each of the prototype trainings (i.e., mobile, virtual classroom, collaborative scenario) developed for this research. Similar to the control condition, participants first reviewed the informed consent and completed the demographic questionnaire and pre-test. Then students began the mobile portion of training on an iPad which started with a 12 item CAT. After the CAT, students progressed through a number of lessons covering different topics regarding the components, features, and functionality of the JEM radio. Lessons also included interim check on learning activities that were designed to reinforce the newly learned principles. At the end of the mobile training, participants completed another 12 item CAT on the iPad. Following this, Soldiers completed a reaction questionnaire to capture their overall opinions about the prototype mobile training.

Next, Soldiers participated in the virtual classroom training. Due to the number of Soldiers in this condition and the limited number of computers available, the virtual classroom was administered twice (on two separate days). Soldiers were randomly assigned to participate in the virtual classroom on one of the two days. The instructor for all virtual classroom sessions was a member of the experimental staff who was familiar with the training content.⁵

⁵ A male staff member who was key to the prototype content development was used as the instructor for the virtual classroom. Prior to conducting the live sessions, this staff member was also extensively trained on the course content, which was scripted, as well as the use of the instructor platform. This was done to minimize the risk of the treatment condition being confounded with instructor usability issues due to limited time for instructor training on-site. However, the course instructor was present at the instructor station with our staff member during the virtual classroom sessions to ensure the accuracy and consistency of the training content presented and to respond to student questions.

The virtual classroom training extended aspects of the mobile training to include the ability to share and deliver content in a ‘virtual schoolhouse’ style interface. For this portion, the instructor led the virtual classroom course using text chat, voice over internet protocol (IP), and had the ability to both manipulate a virtual radio and share content with students in the classroom. Students could communicate with the instructor, communicate with each other, and interact with learning materials throughout the course. Similar to the mobile training, participants also completed interim check on learning activities to reinforce the newly learned material. Students were encouraged to take brief breaks after every two lessons; in all, the virtual classroom portion of training lasted approximately three hours. At the end of the virtual classroom, Soldiers completed the CAT for the third and final time and then completed a reaction measure. Both measures were delivered via the laptop on which Soldiers participated in the virtual classroom.

The final phase of the training was the collaborative scenario. For this portion of training, participants were randomly assigned to groups of three and engaged in an interactive “day in the life” exercise that focused on use of the radio in a scenario-based exercise (Brusso et al., 2014). Each team member sat at their own workstation and worked together collaboratively to complete a series of seven short vignettes. Each vignette focused on different procedures and tasks for operating the radio. The control of the radio rotated among each participant between vignettes so that only one team member was the active user of the radio at a time. As noted previously, all players saw the same screen but only the “active” player was able to “use” the radio while the other two players were in “view only” mode and offered advice and feedback for the task. Participants were scored based on whether their team successfully performed the tasks required for each vignette (i.e., the team receives 1 point for every task completed correctly within the time limit). In addition, participants were scored on individual check on learning items that were automatically delivered and scored between vignettes. Following participation in the collaborative scenario, Soldiers again completed a reaction questionnaire.

After completing all three phases of the prototype training (mobile, virtual classroom, and collaborative scenario) participants completed a 14-item multiple choice post-test and then completed the work sample task. The work sample task required participants to perform a series of five tasks on an actual JEM radio. This task took approximately 45 minutes to complete. Upon completion, participants were debriefed on the research and excused from the experiment. Figure 3 contains a summary of the schedule of events for the treatment condition.

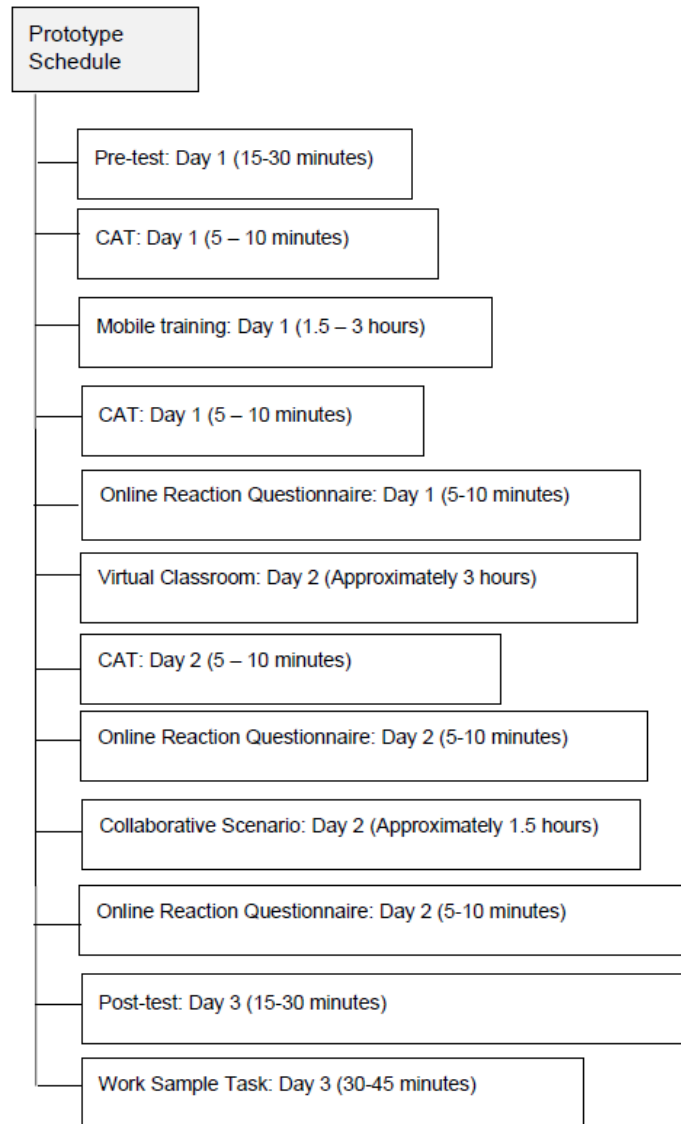


Figure 3: Prototype training schedule.

Results

Equivalence of Groups

A key assumption in the experimental design employed in our study was that the treatment and control groups were comparable prior to treatment. That is, we assumed participants in each condition were similar to each other even though they were not randomly assigned to conditions. The experimental control becomes more effective when the similarity between the treatment and control groups is confirmed by pre-test scores (Campbell & Stanley, 1963, p. 47-48). The first step in our analysis was to check this assumption by examining the equivalence of the control and treatment groups. We compared scores between the two groups on the knowledge-based pre-test, participant experience in using the MBITR/JEM (i.e., whether the participant has experience using an MBITR or JEM and amount of experience in months),

participant MBITR/JEM expertise (i.e., self-rated level of expertise using an MBITR/JEM on a 4-point Likert scale), and participant experience loading keys (i.e., self-rated level of experience loading keys on an MBITR/JEM on a 4-point Likert scale).

Two sets of statistical analyses were conducted to test for the pre-treatment equivalence between the two conditions. First, a series of t-tests were employed to compare the group means of pre-test scores, participant experience and expertise with the MBITR/JEM, and participant experience in loading keys. No significant differences between the two groups on these variables were found. Means and standard deviations for each variable are provided in Table 1.

Table 1

Means, Standard Deviations and t-Values on Equivalence Measures

	Mean (SD)		<i>t</i>
	Control (N = 31 ⁺)	Treatment (N = 33)	
Pre-Test Score*	6.48 (2.28)	7.39 (1.89)	-1.75
Experience with MBITR/JEM (Yes/No)	0.90 (0.30)	0.91 (0.29)	-.08
Experience with MBITR/JEM (in months)	30.64 (26.74)	30.33 (30.29)	.01
MBITR/JEM Expertise**	2.93 (0.98)	2.82 (0.92)	.49
Experience Loading Keys***	3.35 (0.91)	3.30 (0.85)	.24

*Score is out of a possible total of 14;

**Self-rating on a 4-point Likert scale: 1 = Have never used before; 2 = Have used for only standard communication; 3 = Have performed simple troubleshooting in usual circumstances; 4 = Have performed advanced troubleshooting in unusual circumstances.

*** Self-rating on a 4-point Likert scale: 1 = Have never performed or seen it performed; 2 = Have never performed but have seen it performed; 3 = Have performed but not as part of usual duties; 4 = Have performed as part of usual duties.

⁺ One participant in the control condition was a no-show for the demographic questionnaire and the pre-test but participated in the rest of the program.

Second, the condition assignment (i.e., control vs. treatment) was regressed on a number of demographic variables through a binomial logistic regression model to examine whether the condition assignment could be explained by any systematic differences between the groups. Demographic variables entered into the model included rank, number of deployments, MBITR/JEM experience while deployed, experience and expertise with the MBITR/JEM, and experience in loading keys. Results of the logistic regression analysis are presented in Table 2. As expected, neither the overall regression model nor any of the individual predictors were found to be statistically significant, thus supporting the random assignment assumption.

Table 2
Summary of Logistic Regression Analysis for Demographic Variables Predicting Group Assignment

Predictor	<i>B</i>	<i>SE B</i>	<i>Wald</i>	<i>df</i>	<i>p</i>	<i>e^B</i>
Rank	-.91	.56	2.69	1	.10	.40
Number of Deployments	-.12	.24	.26	1	.61	.89
MBITR Experience While Deployed	-1.14	.94	1.48	1	.23	.32
JEM Experience While Deployed	-.23	.77	.08	1	.77	.80
Experience with MBITR/JEM (Yes/No)	.22	1.54	.02	1	.89	1.24
Experience with MBITR/JEM (in months)	.00	.01	.12	1	.73	1.00
MBITR/JEM Expertise	-.03	.51	.00	1	.96	.98
Experience Loading Keys	.24	.47	.27	1	.61	1.28
Constant	2.16	1.61	1.80	1	.18	8.69
χ^2					6.02	
<i>df</i>					8	
Overall % Predicted Correct					58.7	

Note: e^B = exponentiated *B* (odds ratio⁶)

In addition to establishing equivalence between the two conditions, we also examined predictors of the pre-test score to further explore any systematic differences existing in the overall sample prior to treatment. A linear regression analysis revealed that the only significant predictor was self-rated MBITR/JEM expertise, explaining 7.3% of variance in pre-test scores ($F = 4.81, p < .05$). As expected, participants who rated themselves higher on prior MBITR/JEM expertise scored better on the pre-test ($\beta = .27, p < .05$). However, as mentioned previously, there

⁶ An odds ratio less than one means lower odds of outcome associated with increase in the associated predictor; an odds ratio greater than one means higher odds of outcome associated with increase in the associated predictor.

was no difference in pre-test score or MBITR/JEM expertise between the two conditions. Course membership (i.e., ALC vs. SLC) also did not emerge as a significant predictor of pre-test score. This further supports the conclusion that our sampling and condition assignment strategy was adequate in meeting the assumptions required of the experimental design.

Research Question 1: Is the prototype training at least as effective as the traditional classroom training?

A primary goal of this experiment was to better understand the impact of transitioning classroom-based training to a virtual format, utilizing mobile training technologies and virtual world technologies. Specifically, we first compared reactions to the prototype training to reactions to the classroom-based training. In addition, we used pre-/post-test scores to compare overall learning in the classroom setting to overall learning during the prototype training. Finally, we examined scores on a work sample task to determine whether or not there was a difference in *transfer* of learning to the use of an actual JEM radio. Results of these comparisons are presented below.

Participant reactions. There were three dimensions on the general reaction questionnaire (administered at the end of both conditions) that could be used to directly compare between the control and treatment conditions: satisfaction, perceived learning and perceived learning transfer. Table 3 shows the comparisons of mean ratings and standard deviations for each reaction survey item by condition. Results of the regression analyses comparing the two conditions are also included in this table.

Table 3

Means, Standard Deviations, and Treatment Effects on Perceived Learning, Perceived Learning Transfer, and Satisfaction

Dimension/Item	Mean (SD)		r^2 Treatment (r^2 change with MBITR/JEM Expertise)
	Control (N=32)	Prototype (N=33)	
Perceived Learning			
1. Learning this material was fun.	3.38 (1.13)	4.24 (.90)	14%**
2. Overall, I have learned a lot from this training. ⁺	2.72 (.85)	3.88 (.89)	29%*** (6%*)
Perceived Learning Transfer			
1. It is clear to me that the people conducting the training understand how I will use what I learn.	2.50 (.76)	3.82 (.92)	38%***
2. This training was relevant to my job in the Army. ⁺	1.88 (.71)	4.33 (.69)	76% *** (2%*)
3. I believe the training will help me do my current job in the Army better. ⁺	2.47 (1.047)	3.94 (1.17)	30%*** (6%*)
4. I learned something I can apply immediately to my work in the Army.	2.56 (1.01)	3.82 (1.07)	26%***
5. I plan to use what I learned on my job in the Army. ⁺	2.22 (.94)	4.06 (.97)	48%*** (7%**)
6. I am prepared to train other Soldiers on what I learned in this training.	2.09 (.96)	4.09 (.95)	55%***
7. I get excited when I think about trying to use my new learning on my job in the Army.	2.63 (.98)	3.88 (1.11)	26%***
8. I will be using the equipment on my job in the Army after the training. ⁺	2.19 (.93)	4.00 (1.03)	49%*** (5%*)
9. The training was of practical value to me. ⁺	2.19 (1.00)	4.03 (1.05)	44% *** (4%*)
Satisfaction			
1. I enjoyed this training program.	2.78 (1.01)	4.09 (1.13)	26%***
2. My time on the training was well spent.	2.81 (1.08)	4.03 (1.21)	20%***
3. I would recommend this training program to other Soldiers.	2.63 (1.16)	3.21 (.74)	7%*

⁺Items that were also predicted by self-rated MBITR/JEM expertise.

