This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.

Improving the Army's Assessment of Interactive Multimedia Instruction Courseware

Susan G. Straus, Michael G. Shanley, Rachel M. Burns, Anisah Waite, James C. Crowley

Prepared for the United States Army
Approved for public release; distribution unlimited

ARROYO CENTER

The research described in this report was sponsored by the United States Army under Contract No. W74V8H-06-C-0001.

Library of Congress Cataloging-in-Publication Data

Improving the Army's assessment of interactive multimedia instruction courseware / Susan G. Straus ... [et al.].
p. cm.
Includes bibliographical references.
ISBN 978-0-8330-4727-4 (pbk. : alk. paper)
1. Military education—United States—Computer-assisted instruction—Evaluation. 2. Distance education—United States—Computer-assisted instruction—Evaluation. 3. The Army Distributed Learning Program (U.S.) 4. Computer-assisted instruction—United States—Evaluation. 5. Interactive multimedia—United States—Evaluation. 6. Curriculum evaluation—United States. I. Straus, Susan G.

U408.3.I48 2009
355.0071'5—dc22

2009026430

Published 2009 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

Since 1998, the Army's Training and Doctrine Command (TRADOC) has been engaged in establishing and fielding The Army Distributed Learning Program (TADLP) to enhance and extend traditional methods of learning within the Army's training strategy. This report discusses program-level approaches for evaluating the quality of the Army's interactive multimedia instruction (IMI) courseware, which is used in its distributed learning (DL) program. A fiscal year (FY) 2006 study conducted by the Arroyo Center for TRADOC identified several challenges within TADLP and its IMI courseware component, including the lack of a program-level assessment of course quality.

The present study develops and applies a method of assessing the instructional design features of courseware that could be used to evaluate the quality of Army IMI courseware on an ongoing basis. The report demonstrates the feasibility of this approach, illustrates the kinds of information produced by such an evaluation, and shows how the results can be used to identify specific areas for improvement in courseware and to monitor quality at the program level.

This study will be of interest to persons involved in planning, developing, delivering, and evaluating IMI and other forms of distributed learning.

This research was sponsored by U.S. Army Training and Doctrine Command and was conducted within RAND Arroyo Center's Manpower and Training Program. RAND Arroyo Center, part of the RAND Corporation, is a federally funded research and development

center sponsored by the United States Army. Correspondence regarding this report should be addressed to Susan Straus (sgstraus@rand.org).

The Project Unique Identification Code (PUIC) for the project that produced this monograph is ATFCR07213.

For more information on RAND Arroyo Center, contact the Director of Operations (telephone 310-393-0411, extension 6419; FAX 310-451-6952; email Marcy_Agmon@rand.org), or visit Arroyo's website at http://www.rand.org/ard/.

Contents

Tables

Summary

Since 1998, the Army's Training and Doctrine Command (TRADOC) has been engaged in establishing and fielding The Army Distributed Learning Program (TADLP) to enhance and extend traditional methods of learning within the Army's training strategy. Distributed learning (DL) is intended to speed the pace of learning and allow training to take place when and where soldiers need it. The Army has an expansive vision of a greatly increased role for DL over time.

Given this expectation, an important component of TADLP's overall performance is the quality of its courses, which consist primarily of asynchronous interactive multimedia instruction (IMI). An assessment of IMI quality is necessary for strategic planning—to understand TADLP outputs, to manage budgets devoted to increasing quality, and to identify and implement needed improvements to processes that affect quality. Moreover, ensuring and documenting the quality of IMI courseware are important to show the value of this type of instruction, to gain the buy-in of DL stakeholders, and to secure the resources needed to achieve the program's goals.

A comprehensive evaluation of training quality requires several types of measures and methods:

- Measures of outcomes, including reactions (e.g., learner satisfaction), learning (e.g., performance on course tests), behavior (performance on the job or in subsequent training), and results (effects of training on organizational outcomes) (Kirkpatrick, 1959–1960, 1994).

- Courseware evaluations to determine whether courses effectively cover the relevant knowledge, skills, and abilities (KSAs) and to assess technical and instructional design criteria that influence the quality of the learning experience.
- Test evaluation using qualitative and quantitative approaches to ensure that tests cover relevant KSAs and can discriminate among good and poor performers. Good tests are needed to measure learning and to assess whether performance in training transfers to performance on the job.
- Administrative data, which can point to potential negative or positive aspects of course quality. Examples include enrollment rates, completion rates, cost data, cycle time of courseware production, and responsiveness of the training development process to the need for changes in courseware content.