* $p < .05$;

** $p < .01$;

*** $p < .001$.

To test whether participants reacted differently to the traditional classroom training and the prototype training, we employed a series of hierarchical regression models with reaction survey items as the criterion and the condition assignment as the predictor entered during the first step. Additional control variables entered into the regression model during the second step as predictors included course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys. Overall, treatment condition was the only significant predictor of all reaction survey items on the three comparable dimensions (i.e., learning, transfer, satisfaction), explaining approximately 35% of variance on average. As shown in Table 3, participants in the prototype condition reported much more favorable attitudes across all three dimensions. The percent of variance explained by treatment alone ranged from 7% to 76% with an average of 35% across all items. For some items (see Table 2), especially those in the perceived learning transfer dimension, MBITR/JEM expertise also emerged as a significant predictor. However, MBITR/JEM expertise was found to be inversely related to learning and transfer. Participants with more prior MBITR/JEM expertise reported lower levels of learning and transfer, as expected. The effect of MBITR/JEM expertise as a predictor was much smaller compared to treatment condition, only explaining an additional 5% of variance on average.

Pre-/Post-test scores. Another goal of Experiment 1 was to examine whether or not participants increased their understanding of the JEM radio as a result of the prototype training and whether understanding increased at least as much as in the control group. Scores on the pre- and post-tests were used to examine the knowledge gain as a result of the training. Means and standard deviations on the tests are provided in Table 4. Results of the regression analyses comparing the two conditions are also included in this table.

Table 4

Means, Standard Deviations and Treatment Effects on Pre-/Post-test Scores and Delta

	Mean (SD)		r^2 Treatment
	Control (N=31)	Treatment (N=33)	
Pre-Test Score	6.48 (2.28)	7.39 (1.89)	NS
Post-Test Score	8.78 (2.54)	10.12 (2.12)	7%*
Delta	2.29 (2.77)	2.73 (2.53)	NS

* $p < .05$

We first examined the training effects on participants' understanding of the subject matter regardless of treatment condition. A paired sample t-test revealed a significant increase from the pre-test score to the post-test score across both groups ($t = 7.63, p < .001$). Next, we examined the treatment effects on the post-test scores through a linear regression model. In addition to treatment condition, control variables, including course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys, were also entered into the equations as predictors. The only statistically significant predictor of post-test score was treatment condition, explaining approximately 7% of variance. Participants in the treatment condition scored higher

on the post-test than their counterparts in the control condition ($\beta = .267, p < .05$). Lastly, we employed the same regression model with delta (i.e., pre-/post-test score change) as the outcome variable and the results were not statistically significant. We also calculated Cohen's d (Cohen, 1992) to compare the delta of the control to the delta of the treatment condition. Results of this calculation suggest that the size of the effect was very small ($d = .17$).⁷

Work sample task. The last goal of this experiment was to determine whether or not participants in the treatment condition would perform as well as participants in the control condition on a work sample task using an actual JEM radio. To test this question, we compared final scores and task completion time on the work sample task between the treatment and control conditions. Means and standard deviations on the work sample task, results of the regression analyses comparing between the two conditions, and effect sizes, are provided in Table 5.

To test for the treatment effect on the work sample task, we conducted regression analyses to examine whether or not treatment condition predicted work sample task completion scores or completion time. Consistent with the previous tests, control variables, including course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys, were also entered into the equations as predictors in addition to treatment condition. Treatment condition and self-rated MBITR/JEM expertise were found to be the only statistically significant predictors of both work sample task measures. Participants in the treatment condition scored higher on the work sample task ($\beta = .488, p < .001$) and completed it faster ($\beta = -.263, p < .05$) than their counterparts in the control condition. In addition, participants with higher MBITR/JEM expertise also scored higher on the work sample task ($\beta = .337, p < .01$) and completed the tasks faster ($\beta = -.338, p < .01$). We also calculated Cohen's d to measure the effect size between conditions for both measures of the work sample task. Results indicate that there was a large effect size for task completion score and a medium-sized effect for task completion time.

Table 5

Means, Standard Deviations and Treatment Effects on Work sample task

	Mean (SD)		r^2 Treatment (r^2 change with MBITR/JEM Expertise)	Cohen's d
	Control (N=32)	Treatment (N=33)		
Task Completion Score ⁺	16.91 (4.12)	20.79 (2.78)	22%*** (11%**)	1.13
Task Completion Time (in minutes) ⁺⁺	22.71 (3.38)	20.80 (3.52)	6%* (9%*)	.56

* $p < .05$;

** $p < .01$;

*** $p < .001$.

⁺ Score is out of a possible total of 26;

⁺⁺ The limit for task completion time is 30 minutes.

⁷ Typically, an effect size of approximately 0.20 is considered small, 0.50 is considered a medium effect, and approximately 0.80 is considered a larger effect.

Discussion

Our results demonstrated that the prototype training was at least as effective as the classroom training in several critical areas. First, we found that participants in both the classroom training and in the prototype training demonstrated an increase in knowledge about the JEM as a result of training. There was also a small, but significant difference in post-test scores with those in the prototype training scoring slightly higher; however, the change in knowledge was not significantly different between groups. This finding indicates that both modes of training were an effective means of introducing a basic understanding of JEM operation.

Second, we found that participants in the prototype condition performed better than their counterparts in the control condition on the hands-on work sample task. Participants in the prototype training completed tasks on the JEM more accurately and more rapidly than those in the classroom training. These results are promising given the operation of the JEM involves a psychomotor component. One important concern about the prototype training was the possibility of negative training because students were never in contact with the actual physical equipment during the training. If the simulated model of the JEM radio did not contain enough psychological and physical fidelity, there was a risk that participants would practice carrying out a task incorrectly. In such a case, training could do more harm than good. In this study, we did not find any results that would suggest training with the prototype technologies led to negative transfer. Instead, results of our study show promises of using simulated technologies for training and training transfer in cases where gaining access to physical equipment is limited.

Finally, we found that participants who received the prototype training indicated they were more satisfied with the training than their counterparts who received traditional classroom training. Participants in the treatment condition reported higher ratings of perceived learning, perceived transfer, and training satisfaction. This is consistent with the extant literature that shows distributed learning technologies have a positive impact on a range of trainee attitudes (e.g., Chang et al., 2009; Mautone et al., 2010; Topolski et al., 2010).

An important limitation of this experiment was our inability to utilize the technologies in the truest sense of a distributed learning setting; rather we replicated the experience in a classroom setting in which all the students were collocated with each other. The experimental technologies were intended to allow for training anywhere, anytime. The fact that participants were sitting in the same room with an instructor nearby – in a very controlled setting – limits the generalizability of these findings. Ideally, further research should include implementation of the prototype in a true field setting where participants would complete the mobile training from home (or another location of their choice) and on their own time and the virtual classroom and collaborative scenario would include participants who are (at least minimally) geographically dispersed.

EXPERIMENT 2: POTENTIAL BENEFITS OF USING CATS IN AN ARMY TRAINING CONTEXT

The ALM specifically calls for the use of an individually-tailored instructional approach for Soldiers, and it highlights the need for technologies present adaptive learning scenarios and tailored feedback and instructional support (TRADOC, 2011). One important component to effective implementation of this strategy is frequent and reliable assessment enabling accurate tailoring of training content to the learner's level of competency. Although ALM calls for increased use of assessments specifically promoting the use of valid pre-, and post-tests for this purpose, limited detail on specific types of assessments for Army training is provided. As a starting application we aimed to examine the effect of computer adaptive testing (CAT) as the method of assessment for Experiment 2 based on previous research results indicating its benefits.

CATs emerged in the 1960s and provided a method for tailoring a test to a test-taker's ability. A traditional non-adaptive test is characterized by a fixed set of questions or items that are administered to each test-taker. CATs provide a tailored set of test items for each test-taker that most effectively and efficiently measures that person's ability.

CATs offer several advantages over traditional testing. The primary advantage of using a CAT is that it can provide greater measurement precision and overall efficiency. By utilizing data about how an item assesses an underlying ability and information about how the test-taker answered previous items, a CAT presents items that are close to the test-taker's true level of ability. A CAT is typically able to use fewer questions than would be required by a traditional test to achieve the same level of precision because it is able to present items that are close to the test-taker's level of ability (Mardberg & Carlstedt, 1998; Moreno & Segall, 1997). In a military training setting, reduced time for assessment could translate to potential training benefits such as reduced trainee fatigue and higher trainee engagement.

Another benefit of CAT is that it provides improved test security, making it more difficult to cheat. In a CAT, test-takers receive a different set of items during each administration, making it very difficult for one test-taker to provide information to subsequent test-takers or for a test-taker to improve performance on the test overall by learning only a few items (Green, 1983). In the training context, these characteristics are advantageous for pre-/post-testing in that there should be no practice effect from having taken the pre-test.

The main goal of Experiment 2 was to examine the potential benefits of using CATs in an Army training context. With this overall research goal in mind, we were interested in answering the following questions specifically:

- 2a) Is there a lower standard error associated with the CAT vs. a non-adaptive test?
- 2b) Does the CAT better correlate with performance on other independent ability tests, such as the work sample task, than traditional test scores?

Standard error of measurement (SEM) is an individual assessment precision index based on item response theory (IRT). A lower standard error is indicative of greater precision, and this

indicator is commonly used in examining CATs and CAT items (e.g., Weiss, 2004). For this experiment, we expected CAT items to be associated with lower SEMs than items from a non-adaptive test given CATs should achieve greater precision than non-adaptive tests (e.g., Mardberg & Carlstedt, 1998; Moreno & Segall, 1997). We also expected CAT scores to be more strongly correlated with the other ability measures administered independently of the CAT during the experiment.

Method

Participants

A total of 42 Soldiers participated in Experiment 2. All participants were from the MOS 25U except for one who came from Cross Level 25B. Soldiers were in ranks E5 through E7 with an average of 34 months in rank, as well as an average of 8 years in the military and 2 deployments. The CAT condition included a total of 21 Soldiers, including 16 from ALC and five from SLC. The non-adaptive condition included 21 Soldiers, 15 from ALC and six from SLC. As in Experiment 1, Soldiers were recruited at a time where ALC and SLC courses overlapped at Fort Gordon, GA, and Soldiers had availability to participate in this research for one week.

Materials

In Experiment 2, our intent was to replicate the environment created in the prototype condition of Experiment 1; however, at the three points where the CAT was administered in Experiment 1 (before and after the mobile training and just after the virtual classroom) half of the participants were administered a non-adaptive version of the same test. In the non-adaptive version, Soldiers received a random set of items pulled from each of the four training content domains (three items from each domain for a total of 12 items). The remainder of the materials and apparatus used in this experiment were identical to Experiment 1.

Experimental Design and Procedure

Experiment 2 used a quasi-experimental non-equivalent control-group treatment-group design (Campbell & Stanley, 1963). The study included two conditions, an adaptive testing condition (i.e., CAT condition) and a non-adaptive testing condition. Similar to Experiment 1, experimental equivalence of the conditions was confirmed by examining demographic information and pre-test scores.

The procedure in Experiment 2 was identical to the procedure used in the prototype condition in Experiment 1, except that for half of the participants in Experiment 2, the CAT was replaced with the non-adaptive assessment at all three administrations. Additional data from six participants was collected during the team's next site visit one month later because we were not able to collect enough data for both conditions during the one-week site visit.

Results

Equivalence of Groups

Using the same analytical procedures detailed in the Results section of Experiment 1, we compared scores between the two groups on the knowledge-based pre-test, participant experience in using the MBITR/JEM (i.e., whether the participant has experience using an MBITR or JEM and amount of experience in months), participant MBITR/JEM expertise (i.e., self-rated level of expertise using an MBITR/JEM on a 4-point Likert scale), and participant experience loading keys (i.e., self-rated level of experience loading keys on an MBITR/JEM on a 4-point Likert scale).

Results of a series of t-tests conducted to compare the group means of pre-test scores, participant experience and expertise with the MBITR/JEM, and participant experience in loading keys are summarized in Table 6. As expected, there were no significant differences between the two treatment groups on these variables.

Table 6
Means, Standard Deviations and t-Values on Equivalence Measures

	Mean (SD)		
	CAT (N = 21)	Non-Adaptive Test (N = 21)	t
Pre-Test Score*	6.95 (2.25)	6.86 (2.56)	.13
Experience with MBITR/JEM (Yes/No)	.90 (.30)	.81 (.40)	.87
Experience with MBITR/JEM (in months)	23.68 (23.53)	13.95 (18.74)	1.48
MBITR/JEM Expertise**	2.86 (.96)	2.52 (.98)	1.11
Experience Loading Keys***	3.38 (.92)	2.86 (1.15)	1.63

*Score is out of a possible 14;

**Self-rating on a 4-point Likert scale: 1 = Have never used before; 2 = Have used for only standard communication; 3 = Have performed simple troubleshooting in usual circumstances; 4 = Have performed advanced troubleshooting in unusual circumstances.

*** Self-rating on a 4-point Likert scale: 1 = Have never performed or seen it performed; 2 = Have never performed but have seen it performed; 3 = Have performed but not as part of usual duties; 4 = Have performed as part of usual duties.

Next, we used a binomial logistic regression⁸ model to predict the condition assignment (i.e., CAT and non-adaptive test) by a number of demographic variables, including rank, number of deployments, MBITR/JEM experience while deployed, experience and expertise with the MBITR/JEM, and experience in loading keys. Results of this analysis are summarized in Table 7. As in Experiment 1, the overall model was not found to be statistically significant. In examining the individual predictors, only number of deployments was significant ($p < .05$). Overall, this provided little evidence of systematic differences between the two treatment conditions prior to the study.

Table 7
Summary of Logistic Regression Analysis for Demographic Variables Predicting Group Assignment

Predictor	<i>B</i>	<i>SE B</i>	<i>Wald</i>	<i>df</i>	<i>p</i>	<i>e^B</i>
Rank	-.62	.63	.96	1	.33	.54
Number of Deployments	.87*	.55	4.32	1	.04	2.38
MBITR Experience While Deployed	-.18	1.47	.02	1	.90	.84
JEM Experience While Deployed	-.73	1.06	.48	1	.49	.48
Experience with MBITR/JEM (Yes/No)	.17	2.34	.01	1	.94	1.18
Experience with MBITR/JEM (in months)	-.02	.02	.78	1	.38	.98
MBITR/JEM Expertise	.32	.72	.19	1	.66	1.37
Experience Loading Keys	-.69	.68	1.04	1	.31	1.04
Constant	.87	1.58	.30	1	.58	2.38
χ^2			10.46			
<i>df</i>			8			
Overall % Predicted Correct			76			

Note: e^B = exponentiated *B* (odds ratio)

* $p < .05$

⁸ For more details on logistic regression analysis, see Menard (2002).

Finally, the pre-test score was regressed on a list of demographic variables to further reveal any systematic differences existing in this sample prior to the study. Consistent with Experiment 1, the linear regression analysis indicated the only significant predictor was self-rated MBITR/JEM expertise, explaining 34.4% of variance in pre-test scores ($F = 21.02, p < .001$). As expected, participants who rated themselves higher on prior MBITR/JEM expertise scored better on the pre-test ($\beta = .59, p < .001$). However, there was no difference in pre-test score or MBITR/JEM expertise between the CAT and non-adaptive test conditions. Neither number of deployments or course membership emerged as significant predictors of pre-test score. This supports the conclusion that our sampling and condition assignment strategy was adequate in meeting the assumptions required of the experimental design.

Overall Research Question 2: Does a CAT offer value over and above non-adaptive testing in a training context?

Experiment 2 aimed to examine the added value of the adaptive testing process compared to a random selection process of the same assessment items. Specifically, we compared the test SEM on each assessment occasion, which is an individual assessment precision index based on item response theory (IRT), between the CAT and non-adaptive test conditions. The purpose of this analysis is to determine whether the adaptive testing process yielded more precise assessment results given the same amount of items administered on each assessment occasion between conditions. In addition, we investigated whether the CAT scores were more strongly correlated with the other ability measures in the experiment, including the pre- and post-tests and the work sample task. Results are presented below.

2a) Standard error of measurement. The standard error of measurement (SEM) was computed for each individual test-taker's ability estimate using IRT metrics. These metrics (i.e., SEM and trait ability) can be computed for both traditional non-adaptive test items and adaptive test items. Each time an individual takes the test, there is a SEM associated with that particular ability estimate. If a single learner were to take the same test repeatedly, with no new learning taking place between testings and no memory of question effects, the standard deviation of the individual's repeated test scores is denoted as the SEM. The purpose of comparing the SEM between the adaptive and non-adaptive testing processes on each assessment occasion was to confirm that, by administering the same number of items drawn from the same item bank, a higher level of measurement precision can be achieved through an algorithm that is adaptive to the test-taker's ability in comparison to a random item selection process. In our analyses, we conducted a series of t-tests to compare the test scores and associated SEM on all three assessment occasions (i.e., pre-mobile, post-mobile/pre-virtual classroom, post-virtual classroom).

Results, including test score means, SEMs, and t-scores, are summarized in Table 8. Note that the test score means are average ability estimates based on IRT. Typically, a test score based on IRT ranges from -3 to 3; the higher the score, the higher the ability level of the Soldiers.⁹ For an individual Soldier, a score of 0 would be the midpoint of the ability spectrum as determined

⁹ For more details on how trait scores are calculated with IRT models, see Embretson and Reise (2000).

by the IRT model, which was developed based on ratings of item difficulty collected from SMEs in workshops prior to test administration. Specifically, a panel of subject matter experts rated each item on level of difficulty (see Brusso et al., 2014, for a full description of the methodology). The test score mean represents the average ability estimate of all of the Soldiers who completed the test within each condition. As predicted, the average SEM associated with scores in the CAT condition was found to be significantly lower than the average SEM associated with scores in the non-adaptive condition. The actual scores (i.e., ability estimates) were not statistically different between the two conditions.

Table 8

Means, Standard Deviations and t-Values on Participant Scores and SEM

	Mean (SD)		<i>t</i>
	CAT (N = 21)	Non-Adaptive Test (N = 21)	
Pre-Mobile Test Score	-.22 (.76)	-.38 (.61)	.74
Pre-Mobile SEM	.36 (.02)	.44 (.02)	-12.36***
Post-Mobile Test Score	.21 (.88)	-.01 (.65)	.94
Post-Mobile SEM	.35 (.02)	.44 (.03)	-11.28***
Post-Virtual Classroom Test Score	.34 (.66)	.69 (.57)	-1.79
Post-Virtual Classroom SEM	.36 (.04)	.47 (.05)	-8.56***

*** $p < .001$.

2b) Correlations with pre- and post-tests and work sample task. In addition to comparing the SEM between the adaptive test and non-adaptive test conditions, we compared the correlations between the ability estimates on each assessment occasion and the other ability measures in the experiment, including the pre- and post-tests and the work sample task. Because the adaptive test ability estimates have lower levels of SEM associated with them, we expected the scores to be more strongly correlated with the other ability measures. As shown in Table 9, the adaptive ability estimates of both the post-mobile and post-virtual classroom scores were significantly correlated with the work sample task, whereas the non-adaptive ability estimates were not. However, we did not find any evidence of a relationship with the pre-test or post-test.

Table 9

Correlations between Ability Estimates and Pre- and Post-Tests and Work sample task

		Pre-Test	Post-Test	Work sample task
Pre-Mobile Test Score	CAT	.47*	.23	.26
	Non-Adaptive Test	.63**	.07	.41
Post-Mobile Test Score	CAT	.18	.31	.44*
	Non-Adaptive Test	.52*	.57**	.32
Post-Virtual Classroom Test Score	CAT	-.01	.35	.44*
	Non-Adaptive Test	.29	.03	.29

* $p < .05$.

** $p < .01$.

Discussion

Our findings showed that the CAT did yield a lower SEM than the non-adaptive test as expected, indicating that the CAT was more precise than the non-adaptive test given the same number of items administered on each assessment occasion. In addition, the CAT scores were more highly associated with the scores on the work sample task. This indicates that even with a relatively small sample size, the CAT was more precise and a better predictor of performance on this task than the non-adaptive version of the same test. It should be noted that the CAT scores themselves were not different between the two groups, which is expected because Soldiers between the two experimental conditions did not differ in their knowledge and abilities on the JEM. The results confirmed that the CAT is a more reliable test than the non-adaptive test and therefore is a more efficient and effective assessment. The scores from the CAT better reflect the Soldier's ability rather than other random factors, however, a better assessment tool alone cannot increase learners' ability. Rather, it offers a more reliable gauge that can be used with other tools to improve learning. For example, if training was to be tailored based on assessment results, one can be more confident in the learning paths determined by the CAT as opposed to the non-adaptive test, given the same number of assessment items.

There were several limitations of this experiment. The CAT used as part of this effort was a one-parameter logistic model (1PLM), which means that the only property used to select items for presentation was difficulty. Difficulty levels of CAT items are typically determined empirically, using large samples. In this case, we used expert judgment to determine difficulty levels. While there is evidence supporting this method as valid (e.g., Stark, Chernyshenko, & Guenole, 2011), there is also certainly risk. Traditionally, items for a 1PLM CAT would be empirically calibrated with a sample of 150 or even larger. One consideration in using CAT for training purposes should be the feasibility of item calibration. If a 3PLM is used in constructing a CAT, items should be calibrated with samples of 1,000 or more. The time and costs associated with this must be balanced with precision and efficiency needed in a training context.

As noted in our Year 1 Report (Brusso et al., 2014), another limitation of our CAT was that we had to make assumptions about the dimensionality of the training.¹⁰ A CAT fundamentally assumes unidimensionality of the content. In other words, if a test-taker scores high on one CAT item, he or she is likely to score high on another item. In this case, it was not possible to know for sure whether or not the training was unidimensional. We made the assumption that the various content areas are related to each other and all pertain to a higher-order general ability to use the JEM. We did identify four content domains within the training, however. In order to ensure learners received items from each of the domains, we constructed the CAT such that the same number of items from each content area were administered every time.

¹⁰ We relied on an assumption that the various content areas in the JEM training (i.e., Functionality/Basic Procedures, Complex Procedures, Terminology and Specifications, and Troubleshooting) are related to each other and all pertain to a higher-order general ability to use the JEM. The different content areas are represented in the item pool and the same number of items from each content area is drawn for each test. However, the adaptive mechanism functions across content areas (i.e., the item selection from one content area may be based on the response to an item from a different content area).

The fact that the standard error was lower for CAT items than for the non-adaptive test reinforces that this structure was more effective than a traditional test. However, without additional testing on dimensionality of the items, we cannot know for sure that this was the optimal structure.

The issue of dimensionality should be considered carefully in using CATs in a training environment. The first part of that consideration should be whether or not the dimensionality is known – or that investment can be made in determining it. The second issue is the implication of constructing a multi-dimensional CAT. In essence, a separate CAT must be constructed for each dimension, reducing any potential efficiency-related benefits. Even in circumstances in which the training is relatively unidimensional, but includes multiple content domains, the content balancing of the items reduces the efficiency of the CAT (e.g., Kingsbury & Zara; 1991; Weiss, 2004).

EXPERIMENT 3: THE EFFECTS OF PERIODIC TESTING DURING TRAINING

As discussed previously, the ALM (TRADOC, 2011) promotes assessment as a way to ensure learning has occurred to a standard. In addition, tailoring instruction to the learner necessitates periodic assessment to track progress. However, there is some evidence that periodic testing also influences actual learning (e.g., Carrier & Pashler, 1992).

Training assessments are commonly thought of simply as a means of measuring learning or skill. However, there is significant evidence that testing alone influences learning (Carrier & Pashler, 1992). Historically known as ‘the testing effect,’ the term ‘test-enhanced learning’ is currently used to describe an educational strategy where the effect is intentionally employed to positively influence learning outcomes (Larsen, Butler, & Roediger, 2009; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a; 2006b).

One key factor underlying dynamic driving the positive learning outcomes, associated with test-enhanced learning, stems from requiring the learner to rehearse retrieval (Carrier & Pashler, 1992; Karpicke & Roediger, 2008). The idea is that rehearsing retrieval will strengthen the retrieval process, reduce forgetting, and support the long-term retrieval of the learning content. Therefore, targeted learning outcomes will be enhanced when the learning objectives need to be operationalized (e.g., demonstrated during an end of course assessment, or performed on the job) (Karpicke & Roediger, 2008).

Feedback is another key component of test-enhanced learning. Larsen, Butler, and Roediger (2009) demonstrated that repeated interim assessments incorporating feedback supports superior long-term retention in learners better than repeated assessment without feedback. Feedback that helps learners understand how to change their performance is more effective than feedback that simply provides learners with the results of their assessment (e.g., Bransford, Brown, & Cocking, 1999). Such targeted feedback reinforces the training content and can help foster a deeper level of learning (Cannon-Bowers & Bowers, 2010).

There is ample evidence that assessment and feedback positively affect learning (see Cannon-Bowers & Bowers, 2010, for review) leading us to ask how frequently assessment should be provided during training. Larsen, et al. (2008), and Karpicke & Roediger (2008) demonstrated that repeated testing promotes better retention over time than a single test. In addition, the research evidence shows that benefits from multiple assessments are greater when multiple assessments are distributed over time (Karpicke & Roediger, 2007; Larsen, et al., 2008). Karpicke & Bauernschmidt (2011) found evidence that differing spacing schedules had varying impact on learning outcomes. In that study, repeated retrieval tasks having longer intervals between them improved long-term retention over repeated retrieval tasks with no spacing. Regarding timing of feedback, Bransford, et al. (1999) argue that feedback should be frequent or continuous, but there is also evidence that intermittent feedback may help trainees to be less dependent on continuous reinforcement (Schmidt & Bjork, 1992; Shute & Gawlick, 1995).

The purpose of Experiment 3 was to determine whether or not interim assessment during the mobile training and the virtual classroom resulted in higher levels of learning. The prototype mobile training and virtual classroom both included formative assessments in the form of check on learning activities that served to reinforce the training content and provide feedback to Soldiers. There were eight item types that included true/false, matching (one versus multi-part),

traditional multiple choice, choose all that apply, interactive multiple choice, scored manipulation of the virtual radio, and scored programming of the virtual radio. There was a stronger focus on declarative knowledge in the mobile training; in the virtual classroom, there was a stronger focus on demonstrating that knowledge. When learners answered a check on learning item incorrectly, they received feedback that included a brief explanation of the correct answer. For programming items, the feedback presented to the learner provides specific guidance regarding the correct procedural action. In keeping with the overall research goals Experiment 3 was designed to answer the following question:

1. Do learners who receive interim check on learning assessments during the prototype training perform better on learning outcomes than those who do not?

Based on existing evidence supporting the use of frequent assessment and feedback (e.g., Cannon-Bowers & Bowers, 2010; Larsen et al., 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006a, 2006b), we expect that learners who receive check on learning items to perform better on learning outcomes than those who do not.