Purpose of This Report

This report (1) reviews current efforts to evaluate the effectiveness of IMI in TADLP, (2) develops and tests an approach to evaluating the quality of Army IMI courseware, and (3) identifies directions for improvement in IMI courseware and in the evaluation of IMI within TADLP. As in other approaches to courseware evaluation, we focus on the quality of the learning experience (see, e.g., ASTD Certification Institute, 2001–2003) in terms of instructional design and technical features of courseware, rather than on the accuracy or comprehensiveness of substantive course content. In addition to devising evaluation criteria, we demonstrate the feasibility of this assessment approach by applying it to lessons from a sample of IMI courses. We also illustrate the kinds of information produced by such an evaluation and demonstrate how that information can be used to identify areas for improvement in TADLP courseware and to monitor quality at the program level.

TADLP Does Not Have a Systematic Quality Assessment Process

Our research revealed that there are no efforts in TADLP to assess course quality at the program level. Although several types of independent training evaluation are performed in the proponent schools, these assessments are not comprehensive or systematic in terms of the range of evaluation measures collected, standardization of measures, or synthesis of results across courses. In short, TADLP does not currently have sufficient data or capabilities for data integration to gauge overall program effectiveness or to evaluate the effect of quality-improvement initiatives.

RAND Developed an Approach to Courseware Evaluation for TADLP IMI

We developed a set of evaluation criteria and used them to evaluate lessons from a sample of ten high-priority courses fielded under TADLP between 2005 and 2007. Courses represented a range of knowledge and skills. We had online access to two of the courses, and the remaining eight were provided on CDs.

Our evaluation criteria were based on existing standards in the training development community from (1) the American Society for Training & Development Certification Institute E-Learning Courseware Certification Standards (ASTD Certification Institute, 2001–2003), (2) the Center for Cognitive Technology at University of Southern California (provided by Richard Clark, February 5, 2007), and (3) a checklist developed at TRADOC (provided by D. E. MacAllister, Instructional Systems Specialist, May 16, 2007). We drew criteria from these sources to yield a set of standards that (1) focused on what we viewed as the most important issues, (2) struck a balance between level of detail and feasibility, and (3) could be evaluated from the available materials. Our criteria comprised three categories that reflect the major themes of our source checklists:

- Technical criteria, such as the ease of launching a course and navigating through it, the accessibility and usability of supplementary instructional materials, and the availability and quality of technical support.
- Production-quality criteria, such as legibility of graphics and text, characteristics of audiovisual materials, and the use of techniques to maintain learner interest.
- Pedagogical criteria, such as the specificity of learning objectives; characteristics of instruction of concepts, processes, and procedures; the number and quality of checks on learning; the number and quality of opportunities for practice; and the frequency and quality of feedback.

In addition to applying these criteria, we evaluated the IMI level in each lesson.[1]

Three evaluators, or "raters," were trained in how to evaluate the courses. After establishing satisfactory interrater reliability on a sample of lessons, the raters went through two or three modules of each course in much the same way a student would. They rated a total of 79 lessons, which comprised 39 percent of the lessons per course, on average, or approximately 200 hours of instruction. The majority of the lessons focused on instruction of concepts, followed by instruction of procedures.

Some Features of Courseware Quality Do Not Meet Expected Standards

Our content evaluation found both strengths and deficiencies in course quality. The focus of our evaluation was quality at the program level,

[1] The Army Training Support Center (ATSC) defines four levels of interactivity (Wardell, 2006) which involve progressively greater degrees of interaction between the learner and the computer. These range from instruction in which the learner is a passive recipient of information to immersive learning exercises in which the learner interacts with lifelike visual and auditory cues based on realistic mission scenarios.

but ratings can also be reported at the course level to identify specific needs for improvement.

Levels of Interactivity

In general, we found the content somewhat "thin" in some areas, as exemplified by pedagogical shortcomings and lower-than-required levels of IMI, or interactivity. Our ratings show that 4 percent of the lessons (often introductory lessons) were the lowest level (Level 1), 76 percent included some Level 2, and 20 percent included some Level 3. Whereas some basic material may not lend itself to Level 3 IMI, instruction of even higher-level content consisted largely of Level 2 IMI.