Method

Participants

A total of 60 Soldiers participated in Experiment 3. All participants were from the MOS of 25U and in ranks E5 through E7 with an average of 29 months in rank. Soldiers had an average of 10 years in the military and 3 deployments. The treatment condition of this study (those who completed training with interim check on learning assessments), was identical to the prototype condition of Experiment 1. Therefore, the interim check on learning condition in Experiment 3 included 33 Soldiers, 20 from ALC and 13 from SLC (same data collection as the prototype in Experiment 1). The control condition (without interim check on learning assessments) included 27 Soldiers, all from ALC. As in the previous studies, Soldiers were recruited for the treatment condition in Experiment 3 during a week where ALC and SLC courses overlapped at Fort Gordon, Georgia, having availability to participate in this research for one week.

Materials

The materials used in this experiment were similar to those used in Experiment 1. A total of 10 iPads were used to deliver the mobile training and 27 laptops were used to deliver the virtual classroom and collaborative scenario. The questionnaires and test items were identical to those described in Experiment 1.

Experimental Design

Experiment 3 also used a quasi-experimental non-equivalent control-group treatment-group design (Campbell & Stanley, 1963). The study included two conditions, a treatment condition in which participants received interim check on learning assessments during training (interim check on learning assessment condition) and a control condition in which no check on learning assessments were presented. As in the previous studies, experimental equivalence of the conditions was confirmed by examining demographic information and pre-test scores.

The procedure used to collect data for the Experiment 3 control condition was identical to the procedure used in the prototype condition in Experiment 1 (the treatment condition of Experiment 3) except that all interim check on learning assessments were removed from the mobile and virtual classroom trainings

Results

Equivalence of Groups

As in the previous experiments, in Experiment 3, we compared scores between the control (no interim check on learning assessments) and treatment (with interim check on learning assessments) groups on the knowledge-based pre-test, participant experience in using the MBITR/JEM (i.e., whether the participant has experience using an MBITR or JEM and amount of experience in months), participant MBITR/JEM expertise (i.e., self-rated level of expertise using an MBITR/JEM on a 4-point Likert scale), and participant experience loading keys (i.e., self-rated level of experience loading keys on an MBITR/JEM on a 4-point Likert scale). Two statistical tests were conducted to test for the pre-treatment equivalence between the two conditions. First, a series of t-tests were employed to compare the group means of pre-test scores, participant experience and expertise with the MBITR/JEM, and participant experience in loading keys. An examination of the t-test results revealed that there were significant differences between the two groups in their pre-test score. This may, in part, be explained by the lack of inclusion of SLC in the control condition. Due to constraints at the data collection site, we had been permitted to only collect data from ALC Soldiers. To account for this difference, we examined the change in score from pre-test to post-test rather than just differences in post-test score. We also controlled for treatment condition and course in analyses. Means and standard deviations for each variable are provided in Table 10.

Table 10

Means, Standard Deviations and t-Values on Equivalence Measures

	Mean (SD)		
	Control (N = 27)	Treatment (N = 33)	<i>t</i>
Pre-Test Score*	5.89 (1.89)	7.39 (1.89)	3.07 ⁺
Experience with MBITR/JEM (Yes/No)	0.96 (0.19)	0.91 (0.29)	-.82
Experience with MBITR/JEM (in months)	28.26 (29.01)	30.33 (30.29)	-.09
MBITR/JEM Expertise**	2.93 (0.78)	2.82 (0.92)	-.50
Experience Loading Keys***	3.30 (0.99)	3.30 (0.85)	.03

*Score is out of a possible 14;

**Self-rating on a 4-point Likert scale: 1 = Have never used before; 2 = Have used for only standard communication; 3 = Have performed simple troubleshooting in usual circumstances; 4 = Have performed advanced troubleshooting in unusual circumstances.

*** Self-rating on a 4-point Likert scale: 1 = Have never performed or seen it performed; 2 = Have never performed but have seen it performed; 3 = Have performed but not as part of usual duties; 4 = Have performed as part of usual duties.

⁺ $p < .01$

Second, the condition assignment (i.e., control vs. treatment) was regressed on a number of demographic variables through a logistic regression model to examine whether the condition assignment could be explained by any systematic differences between the groups. Demographic variables entered into the model included rank, number of deployments, MBITR/JEM experience while deployed, experience and expertise with the MBITR/JEM, and experience in loading keys. Results of the logistic regression analysis indicated that the overall regression model was not statistically significant; however, both rank and deployments were found to be significant predictors of group assignment (see Table 11). This may be a result of including only ALC students in the control condition. Although there were some demographic differences between groups, given that most variables related to skill with the JEM were not significant, and given that the model as a whole was not significant, we proceeded with further investigation of the third research question.

Table 11

Summary of Logistic Regression Analysis for Demographic Variables Predicting Condition Assignment

Predictor	<i>B</i>	<i>SE B</i>	<i>Wald</i>	<i>df</i>	<i>p</i>	<i>e^B</i>
Rank	.85*	.42	4.19	1	.04	2.34
Number of Deployments	-.61*	.30	4.21	1	.04	.54
MBITR Experience While Deployed	-.20	.78	.07	1	.80	.82
JEM Experience While Deployed	1.37	.80	2.97	1	.09	.80
Experience with MBITR/JEM (Yes/No)	1.95	1.78	1.20	1	.27	1.24
Experience with MBITR/JEM (in months)	.01	.01	.73	1	.39	1.00
MBITR/JEM Expertise	-.07	.65	.01	1	.92	.98
Experience Loading Keys	-.51	.58	.77	1	.38	1.28
Constant	-1.22	1.67	.53	1	.47	.29
χ^2			10.31			
<i>df</i>			8			
Overall % Predicted Correct			70.0			

Note: *e^B* = exponentiated *B* (odds ratio)

**p* < .05

Next, we examined predictors of the pre-test score to further explore any systematic differences that may have existed in the overall sample prior to treatment. A linear regression analysis revealed two significant predictors: condition, explaining 14.0% of the variance in pre-test scores ($F = 9.45$, $p < .05$) and self-rated MBITR/JEM expertise, explaining 25.2% of variance in pre-test scores ($F = 9.61$, $p < .05$). As expected, participants who rated themselves higher on prior MBITR/JEM expertise scored better on the pre-test ($\beta = .34$, $p < .05$). However, as mentioned previously, there was no difference in MBITR/JEM expertise between the two conditions. Soldiers in the treatment condition (with interim learning checks) scored higher on the pre-test ($\beta = .37$, $p < .05$). Course membership (i.e., ALC vs. SLC), rank, and deployments did not emerge as significant predictors of pre-test score. This provides mixed evidence to

support that our two experimental groups were equivalent in knowledge of the JEM entering the experiment. The analyses that follow statistically control for a number of variables, including pre-test score, rank, deployments, course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys, to account for any existing systematic differences between the conditions.

Research Question 3: Do interim assessments lead to better learning outcomes?

The primary goal of this experiment was to better understand the impact of interim assessments; specifically, how including interim check on learning assessments through the mobile and virtual classroom trainings impacted learning. To investigate this, we used pre-/post-test scores to compare overall learning in the control (without interim check on learning assessments) to overall learning during the treatment condition (the full prototype with interim check on learning assessments in the mobile and virtual classroom trainings). We then examined scores on the work sample task to determine whether or not there was a difference in *transfer* of learning to the use of an actual JEM radio. Results of these comparisons are presented below.

Pre-/Post-test scores. To assess Research Question 3 we first examined whether or not participants exhibited a greater degree of learning the JEM radio as a result of including interim check on learning assessments in the mobile and virtual classroom trainings. Scores on the pre- and post-tests were used to examine the knowledge gain as a result of the training. Means and standard deviations on the tests are provided in Table 12. Results of the regression analyses comparing the two conditions are also included in this table.

We first examined the training effects on participants' understanding of the subject matter regardless of treatment condition. A paired sample t-test revealed a significant increase from the pre-test score to the post-test score ($t = 9.01, p < .001$). When we compared the pre-/post-test score increase for each condition, the results were also statistically significant ($t = 6.52, p < .001$ for Control and $t = 6.20, p < .001$ for Treatment). Next, we examined the treatment effects on the post-test scores through a linear regression model. In addition to treatment condition, control variables, including pre-test score, rank, deployments, course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys, were also entered into the equation as predictors. None of the variables were statistically significant predictors of post-test score. We also employed the same regression model with delta (i.e., pre-/post-test score change) as the outcome variable and the results were also not statistically significant.

Table 12

Means, Standard Deviations and Treatment Effects on Pre-/Post-test Scores and Delta

	Mean (SD)		r^2 Treatment (r^2 change with MBITR/JEM Expertise)
	Control (N=27)	Treatment (N=33)	
Pre-Test Score	5.89 (1.89)	7.39 (1.89)	14.0%* (9%**)
Post-Test Score	7.39 (1.89)	10.12 (2.12)	NS
Delta	2.96 (2.36)	2.73 (2.53)	NS

* $p < .01$

** $p < .001$

Work sample task. The last goal of this study was to determine whether or not participants who received check on learning assessments were able to use an actual JEM radio better than their counterparts in the control condition who did not receive check on learning assessments. Specifically, we compared final scores and task completion time on the work sample task between the treatment and control conditions. Means and standard deviations on the work sample task, results of the regression analyses comparing between the two conditions, and effect sizes, are provided in Table 13.

Table 13

Means, Standard Deviations and Treatment Effects on Work sample task

	Mean (SD)		r^2 Treatment (r^2 change with control variables ⁺⁺⁺)	Cohen's <i>d</i>
	Control (N=27)	Treatment (N=33)		
Task Completion Score ⁺	18.63 (4.28)	20.79 (2.78)	9%* (20%***)	.62
Task Completion Time (in minutes) ⁺⁺	23.20 (3.70)	20.80 (3.52)	10%* (17%**)	.68

* $p < .05$

** $p < .005$

*** $p < .001$

⁺ Score is out of a possible total of 26;

⁺⁺ The limit for task completion time is 30 minutes.

⁺⁺⁺ For Task Completion Score, MBITR/JEM Expertise and pre-test scores were both statistically significant control variables; for Task Completion Time, the only statistically significant control variable was MBITR/JEM Expertise.

To test for the treatment effect on the work sample task, we conducted regression analyses to examine whether or not treatment condition predicted work sample task completion scores or completion time. Consistent with the previous tests, control variables, including pre-test score, course membership, rank, deployments, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys, were also entered into the equations as predictors in addition to treatment condition. Treatment condition was found to be a statistically significant predictor of both Task Completion Score and Task Completion Time. Participants in the treatment condition scored higher on the work sample task ($\beta = .300, p < .05$) and completed it faster ($\beta = -.320, p < .05$) than their counterparts in the control condition. However, for Task Completion Score, the treatment effect is no longer statistically significant after controlling for all the other variables aforementioned. Instead, participants with higher pre-test scores and greater MBITR/JEM expertise scored higher on the work sample task ($\beta = .286, p < .05, \beta = .283, p < .05$, respectively). For Task Completion Time, the treatment effect remained statistically significant after controlling for all the other variables aforementioned. Participants completed it faster ($\beta = -.415, p < .01$). We also calculated Cohen's *d* to measure the size of the effect between conditions for both measures of the work sample task. Results indicate that there was a medium sized effect for both task completion score and task completion time.

Exploratory Analyses

Although it was not one of our initial research questions, we conducted additional analyses to investigate whether including interim check on learning assessments impacted

reactions to the prototype trainings. It was believed that including interim check on learning assessments may have increased Soldier engagement throughout the prototypes. As in Experiment 1, we compared the two conditions on three dimensions from the general reaction questionnaire: satisfaction, perceived learning, and perceived learning transfer.

As in Experiment 1, reactions to the control and treatment conditions were compared on the satisfaction, perceived learning, and perceived learning transfer measures developed for this study (see Appendix E). Table 14 shows the comparisons of mean ratings and standard deviations for each reaction survey item by condition. Results of the regression analyses comparing the two conditions are also included in this table.

To test whether participants reacted differently to the prototypes with interim checks on learning and the prototypes without checks on learning, we employed a series of linear multiple regression models with reaction survey items as the criterion and the condition assignment as the predictor. Additional control variables entered into the regression model as predictors included pre-test score, course membership, MBITR/JEM experience, MBITR/JEM expertise, and experience loading keys. Overall, treatment condition was only a significant predictor for scores on two reaction items: “This training was relevant to my job in the Army” (perceived learning transfer) and “I would recommend this training program to other Soldiers (satisfaction). Although participants in the prototype condition generally reported much more favorable reactions across all three dimensions, the only statistically significant item was for satisfaction, and the result was not in the predicted direction, as those in the control condition were more likely to recommend the training (see Table 14 on next page). This may have occurred because the training was shorter in duration due to the removal of the interim check on learning assessments. Soldiers did report that the version of the training with interim checks on learning was more relevant to their job in Army, perhaps because by completing the check on learning assessments Soldiers had much more “hands on” practice with the virtual JEM radio.

Table 14

Means, Standard Deviations, and Treatment Effects on Perceived Learning, Perceived Learning Transfer, and Satisfaction

Dimension/Item	Mean (SD)		r^2 Treatment (r^2 change with MBITR/JEM Expertise)
	Control (N=24 ⁺)	Prototype (N=33)	
Perceived Learning			
1. Learning this material was fun.	4.12 (.90)	4.24 (.90)	NS
2. Overall, I have learned a lot from this training.	3.46 (1.18)	3.88 (.89)	NS
Perceived Learning Transfer			
1. It is clear to me that the people conducting the training understand how I will use what I learn.	3.75 (.90)	3.82 (.92)	NS
2. This training was relevant to my job in the Army.	3.92 (.83)	4.33 (.69)	NS
3. I believe the training will help me do my current job in the Army better.	3.71 (1.12)	3.94 (1.17)	NS
4. I learned something I can apply immediately to my work in the Army.	3.63 (1.17)	3.82 (1.07)	NS
5. I plan to use what I learned on my job in the Army.	3.75 (.90)	4.06 (.97)	NS
6. I am prepared to train other Soldiers on what I learned in this training.	3.88 (.80)	4.09 (.95)	NS
7. I get excited when I think about trying to use my new learning on my job in the Army.	3.63 (1.10)	3.88 (1.11)	NS
8. I will be using the equipment on my job in the Army after the training.	3.88 (.80)	4.00 (1.03)	NS
9. The training was of practical value to me.	3.79 (1.06)	4.03 (1.05)	NS
Satisfaction			
1. I enjoyed this training program.	3.96 (.95)	4.09 (1.13)	NS
2. My time on the training was well spent.	3.67 (.92)	4.03 (1.21)	NS
3. I would recommend this training program to other Soldiers. ⁺⁺	3.79 (1.18)	3.21 (.74)	9%* (8%*)

⁺ Sample size was only 24 because three participants from the control condition did not complete the general reactions questionnaire.

⁺⁺ Items that were also predicted by self-rated MBITR/JEM expertise.

* $p < .05$

Discussion

Our findings provide mixed evidence that interim assessments contributed to learning. Interim check on learning assessments did not appear to impact knowledge at the end of the training or change in knowledge (as measured by pre/post-tests). However, they did appear to impact application of knowledge to a real radio. Specifically, the treatment effect was statistically significant even after controlling for the existing sample group differences in pre-test scores. One explanation for this finding is the interim check on learning assessments tended to be more application-based as opposed to knowledge-based. Many included hands-on activities such as programming a radio, rather than simple knowledge-recall items. The difference in the types of items on the different measures (i.e., post-test vs. work sample) may explain why the effect of testing was more pronounced in the work sample scores. Further, we did not remove the CAT items from the control condition. It is possible that, although there was no feedback given during CAT assessments, the recall required to complete the CAT had more of an impact on the knowledge-based post-test than the more hands-on check on learning items. Thus

One key limitation of this study was that our groups were not equivalent. We statistically controlled for demographic and pre-test differences to account for this. However, ideally, this study would be replicated with equivalent groups.

Another constraint of this study was the brevity of the training. The total training time was less than eight hours. We would expect a greater effect of the interim assessments for a longer-term course. Ideally, this study should be replicated using a course in which retention of material across the course would be of greater concern. A week or month would also allow for examination of an optimal frequency of assessment.

Another interesting follow-up to this study would be to examine assessment with no feedback, assessment with minimal, and assessment with feedback that includes *why* an answer is correct or incorrect. While there is evidence that both the act of engaging in assessment and the receipt of feedback can result in positive learning outcomes (e.g., Carrier & Pashler, 1992; Karpicke & Roediger, 2008; Cannon-Bowers & Bowers, 2010), additional research on the utility of specific assessment and feedback features is an important next step in understanding how to best implement these types of tools in the Army.

GENERAL DISCUSSION

Review of overall findings

The primary purpose of this effort was to conduct research in support of the ALM. One goal was to examine the effectiveness of mobile, virtual, and game-based trainings with respect to the classroom-based training these technologies are intended to replace. An additional goal was to examine the most effective ways to utilize assessments in training. Specifically, we examined the utility of CATs as an alternative to traditional tests and we examined the effectiveness of interim assessments in training.

Our results support the principles of the ALM in several ways. First, we found that participants in our prototype training had more positive attitudes towards training and scored higher on work sample task measures than their counterparts in the classroom training. This

sheds light on the overall benefits of Soldier-centered learning, including concepts such as allowing Soldiers to proceed through the training at their own pace and learning through collaborative exercises and scenarios. In addition, we were able to explore specific assessment strategies in response to the ALM's call for more frequent assessment that enable accurate tailoring of training content to the learner's level. In particular, we found that our CAT was a more precise measure of student learning than a traditional test; it was also more highly associated with scores on a work sample task than a traditional test. Finally, we found some evidence that the use of interim assessments was associated with better performance on our work sample task, providing some preliminary evidence in support of more frequent assessments throughout the training.

Lessons Learned

Although the results of our study lend support to the ALM's charge to shift from classroom-based instruction to more mobile, virtual, and game-based instruction, we offer several lessons learned, limitations, and considerations. Reactions were generally more positive to the prototypes than the classroom learning and the research team reported perceiving much higher levels of engagement during the prototypes especially the game-based collaborative scenario. This observation supports to the belief that game-based training can increase motivation (e.g., Chang et al., 2009; Mautone et al., 2010; Topolski et al., 2010). One lesson learned pertaining to this observation, however, is that game-based training developers must not lose sight of the learning objectives. We believe one reason for the success of these prototypes was that each component and feature was carefully linked back to learning objectives. Another feature we believe contributed to success was the tutorial at the beginning of the collaborative scenario. It was critical to ensuring that users had both the preliminary JEM knowledge and understanding of the interface before engaging in the collaborative scenario.

Another important lesson was that maintaining engagement and facilitating interaction in a virtual classroom can be challenging. In a traditional classroom, an instructor is able to monitor the attention of the learners and keep learners on-task. In a live environment, an instructor can "call on" learners and ask learners to demonstrate a task at-will. In a virtual environment, this is much more challenging because instructors cannot physically see or interact with learners. It is difficult to know if the learner is actually watching or listening at all. It would be beneficial to include features that allow instructors to make the classroom as active as possible. For example, a virtual classroom could include a function to allow the instructor to assign tasks to students and have them complete the task "in front" of the class. An example of this in the present application would have been allowing the instructor to "pass" the radio to a student so that she or he could demonstrate a task for the class. Similarly, our virtual classroom was better received by learners when the design moved away from a step-by-step format to a format that the instructor could use more flexibly. They reported that the step-by-step format was closer to a computer-based PowerPoint presentation, whereas the more flexible approach allowed the instructor to tailor the instruction appropriately.

As mentioned in the Experiment 1 discussion, not using the prototype trainings in the truest sense of a distance learning setting was a limitation of this effort. The control of having the participants complete the training in a classroom type environment was beneficial for ensuring consistency in Experiments 2 and 3. However, a key benefit of the technologies used here is the

ability to learn anytime, anywhere. A logical next step of this program of research would be to test the effectiveness of the training and assessment prototypes in a training setting to determine the generalizability of our findings.

A related limitation was that not all participants were novices with the JEM radio. Participants came into this study with a wide range of experience with the radio, and those with a high level of experience reported being slightly less engaged and found the training to be less useful but the overall perceptions were still positive (i.e., the average ratings being greater than three on a five-point Likert scale). The control and experimental groups were equivalent in terms of participant experience, so we do not expect that this impacted the overall results of this study. However, we expect that our findings regarding overall change in learning as a result of the prototype trainings would have been more robust had the participants all been novices.

A similar observation was that there were several instances in which learners who verbally reported considerable experience with gaming became frustrated when the interface or a particular feature did not function exactly the way the learner expected. Although this was anecdotal, this observation is consistent with the thinking that there is inherent risk of unmet expectations when implementing game-based training (Cannon-Bowers & Bowers, 2010). Specifically, frequent “gamers” may have set of assumptions and expectations about what a game is and how they function. Game-based training may not meet all of those expectations, so it is important to adjust those expectations when implementing this technology (Cannon-Bowers & Bowers, 2010).

Implications for Army Training

The results of these experiments provide evidence that mobile technologies, virtual environments, and game-based training can be effective, engaging tools. As the Army moves towards wider use of these technologies, we propose several considerations. First, training designers must consider the match between mode and content or task to be trained in order to maximize effectiveness. The Office of the Secretary of Defense is currently developing a decision tool to help instructors determine the suitability of individual and collective tasks for live or virtual instruction (Curnow, Paddock, Wisher, DiGiovanni, & Rosengrant, 2012). This decision tool includes factors such as:

- Interaction/Fidelity: Degree of interaction with data, things, people – and the implicit requirements for the training media to reflect that interaction
- Learning Complexity: Index of the complexity that includes number of steps involved, the complexity of the steps or subtasks, and the degree of decision-making involved.
- Task Certainty/Feedback: Degree to which feedback on task completion/success is provided synchronously, is somewhat delayed, or is very delayed or never available.

Based on our experience with the current studies, other factors may need to be added to such as model. For example, the purpose or primary use of the training may also be a factor of consideration – participants in our studies provided feedback on the utility of the mobile application as a means for refresher training in the field. Additional research on how to optimize

the match between training content and mode should be an important component of the shift to more Soldier-centered learning. Ultimately, if training content and mode are mismatched, effectiveness is likely to be lower than when training content and mode are optimally matched.

Another important consideration is that training is only as effective as its design, regardless of technology. One ongoing challenge in implementing distributed learning technologies is aligning them with learner cognitive capacity (e.g., Mayer, 2010). Although mobile applications, virtual classrooms, and game-based learning technologies have unprecedented capacity for flexibility, interactivity, and accessibility, they also carry potential for cognitive overload. Designers must work to strike the right balance between including interesting, interactive content with reducing extraneous cognitive processing for this type of technology to be effective (Mayer, 2010).

Finally, this study supports previous research indicating that distributed learning technologies can be effective tools for learning. Researchers are only beginning to examine the effectiveness of individual features that distributed learning technologies, especially game-based technologies, offer. Cannon-Bowers and Bowers (2010) noted that there are many unknowns regarding the optimal fidelity of a simulation, the most effective methods for embedding scoring into game-based training, and the most effective feedback mechanisms. The Army must continue to explore the possibilities for implementing the principles of ALM to insure the best training for Soldiers is made available.

Conclusion

The ALM calls for the Army to reduce instructor-led training and increase the use of blended approaches that incorporates virtual and constructive simulations, intelligent tutoring, gaming technology, and other technology-delivered instruction. The goal of this research was to evaluate the effectiveness of three of those technologies, mobile, virtual, and gaming technology, compared with classroom-based instruction. A second goal was to examine how to best utilize and incorporate assessments into technology-based training. We found that the prototype training developed for this effort was a successful alternative to its classroom-based counterpart. We also found that CATs offer promise as an alternative to more traditional tests in a training context and that interim assessments appeared to enhance both engagement and learning. However, our findings highlight the need to further examine the most appropriate uses of mobile, virtual classroom, and game-based technologies and the most important and optimal features in these types of technologies. In addition, the promise that these technologies hold also underscores the need to attend to sound design principles so that their potential can be realized.

References

- Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Walseth, P. A., Fiset, M., & Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research* 74 (3), 379–439.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academy Press.
- Brusso, R., Barnieu, J., Huang, J., Lodato, M., Mulvaney, R., Cummings, P., Zoellick, C., Thieme, K., & Spain, R. (2014). *Delivering Training Assessments in a Soldier-Centered Learning Environment: Year One* (Technical Report 1346). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Cannon-Bowers, J., & Bowers, C. (2010). On developing a science of simulation, games, and virtual worlds for training. In S. W. Kozlowski & E. Salas (Eds.), *Learning, Training, and Development in Organizations*. (pp. 229–261). New York: Taylor and Francis Group.
- Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment*, 8(4), 248–260.
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642.
- Chang, V., Gütl, C., Kopeinik, S., & Williams, R. (2009). Evaluation of collaborative learning settings in 3D virtual worlds. *iJet*, 4(3), 6–17.
- Cohen, J. (1992). "A power primer". *Psychological Bulletin*, 112 (1): 155–159.
- Colquitt, J. A., LePine, J. A., Noe, R. A. (2000). Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5), 678–707.
- Curnow, C. K., Paddock, A. F., Wisher, R. A., DiGiovanni, F. C., & Rosengrant, C. (2012). Live or virtual military training? Developing a decision algorithm. *Presentation at the 2012 Interservice/Industry Training, Simulation, and Education Conference*. Orlando, FL.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Green, B.F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. (pp. 69-80). Mahwah, NJ: Lawrence Erlbaum Associates.
- Holden, C. L., & Sykes, J .M. (2011). Leveraging mobile games for place-based language learning. *International Journal of Game-Based Learning*, 1(2), 1-18.
- Karpicke, J., & Bauernschmidt, A. (2011). Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 37(5), 1250-1257. doi:10.1037/a0023436
- Karpicke, J., & Roediger, H. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 33(4), 704-719.
- Karpicke, J., & Roediger, H. (2008). The critical importance of retrieval for learning. *Science*, 319, 966-968.
- Kiger, D., Herro, D., & Prunty, D. (2012). Examining the influence of a mobile learning intervention on third grade math achievement. *Journal of Research on Technology in Education*, 45 (1), 61-82.
- Kingsbury, C. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4(3), 241–261.
- Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, 42(10), 959-966.
- Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, 43(12), 1174-1181.
- Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge, UK: Cambridge University Press.
- Mardberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment*, 6(2), 107–114.
- Mautone, T., Spiker, A., Karp, M. R., & Conkey, C. (2010). Using games to accelerate aircrew cognitive training. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*, (pp. 1898-1909). Orlando, FL.

- Mayer, R. E. (2010). Research-based solutions to three problems in web-based training. In S.W. Kozlowski & E. Salas (Eds.), *Learning, Training, and Development in Organizations*. (pp. 203-227). New York: Taylor and Francis Group.
- McCloy, R. A., & Gibby, R. E. (2011). Computerized adaptive testing. In N. T. Tippins & S. Adler (Eds.), *Technology-Enhanced Assessment of Talent* (pp. 153-189). San Francisco: Jossey-Bass.
- McDaniel, M., Roediger, H., & McDermott, K. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14(2), 200-206.
- Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. U.S. Department of Education.
- Menard, S. (2002). *Applied Logistic Regression Analysis (2nd Ed)*. Thousand Oaks, CA: Sage.
- Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117-135). Mahwah NJ: Erlbaum.
- Moreno, K. E. (1997). CAT-ASVAB operational test and evaluation. In W. A. Sands, B. K. Waters, & R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 199-205). Washington, DC: American Psychological Association.
- Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169-179). Washington, DC: American Psychological Association.
- Norris, C., & Soloway, E. (2004). Envisioning the handheld centric classroom. *Journal of Educational Computing Research*, 30(4), 281-294.
- Olson, L. (2003, May 8). Legal twists, digital turns: Computerized testing feels the impact of "No Child Left Behind." *Education Week's Technology Counts*, 22(35), pp. 11-14, 16.
- Roediger, H., & Karpicke, J. (2006a). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255.
- Roediger III, H. L., & Karpicke, J. D. (2006b). The power of testing memory basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181-210.
- Roschelle, J., & Pea, R. (2002) A walk on the WILD side: How wireless handhelds may change computer- supported collaborative learning. *International Journal of Cognition and Technology*, 1, 145-168.

- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207-217.
- Shute, V., & Gawlick, L. (1995). Practice effects on skill acquisition, learning outcome, retention, and sensitivity to relearning. *Human Factors*, 37(4), 781-803.
- Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel Psychology*, 59(3), 623-664.
- Smith, E.M., Ford, J.K., & Kozlowski, S.W.J. (1997). Building adaptive expertise: Implications for training design strategies. In M. A. Quinones & A. Ehrenstein (Eds.), *Training for a rapidly changing workplace: Applications of psychological research* (pp. 89-118). Washington, DC: American Psychological Association.
- Stark, S., Chernyshenko O. S., & Guenole, N. (2011). Can subject matter experts ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods*, 14, 256-278.
- Taylor, J. C. (2001) Fifth generation distance education. *e-Journal of Instructional Science and Technology (e-JIST)*, 4 (1), 1-14.
- Topolski, R., Leibrecht, B., Cooley, S., Rossi, N., Lampton, D., & Knerr, B. (2010). *Impact of game-based training on classroom learning outcomes*. (Study Report 2010-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Triantafillou, E., Georgiadou E., & Economides A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, 50, 1319-1330.
- U.S. Army Training and Doctrine Command. (2011). *The U.S. Army learning concept for 2015* (TRADOC Pamphlet 525-8-2). Fort Monroe, VA: U.S. Army.
- Vogel, J. J., Greenwood-Ericksen, A., Cannon-Bowers, J. A., & Bowers, C. A. (2006). Using virtual reality with and without gaming attributes for academic achievement. *Journal of Research on Technology in Education*, 39(1), 105-118.
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70-84.
- Zhao, Y., Lei, J., Yan, B., Lai, C., & Tan, H. S. (2005). What makes the difference? A practical analysis of research on the effectiveness of distance learning. *Teachers College Record*, 107 (8), 1836-1884.

Appendix A: Demographics Questionnaire

TIME IN MILITARY

How many total years of military service have you completed? (Include time in current and previous tours and services)

_____ years

RANK

What is your current rank or grade?

- | | | |
|------------------------------|------------------------------|------------------------------|
| ☐ CW5 | ☐ CSM | ☐ SSG |
| ☐ CW4 | ☐ SGM | ☐ SGT |
| ☐ CW3 | ☐ 1SG | ☐ SPC |
| ☐ CW2 | ☐ MSG | ☐ PFC |
| ☐ WO1 | ☐ SFC | ☐ PV2 |
| | | ☐ PV1 |

TIME IN RANK

How many months have you served in your current rank, grade, or pay level?

_____ months

MOS

What is your Military Occupational Specialty (MOS)?

- ☐ 25U Signal Support Systems Specialist
☐ Other (please specify):

DEPLOYMENTS

How many times have you been deployed? Include current deployment if applicable. Enter a whole number (e.g., 2).

Number of deployments: _____

RADIO TRAINING

For which of the following radios have you received training? (check all that apply)

- ☐ AN/PRC-25
☐ AN/PRC-77
☐ AN/PRC-117F (SATCOM)

- ☐ SINCGARS
 - ☐ AN/PRC-148 V1/V2 (MBITR)
 - ☐ AN/PRC-148 V3 or newer (JEM)
 - ☐ AN/PRC-152
 - ☐ AN/PRC-154
 - ☐ Other(s) – Please specify
-

RADIO EXPERIENCE WHILE DEPLOYED

Which of the following radios have you operated while deployed? (check all that apply)

- ☐ AN/PRC-25
 - ☐ AN/PRC-77
 - ☐ AN/PRC-117F (SATCOM)
 - ☐ SINCGARS
 - ☐ AN/PRC-148 V1/V2 (MBITR)
 - ☐ AN/PRC-148 V3 or newer (JEM)
 - ☐ AN/PRC-152
 - ☐ AN/PRC-154
 - ☐ Other(s) – Please specify
-

MBITR EXPERIENCE

Have you used an MBITR or JEM prior to this training? If yes, how many months of experience do you have using an MBITR or JEM?

- ☐ Yes with _____ months of experience
- ☐ No

MBITR/JEM EXPERTISE

Rate your level of expertise using an MBITR/JEM (choose one answer that best describes your expertise).

- ☐ I have never used an MBITR/JEM
- ☐ I have used an MBITR/JEM for only standard communication
- ☐ I have used an MBITR/JEM and performed simple troubleshooting in usual circumstances to ensure I was able to communicate with others
- ☐ I have used an MBITR/JEM and performed advanced troubleshooting in unusual circumstances that involved adapting and making complex decisions

EXPERIENCE LOADING KEYS

Rate your level of experience loading keys on an MBITR/JEM (choose one answer that best describes your experience).

- () I have never performed a keyfill on an MBITR/JEM and have never seen it being performed
- () I have never performed a keyfill on an MBITR/JEM but I have seen it being performed
- () I have performed a keyfill on an MBITR/JEM but it is not part of my usual duties
- () I have performed a keyfill on an MBITR/JEM and it is part of my usual duties

JEM-RELATED TASK EFFICACY

Rate your degree of confidence by recording a number from 0 to 100 using the scale given below:

0	10	20	30	40	50	60	70	80	90	100
Cannot do at all					Moderately can do					Highly certain can do

How confident are you that you can complete the following tasks on the JEM as of now? Confidence (0-100)

- Preventive Maintenance Checks and Services (PMCS) and startup procedures _____
- Keyfill for frequency hopping (full load set) _____
- Programming channels in basic/plain-text and SINCGARS/secure modes _____
- Transferring information from one radio to another through cloning _____
- Zeroizing the radio to clear programming, COMSEC keys or both _____

Appendix B: Pre-Test

Multiple Choice Questions: Please select only ONE best answer to each question.

Q1. The number of programmable channels for the JEM is:

- A. 32
- B. 64
- C. 128
- D. 256

Q2. Which of the following is true of both COMSEC and TRANSEC?

- A. They improve the security of communications between military forces by limiting the opportunity for messages to be intercepted and exploited.
- B. Messages cannot be received without cryptoanalytic equipment to decode the signal.
- C. They rely on the Advanced INFOSEC Machine to perform robust algorithm processing.
- D. They rely on Traffic Encryption Keys to encrypt and decrypt voice and data.



Q3. The arrow in the image is pointing towards which of the following components:

- A. Push-to-Talk Switch
- B. Mechanical Interlock
- C. Squelch Disable
- D. Programmable Function Keys

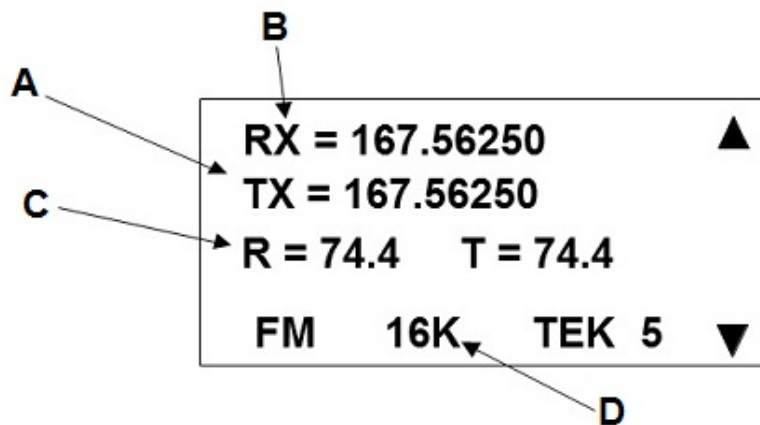
- Q4. Which of the following cables enables two JEMs to be connected to receive and send voice or data traffic?**
- A. Digital data cable
 - B. Personal Data Controller cable
 - C. SINCGARS data adapter cable
 - D. Retransmission cable
- Q5. If the JEM is unexpectedly rebooting AND showing lower than normal transmission range, which area should you first inspect in order to resolve the problem?**
- A. The connection between the radio and antenna
 - B. The side connector
 - C. The audio connector
 - D. The hardware screws on the radio and battery connectors
- Q6. Which of the following statements accurately describes the Frequency Hopping feature?**
- A. It automatically changes the frequency if enemy forces are using that frequency.
 - B. It requires both a Storage KEK and at least one TEK to operate.
 - C. It can be used to transmit and receive voice, but not other data.
 - D. It improves resistance to jamming and interference.
- Q7. PJC is an acronym for:**
- A. Principal JRTC Communication
 - B. Primary Joint Commander
 - C. Personal JEM Compilation
 - D. Private JOSEKI Component
- Q8. Which of the following is an appropriate corrective action if the JEM display has an "ERROR" message?**
- A. Enable the side connector.
 - B. Panic zeroize the JEM.
 - C. Load frequency hopsets.
 - D. Check for low battery power.

Q9. The storage temperature of the JEM ranges from:

- A. -43° C to +61° C (-45° F to 142° F)
- B. -33° C to +71° C (-27° F to 160° F)
- C. -23° C to +81° C (-9° F to 179° F)
- D. -13° C to +91° C (9° F to 196° F)

Q10. The JEM replaces the combined use of which of the following types of equipment:

- A. Multiple Man Pack, Handheld Radios, and COMSEC equipment.
- B. Multiple Man Pack, Handheld Radios, and Infrared tracking.
- C. Infrared tracking, Multiple Man Pack, and COMSEC equipment.
- D. Infrared tracking, Handheld Radios, and COMSEC equipment.



Q11. The image shows Basic Programming Screen 2. Which of the labeled functions sets the frequency of the sub-audible squelch tone that accompanies the voice signal?

- A. A
- B. B
- C. C
- D. D

Q12. If the JEM does not respond to the PTT button for keyfill, the appropriate corrective action is to:

- A. Power down the JEM.
- B. Switch to secure mode.
- C. Change the current frequency.
- D. Set the radio for internal audio operation.

Q13. The primary purpose of groups in the JEM is to:

- A. Maximize the number of channels available.
- B. Differentiate channels used by different units on the battlefield.
- C. Group channels together that use similar frequencies.
- D. Allow users to organize channels to suit their needs.

Q14. Each JEM channel may be named using which of the following formats:

- A. 5-digit alphanumeric
- B. 5-digit numeric
- C. 7-digit alphanumeric
- D. 7-digit numeric

Short Answers: Please provide a brief written response to each of the questions below.

1. List the standard components of the AN/PRC-148 (JEM).
2. Describe one potential consequence of a missing O-ring on the battery.
3. When would you choose to use the 1.2 meter VHF FM antenna instead of the 34 centimeter broadband UHF/VHF antenna?
4. What are the basic procedures for loading COMSEC keys on the AN/PRC-148 (JEM radio)?
5. Explain why you would select Mode23 over the other keyfill options, such as DS-101 or DS-102.
6. While programming in basic plain text mode, in which field would you enter the designated frequency?
7. While programming in SINCGARS frequency hopping mode, what do you enter in place of a frequency?
8. What are the general procedures for Cloning the AN/PRC-148 (JEM)?
9. Which of the following are transferred during cloning operations: channel programming, COMSEC keys, Julian date, Zulu time, global programming such as TX timeout.
10. How would you panic zeroize the AN/PRC-148 (JEM)?
11. What zeroize option would you use to delete both programming and COMSEC keys on the AN/PRC-148 (JEM)?

Appendix C: Post-Test

Multiple Choice Questions: Please select only ONE best answer to each question.

- Q1. Which of the following is a true statement regarding TEKs and KEKs?**
- A. Zeroizing the TEKs will prevent the JEM from loading KEKs over the air.
 - B. Both TEKs and KEKs can be transferred to the JEM over the air.
 - C. Loading KEKs onto the JEM allows TEKs to be encrypted and decrypted.
 - D. The JEM can transmit and receive encrypted voice and data communications as long as a TEK or a KEK is present.
- Q2. The RF connector is used to connect which of the following external components:**
- A. Cloning cable
 - B. PC programming cable
 - C. Digital data cable
 - D. Vehicle adapter
- Q3. Which of the following statements is true of both SINCGARS and HAVEQUICK I/II mode?**
- A. They operate in the VHF band.
 - B. They are used for air-to-air and air-to-ground communication with high-speed aircraft.
 - C. They are capable of frequency hopping to provide transmission security (TRANSEC).
 - D. They are capable of transmitting both voice and data.
- Q4. Which of the following outcomes would result if the PJC is lost?**
- A. The JEM will not accept KEKs or TEKs.
 - B. The JEM will not operate using TRANSEC frequency hopping.
 - C. All previously stored group and channel information will be lost.
 - D. The Frequency Hopping Net Synch Time will be reset.

Q5. The JEM battery life is:

- A. 4 hours at 5 watts
- B. 6 hours at 5 watts
- C. 8 hours at 5 watts
- D. 10 hours at 5 watts

Q6. What happens to the TX power if the operating mode or modulation type is changed?

- A. Nothing--it remains the same.
- B. It clears to 0 watts.
- C. It resets to 1 watt.
- D. It resets to 5 watts.

Q7. What should be done to clear an "ALARM" message?

- A. Slide down the mechanical interlock switch and turn the ON/OFF switch.
- B. Press PTT.
- C. Press ESC.
- D. Press and hold the squelch disable button for 5 seconds.

Q8. The ability to transmit analog or digitized voice is a JEM transmitter characteristic available in:

- A. Plain mode only
- B. Secure mode only
- C. Both plain and secure modes
- D. Neither plain mode nor secure mode

Q9. The operating temperature of the JEM ranges from:

- A. -31° C to +60° C (-24° F to 140° F)
- B. -21° C to +70° C (-6° F to 158° F)
- C. -11° C to +80° C (12° F to 176° F)
- D. 0° C to +90° C (32° F to 194° F)

Q10. The power output of the JEM ranges from

- A. 0.1 - 5 watts
- B. 0.1 - 15 watts
- C. 0.1 - 45 watts
- D. 0.1 - 135 watts

Q11. Which of the following describes a scenario for selecting "Single Channel" in the SINGARS programming screen?

- A. When you are unable to encrypt communications due to lost TEKs or KEKs.
- B. When enemy jamming equipment is active in the area of operations.
- C. When a frequency hopping net is not active.
- D. When you know that you will not need to switch channels for an extended period of time.

Q12. The SCAN function allows the operator of the JEM to:

- A. Monitor traffic on numerous single channel AM and FM voice channels.
- B. Scan the JEM's systems for faults or corrupt hardware components.
- C. Monitor multiple FH channels to synchronize the JEM to the FH Net Synch Time.
- D. Scans FH net channels to locate the home channel for Electronic Remote Fill.

Q13. If the FH Net Synch Time is not performed correctly, what could result?

- A. Inability to communicate with other JEMs in frequency hopping HAVEQUICK mode.
- B. Delay in transmission with other JEMS.
- C. Inability to communicate with other JEMs in frequency hopping SINGARS mode.
- D. Interference between JEMs during retransmission

Q14. All of the following are appropriate corrective actions when troubleshooting in BASIC PLAIN mode EXCEPT:

- A. Check that CTCSS tones are set correctly.
- B. Check that your channel is set to the correct frequency.
- C. Check that the modulation type is set correctly.
- D. Check that COMSEC is loaded.

Appendix D: Work sample task

Training Research Assessment Project

JEM Assessment

This assessment consists of 5 tasks you will be asked to perform on the JEM. Each task contains a series of actions you must perform. Instructions for each task are presented on a separate page within this package. Please do not turn to the next page until you are instructed to do so.

At the beginning of each task, you will be given time to review the actions required to perform the given task. You can ask questions for clarification only at this time. Once you are ready to begin the task, the test administrator will start the clock. You will then proceed with completing all the actions listed in the instructions for the task.

While you are performing the actions for a task, the test administrator will observe and rate your performance. During this time, you cannot ask the test administrator any questions.

When the time limit is reached, the test administrator will stop the clock and you will be instructed to stop any actions immediately.

At the end of each task, the test administrator may ask you one or more follow-up questions related to the task. You will be asked to respond either through oral answers or demonstrated actions. This portion of the test will not be timed.

We ask that you try your best to complete this assessment. However, your score on this assessment will be used for our research purposes only and will not be in any way associated with your name upon its completion.

Any questions?

Do NOT turn the page until you are instructed to do so.
--

Task 1: PMCS

Actions required:

1. Identify verbally and inspect each component on the checklist. You must verbally identify any missing components or issues with the components.

JEM Checklist
• External mic • RTU • Battery • VHF FM antenna • Keyfill cable • Broadband UHF/VHF antenna • Cloning cable

2. Once you are complete with identifying and inspecting the components, assemble the radio by attaching the battery, antenna and external mic.

3. Turn on the radio.

4. Demonstrate how to turn on the backlight.

5. Demonstrate how to adjust the tone volume for the radio.

6. Demonstrate how to adjust the brightness of the display.

Time limit:

You will have 6 minutes to complete this task

Do NOT turn the page until you are instructed to do so.
--

Task 2: Keyfill

Actions required:

1. Perform a keyfill for frequency hopping (full load set). Indicate verbally which keyfill option you would use.
2. Indicate verbally when the radio is ready to receive the load set. You are not required to operate the SKL.

Time limit:

You will have 3 minutes to complete this task.

Do NOT turn the page until you are instructed to do so.

Task 3: Programming

Actions required:

1. Program Channel 001 in basic/plain-text using the following information:

- Channel name: PILOTA
- Frequency: 35.000
- Power: 3.0W
- TEK: 4
- Squelch: 4
- Fade: 1.0S
- Traffic Rate: 12K
- MOD Type: FM

2. Program Channel 002 in SINCGARS/secure using the following information:

- Channel name: PILOTB
- net ID: 333
- SINCGARS: 1
- Offset: 0
- Power: 3.0W
- TEK: 4
- Squelch: 4
- Fade: 1.0S
- DATA Rate: 9600N
- MOD Type: FM
- ECCM: FH

3. Enter the Julian date and Zulu time provided by the test administrator;

4. Establish communication with target radio through the SINCGARS channel programmed above (PILOTB);

5. Troubleshoot any issues that may arise.

Time limit:

You will have 14 minutes to complete this task.

Do NOT turn the page until you are instructed to do so.

Task 4: Cloning

Actions required:

1. Transfer all programming information from your radio to a new radio through cloning;
2. Return both radios to normal operations after cloning;
3. Verify that cloning is successful by doing a radio check on PILOTA (call the TX radio from RX radio).
4. Troubleshoot any issues that may arise.

Time limit:

You will have 5 minutes to complete this task.

Do NOT turn the page until you are instructed to do so.

Task 5: Zeroizing

Actions required:

1. Demonstrate how you would panic zeroize the TX radio.
2. Demonstrate how you would zeroize the RX radio to keep all keys but delete all programming.

Time limit:

You will have 3 minutes to complete this task.

This is the end of the test packet. Thank you for your participation!

Last 4 digits of SSN: _____

Administration Guide

Prior to the tasks

Setup:

1. The second radio (Cloning RX) will need to be keyfilled.
2. The second radio (Cloning RX) will need to be set up with the target frequency (SINCGARS – Net ID: 333; Julian date & Zulu time) for the Soldier to call out to.
3. Disable the side connector on both radios (Cloning RX & TX).
4. Prepare the SKL for keyfill.
5. Be sure to read and follow the setup for each task prior to beginning the task.

Task	Score	Time of completion
Task 1: PMCS		
Task 2: Keyfill		
Task 3: Programming		
Task 4: Cloning		
Task 5: Zeroizing		
Total		

Task 1: PMCS

Setup:

1. A battery with a missing O-ring, 34 centimeter VHF/UHF broadband antenna, external mic, keyfill cable, cloning cable are laid out (in that order) or placed in a box.
2. One O-ring from the battery and the 1.2 meter VHF antenna, if available, should be kept out of sight.

Hand Soldier instructions, allow her/him to read over instructions, ask if s/he has any questions. Instruct Soldier to complete all listed actions within 6 minutes and hand over radio upon completion of all actions in this task. Say “start” and start clock.

Fill out scoring sheet below as Soldier completes task.

Scoring:

Identify all the components		____/1
• RTU • External mic • Battery • 34 centimeter Broadband UHF/VHF antenna • Keyfill cable • Cloning cable	_____ _____ _____ _____ _____ _____	
Identify the missing O-ring in the battery		____/1
Identify the 1.2 meter VHF FM antenna missing		____/1
Turn on backlight		____/1
Adjust tone volume		____/1
Adjust brightness of display		____/1
Total time for completion		____:____
Oral question #1: Antenna		____/1
Oral question #2: Missing O-ring		____/1
Total points for this task		____/8

Stop clock, record time for completion above, check all actions, then ask the following question.

Questions:

1. When would you choose to use the 1.2 meter VHF FM antenna instead of the 34 centimeter broadband UHF/VHF antenna?
2. Describe one potential consequence of a missing O-ring on the battery?

Record Soldier responses:

Answer key:

1. Distance (1.2 meter antenna will reach farther because it's a line-of-sight radio).
2. Acceptable answers include: battery will not be sealed properly (but can still be attached), dust/moisture in the radio – causes interference, unexpected shut-down, limited frequency range...

Score Soldier response, record score on scoring sheet, and make expert rating if Soldier completed everything correctly.

Task 2: Keyfill

Setup:

1. Same setting from previous step
2. Prepare the SKL for keyfill (have a keyfill operator stand by and hold keyfill device) – **ensure it is on, logged in and ready to go before starting the timer.**
3. Have the keyfill operator press “load” on SKL when Soldier indicates the radio is ready for keyfill.

Hand Soldier instructions, allow her/him to read over instructions, ask if s/he has any questions. Instruct Soldier to complete all listed actions within 3 minutes and hand over radio upon completion of all actions in this task. Instruct Soldier to indicate when the radio is ready for keyfill. Say “start” and start clock.

Fill out scoring sheet below as Soldier completes task. If Soldier indicates an incorrect keyfill option (any option other than “Mode23”), instruct her/him to select “Mode23” instead and allow her/him to continue the task. The Soldier will not get the point for programming the radio but may still receive the remaining points for the task.

Scoring:

Program the radio for keyfill	____/1
Press PTT to complete keyfill	____/1
Total time for completion	____:____
Oral question #1: Classification of keys	____/1
Oral question #2: Mode23	____/1
Total points for this task	____/4

Stop clock (after PTT pressed), record time for completion above, check all actions, then ask the following question.

Questions:

1. Demonstrate on the radio how to check the classification of keys currently loaded in the radio.
2. Explain why you would select Mode23 over the other keyfill options, such as DS-101 or DS-102.

Record Soldier response:

Answer key:

1. Alt-Mode --> Status --> key database --> Saville
2. Mode 23 keyfill is a method for loading all COMSEC keys and TRANSEC (SINCGARS) hopsets and lockout sets at one time.

Task 3: Programming

Setup:

1. Same setting from previous step.
2. Complete the keyfill task if the learner failed to do so in the previous task.
3. **Switch the radio to side audio.**
4. Provide Julian date and Zulu time (display on phone from www.zulutime.net).

Hand Soldier instructions, allow her/him to read over instructions, ask if s/he has any questions. Instruct Soldier to complete all listed actions within 14 minutes and hand over radio upon completion of all actions in this task. Say “start” and start clock.

Fill out scoring sheet below as Soldier completes task.

Scoring:

Program the basic channel		____/1
• Channel name: PILOTA • Frequency: 35.000 • Power: 3.0W • TEK: 4 • Squelch: 4 • Fade: 1.0S • Traffic Rate: 12K • MOD Type: FM • Plain	_____ _____ _____ _____ _____ _____ _____ _____ _____	
Program the SINCGARS channel		____/1
• Channel name: PILOTB • net ID: 333 • SINCGARS: 1 • Offset: 0 • Power: 3.0W • TEK: 4 • Squelch: 4 • Fade: 1.0S • DATA Rate: 9600N • MOD Type: FM • ECCM: FH • Secure	_____ _____ _____ _____ _____ _____ _____ _____ _____ _____ _____ _____	
Enter the Julian date and Zulu time		____/1
Switch the radio to internal audio		____/1
Press PTT to establish communication		____/1

Total time for completion	____ : ____
Total points for this task	____ / 5

Stop clock, record time for completion above, and check all actions.

Task 4: Cloning

Setup:

1. Same setting from previous step. Present a second radio as the receiving radio.
2. **Check and make sure the side connector is disabled on both TX and RX radios.**
3. **Lock the keypad on the RX radio.**

Hand Soldier instructions, allow her/him to read over instructions, ask if s/he has any questions. Instruct Soldier to complete all listed actions within 5 minutes and hand over radio upon completion of all actions in this task. Say “start” and start clock.

Fill out scoring sheet below as Soldier completes task.

Scoring:

Connect the cloning cable		____/1
• Connect the cloning cable to the TX radio • Connect the cloning cable to the RX radio	_____ _____	
Enable side connectors		____/1
• Enable side connector on the TX radio • Enable side connector on the RX radio	_____ _____	
Program radios for cloning		____/1
• Program sending radio for cloning (set to Cloning TX) • Program receiving radio for cloning (set to Cloning RX)	_____ _____	
Unlock the keypad on the receiving radio (ALT + ESC)		____/1
Radio check on PILOTA		____/1
Total time for completion		____:____
Oral question		____/1
Total points for this task		____/6

Stop clock, record time for completion above, check all actions, then ask the following question.

Question:

1. Which of the following are transferred during cloning operations: channel programming, COMSEC keys, Julian date, Zulu time, global programming such as TX timeout.

Record Soldier response:

Answer key:

1. Channel programming and global programming such as TX timeout.

Task 5: Zeroizing

Setup:

1. Same setting from previous step.

Hand Soldier instructions, allow her/him to read over instructions, ask if s/he has any questions. Instruct Soldier to complete all listed actions within 3 minutes and hand over radio upon completion of all actions in this task. Say “start” and start clock.

Fill out scoring sheet below as Soldier completes task.

Scoring:

Panic zeroize the Cloning TX radio (twist on/off switch while pulling down mechanical interlock)	____/1
Zeroize the Cloning RX radio through programming (set defaults)	____/1
Total time for completion	____:____
Oral question	____/1
Total points for this task	____/3

Stop clock, record time for completion above, check all actions, then ask the following question.

Question:

1. Demonstrate on the TX radio how you would zeroize the radio to delete both programming and COMSEC keys.

Record Soldier response:

Answer key:

1. Select "clear all" in the zeroize menu

Appendix E: Reactions Measures

Reaction Survey: Mobile (administered in prototype only)

Item	Format	Dimension
The technology interface was easy to use.	5-point disagree/agree scale	Usability
The technology allowed for easy review.	5-point disagree/agree scale	Usability
I was able to access the training with minimum assistance.	5-point disagree/agree scale	Usability
I was able to successfully operate the functionality within the training.	5-point disagree/agree scale	Usability
I am satisfied with the technology interface.	5-point disagree/agree scale	Usability
I have used technology like this before.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I am familiar with this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I felt comfortable using the technology from the very beginning.	5-point disagree/agree scale	Comfort/ familiarity with Technology
It took me a while to get used to this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
The session objectives were met.	5-point disagree/agree scale	Content/ISD
The design of the training was an effective way to present the subject matter.	5-point disagree/agree scale	Content/ISD
The material was presented in a logical sequence so that it helped me understand the subject matter.	5-point disagree/agree scale	Content/ISD
The media (i.e., graphics and animated sequences) appropriately illustrate the points being discussed.	5-point disagree/agree scale	Content/ISD
The design and presentation of material motivated me to learn.	5-point disagree/agree scale	Content/ISD
Overall, I am pleased with the way the training was presented.	5-point disagree/agree scale	Content/ISD
I was engaged with the topic at hand throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I found it difficult to focus throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training

Item	Format	Dimension
I can recall most of what was taught during the training.	5-point disagree/agree scale	Engagement/ Attention during training
I paid close attention to all the material throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I tried to access all training resources available to me, including optional contents such as the glossary.	5-point disagree/agree scale	Engagement/ Attention during training
The training kept me focused the entire time.	5-point disagree/agree scale	Engagement/ Attention during training
Did you experience any problems accessing the training?	Yes/No	Technical Issues
If yes, please describe.	Open-ended	Technical Issues
Would you expect any challenges in completing the mobile training on your own? What features or guidelines can you think of that would lessen these challenges?	Open-ended	User Comment
What are some potential uses of the mobile application in the field?	Open-ended	User Comment

Reaction Survey: Virtual Classroom (administered in prototype only)

Item	Format	Dimension
The technology interface was easy to use.	5-point disagree/agree scale	Usability
The technology allowed for easy review.	5-point disagree/agree scale	Usability
I was able to access the training with minimum assistance.	5-point disagree/agree scale	Usability
I was able to successfully operate the functionality within the training.	5-point disagree/agree scale	Usability
I am satisfied with the technology interface.	5-point disagree/agree scale	Usability
I have used technology like this before.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I am familiar with this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I felt comfortable using the technology from the very beginning.	5-point disagree/agree scale	Comfort/ familiarity with Technology
It took me a while to get used to this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
Since my instructor could not read my confusion and call it out, I kept quiet about questions I had.	5-point disagree/agree scale	Interactivity
I felt more comfortable asking questions virtually.	5-point disagree/agree scale	Interactivity
The technology enabled me to interact with instructor without any issues.	5-point disagree/agree scale	Interactivity
The technology allowed me to interact with my fellow classmates easily.	5-point disagree/agree scale	Interactivity
I found it difficult to understand the material since my instructor was in a remote location.	5-point disagree/agree scale	Interactivity
The session objectives were met.	5-point disagree/agree scale	Content/ISD
The design of the training was an effective way to present the subject matter.	5-point disagree/agree scale	Content/ISD
The material was presented in a logical sequence so that it helped me understand the subject matter.	5-point disagree/agree scale	Content/ISD
The media (i.e., graphics and animated sequences) appropriately illustrate the points being discussed.	5-point disagree/agree scale	Content/ISD
The design and presentation of material motivated me to learn.	5-point disagree/agree scale	Content/ISD
Overall, I am pleased with the way the training was presented.	5-point disagree/agree scale	Content/ISD

Item	Format	Dimension
How satisfied are you with the instructor's knowledge of training material and subject matter?	5-point dissatisfied/satisfied scale	Instructor
How satisfied are you with the instructor's ability to keep interest of the Soldiers?	5-point dissatisfied/satisfied scale	Instructor
How satisfied are you with the instructor's presentation and explanation of training materials?	5-point dissatisfied/satisfied scale	Instructor
How satisfied are you with the instructor's responsiveness to Soldier questions and problems?	5-point dissatisfied/satisfied scale	Instructor
How satisfied are you with instructor's overall effectiveness?	5-point dissatisfied/satisfied scale	Instructor
I was engaged with the topic at hand throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I found it difficult to focus throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I can recall most of what was taught during the training.	5-point disagree/agree scale	Engagement/ Attention during training
I paid close attention to all the material throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
The training kept me focused the entire time.	5-point disagree/agree scale	Engagement/ Attention during training
I found I lost interest since my instructor was in a remote location.	5-point disagree/agree scale	Engagement/ Attention during training
I found I lost interest since my classmates were in different locations.	5-point disagree/agree scale	Engagement/ Attention during training
Did you experience any problems accessing the training?	Yes/No	Technical Issues
If yes, please describe.	Open-ended	Technical Issues
Were there any portions of the training that were less effective due to the instructor being virtual? Were you satisfied with the capability of the virtual instructor in answering your questions and demonstrating radio functionality?	Open-ended	User Comment

Reaction Survey: Collaborative Scenario (administered in prototype only)

Item	Format	Dimension
The technology interface was easy to use.	5-point disagree/agree scale	Usability
The technology allowed for easy review.	5-point disagree/agree scale	Usability
I was able to access the training with minimum assistance.	5-point disagree/agree scale	Usability
I was able to successfully operate the functionality within the training.	5-point disagree/agree scale	Usability
I am satisfied with the technology interface.	5-point disagree/agree scale	Usability
I have used technology like this before.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I am familiar with this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
I felt comfortable using the technology from the very beginning.	5-point disagree/agree scale	Comfort/ familiarity with Technology
It took me a while to get used to this type of technology.	5-point disagree/agree scale	Comfort/ familiarity with Technology
The session objectives were met.	5-point disagree/agree scale	Content/ISD
The design of the training was an effective way to present the subject matter.	5-point disagree/agree scale	Content/ISD
The material was presented in a logical sequence so that it helped me understand the subject matter.	5-point disagree/agree scale	Content/ISD
The media (i.e., graphics and animated sequences) appropriately illustrate the points being discussed.	5-point disagree/agree scale	Content/ISD
The design and presentation of material motivated me to learn.	5-point disagree/agree scale	Content/ISD
Overall, I am pleased with the way the training was presented.	5-point disagree/agree scale	Content/ISD
I was engaged with the topic at hand throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I found it difficult to focus throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
I can recall most of what was taught during the training.	5-point disagree/agree scale	Engagement/ Attention during training

Item	Format	Dimension
I paid close attention to all the material throughout the training.	5-point disagree/agree scale	Engagement/ Attention during training
The training kept me focused the entire time.	5-point disagree/agree scale	Engagement/ Attention during training
Did you experience any problems accessing the training?	Yes/No	Technical Issues
If yes, please describe.	Open-ended	Technical Issues
Did you find the collaborative setup for the game-based exercise helpful or distracting (vs. if you had completed the exercise on your own). Why?	Open-ended	User comment

Reaction Survey: General (administered in all conditions)

Item	Format	Dimension
Learning this material was fun.	5-point disagree/agree scale	Perceived Learning
Please rate the following:		
My knowledge of the subject PRIOR to taking this lesson:	5-point poor/excellent scale	Perceived Learning
My knowledge of the subject AFTER to taking this lesson:	5-point poor/excellent scale	Perceived Learning
My ability to apply the strategies and techniques presented to an actual situation in this subject area PRIOR to taking this lesson:	5-point poor/excellent scale	Perceived Learning
My ability to apply the strategies and techniques presented to an actual situation in this subject area AFTER to taking this lesson:	5-point poor/excellent scale	Perceived Learning
To what do you attribute the differences in your PRIOR and AFTER responses?	Open-ended	Perceived Learning
Overall, I have learned a lot from this training.	5-point disagree/agree scale	Perceived Learning
It is clear to me that the people conducting the training understand how I will use what I learn.	5-point disagree/agree scale	Learning transfer
This training was relevant to my job in the Army.	5-point disagree/agree scale	Learning transfer
I believe the training will help me do my current job in the Army better.	5-point disagree/agree scale	Learning transfer
I learned something I can apply immediately to my work in the Army.	5-point disagree/agree scale	Learning transfer
I am prepared to train other Soldiers on what I learned in this training.	5-point disagree/agree scale	Learning transfer
I plan to use what I learned on my job in the Army.	5-point disagree/agree scale	Learning transfer
I get excited when I think about trying to use my new learning on my job in the Army.	5-point disagree/agree scale	Learning transfer
I will be using the equipment on my job in the Army after the training.	5-point disagree/agree scale	Learning transfer
The training was of practical value to me.	5-point disagree/agree scale	Learning transfer
I enjoyed this training program.	5-point disagree/agree scale	Satisfaction
My time on the training was well spent.	5-point disagree/agree scale	Satisfaction
I would recommend this training program to other Soldiers.	5-point disagree/agree scale	Satisfaction
What, if anything, would you change about the training?	Open-ended	User Comment

Item	Format	Dimension
Please provide any additional comments you have about the training.	Open-ended	User Comment
What are some capabilities/features that would have made the training more realistic? More user-friendly?	Open-ended	User Comment
Did you experience any unanticipated outcomes (good or bad) in using this training application? Please explain.	Open-ended	User Comment

Appendix F: Privacy Act Statement

Project Title: Evaluation of Mobile and Virtual Training and Assessments for JEM Radio

Authority: The Department of the Army may collect the information requested in this session under the authority of 10 United States Code, Section 2358, “Research and Development Projects.” In accordance with the Privacy Act of 1974 (Public Law 93-579), this notice informs you of the purpose, use, and confidentiality of this session.

Purpose: Over the past year, the U.S. Army Research Institute and contracting firm ICF International have developed prototype training for Soldiers to use the JTRS Enhanced MBIR (JEM), a common piece of signal equipment. The prototype includes training on a mobile device, a virtual classroom, and a virtual world. The purpose of this effort is to determine the effectiveness of the prototype training. Within this line of research we will be investigating (1) whether the prototype training is at least as effective as the traditional classroom-based training, (2) whether computer adaptive testing (CAT) offers value above non-adaptive testing in a training context and (3) whether interim assessments lead to better training outcomes.

Routine Uses: The information and feedback collected from this session will be used to answer key questions regarding how technology based training and assessments can be used to meet the vision of the Army Learning Model (ALM).

Disclosure: Participating in this session is voluntary and you may choose at any time not to participate. There is no penalty for choosing not to participate.

Confidentiality: We will not identify you or attribute comments to any particular participant made in this evaluation. Your responses are anonymous.

Contact: For further information about this project or your rights as a participant, send e-mail to: ARI_RES@conus.army.mil (type “ARI TRAP experiment” in the subject line).

Appendix G: Informed Consent

Project Title: Evaluation of Mobile and Virtual Training and Assessments for JEM Radio

Purpose of the research project: Over the past year, the U.S. Army Research Institute (ARI) and contracting firm ICF International have developed prototype training for Soldiers to use the JTRS Enhanced MBIR (JEM), a common piece of signal equipment. The prototype includes training and assessments on a mobile device, virtual classroom, and virtual world. The purpose of this effort is to determine the effectiveness of the prototype training. Within this line of research we will be investigating (1) whether the prototype training is at least as effective as the traditional classroom-based training, (2) whether computer adaptive testing (CAT) offers value above non-adaptive testing in a training context and (3) whether interim assessments lead to better training outcomes.

What you will be asked to do in this project: You will be asked to voluntarily participate in either the traditional classroom training for the JEM or the prototype JEM training. The classroom training will consist of a lecture and hands-on lab. The prototype training condition will consist of three training applications: The JEM Training mobile application, the JEM Training Virtual Classroom, and the JEM Collaborative Assessment. Within each of these training types you will also be asked to complete a numbers of assessments. You will also be asked to answer questions regarding your satisfaction with the training content and structure. Following this you will be asked to completed several tasks using a JEM radio. Your performance on the training and assessments will not be connected to your name in any way; data collected with be used for research purposes only to evaluate the effectiveness of the prototype training.

Location: Ft. Gordon, GA.

Voluntary participation: Your participation is voluntary; there is no penalty for not participating. You have the right to withdraw from the event at any time. You must be 18 years of age or older to participate.

Time required: Approximately 6-8 hours, spread across two days.

Risks: There are no risks greater than those encountered in everyday computer activities.

Benefits: Information from this session will be used answer critical questions regarding the use of technology-based training and assessments in the Army Learning Model (ALM). Benefits to individual participants include the opportunity to learn about the JEM radio, participate in prototype training, and express opinions about ways in which training can be improved.

Compensation: No compensation will be provided for your participation.

Whom to contact if you have questions about the project: You should send your questions to ARI_RES@conus.army.mil. Reference project name: ARI TRAP Experiment 2013.

Whom to contact about your rights in the project: You should send your questions to ARI_RES@conus.army.mil. Reference project name: ARI TRAP experiment 2013.

Appendix H: List of Acronyms

ALC	Advanced Leadership Course
ALM	Army Learning Model
AN/PRC	Army Navy/Portable Radio Communication
CAT	Computer-adaptive test
IRT	Item response theory
JEM	Joint Enhanced Multiband
MBITR	Multiband Inter/Intra Team Radio
MOS	Military Occupational Specialties
MUVE	Multi-user virtual environment
NCO	Non-Commissioned Officer
NCOA	Non-Commissioned Officer Academy
PLM	Parameter logistic model
PMCS	Preventive maintenance checks and services
SEM	Standard error of measurement
SLC	Senior Leadership Course
TR	Transfer Ratio