Technical Features

The technical features of the courses were found to be generally strong. All the courses were easy to navigate, and cues to the learner's position in the course were easily accessible. However, we identified two technical features needing substantial improvement. First, we could not launch most of the courses provided on CDs without substantial technical assistance. We expect that if Army learners experience similar problems and do not have access to technical support, many of them will give up, which would be reflected in low course completion rates. Second, supplementary instructional resources (i.e., reference materials) were difficult to use. Although most courses had a menu through which the learner could easily access resources such as field manuals (FMs), the concepts on any particular page of instruction were not linked directly to an FM. Therefore, using these resources required substantial effort.

Production-Quality Features

Production-quality features of the courses were strong in many areas. Narration was easy to understand, courses had minimal irrelevant content, and most graphics and text were typically legible. However, some aspects of the audiovisual features need improvement. Courses were rated only moderately effective in the use of animation/video to demonstrate processes and procedures. Narration, while easy to understand, was often very slow and could not always be disabled, which inhibited

the pace of learning. Significant improvement is also needed in using media to engage learners and in eliminating sensory conflicts.

Pedagogical Features

The pedagogical aspects of the courseware are the most important criteria in our quality assessment. Pedagogical strengths of the courses include clear descriptions of lesson objectives, an appropriate order of lessons, clear and comprehensive definitions of concepts, and opportunities for learners to correct their strategies in checks on learning and practical exercises.

However, pedagogy was also the area most in need of improvement. A pervasive problem in many of the courses was a lack of examples from the job or mission environments; this occurred in instruction of both concepts and procedures. Courses also need to do a better job of demonstrating procedures and providing explanations of why procedures work the way they do, so that learners can better understand the concepts and skills taught and can thus be prepared to adapt their behavior in nonstandard situations. Finally, in most of the courses we evaluated, practical exercises did not provide sufficient opportunities to integrate concepts and to practice procedures; there were typically too few opportunities, the exercises did not progress from simple to complex problems, and they did not include both part-task and whole-task practice. In short, the courseware was deficient with respect to two critical aspects of training: effective explanations of procedures, and opportunities for practice.

IMI Courseware Should Be Improved in Several Ways

Our analysis suggests a number of potential improvements to IMI courseware:

- Directly linking course content to relevant information in FMs and other supplemental resources will provide learners with a powerful tool for rapidly deepening or sustaining their knowledge in specific task areas. Allowing learners to replace narration

with text could also increase their pace of proceeding through the material.

- Correctly applied, higher levels of interactivity can support improvements in instruction by providing more powerful techniques to improve the relevance and realism of the courseware and the fidelity of practical exercises, thereby enhancing opportunities for transfer of training. Higher-level IMI also can increase learner engagement.
- More generally, in the context of the current phased approach to structured training (in which a self-paced DL phase can be completed up to a year prior to a residential phase), the Army should consider emphasizing the use of IMI to teach concepts and processes rather than procedures. Whereas it is possible to provide practice opportunities for administrative tasks such as filling out forms or logs in IMI (as some courses did), it is clearly much more difficult to provide opportunities to practice physical performance tasks, such as entering and clearing a building, using hand grenades, immobilizing equipment, or tying an eye splice. The use of IMI may be best reserved for training procedures when (1) the procedures can be realistically practiced within the context of IMI (e.g., using software, completing forms, performing calculations) or with the addition of simple and inexpensive job aids; (2) the learning is not subject to rapid decay or is easily refreshed; (3) DL can serve as a supplement to residential training, e.g., when the IMI can be assigned as "homework" immediately preceding a practical exercise; or (4) exported training can be supported by a high level of instructor-student interaction.

We note that TADLP has implemented changes in processes that may have translated into improvements in the quality of courses being developed today. However, the success of these new efforts has not been documented. Moreover, given the range of these initiatives, it is unlikely that all the deficiencies noted in our sample have been corrected. Research is needed to determine the success of new initiatives aimed at improving quality and to guide the Army toward the most effective use of IMI.

TADLP's Assessment Program for IMI Should Also Be Improved

One of the conclusions from our research is that program-level assessments of courseware quality are needed on an ongoing basis. The method used in this study has a number of strengths and can effectively fill at least a part of that need for TADLP. Our approach provides a systematic method of evaluation using multiple raters and a comprehensive set of criteria based on standards proposed by experts in training development and assessment. The method yields quantifiable data, enabling integration of results across courses, schools, or other units of analysis. Once criteria are developed and raters are trained, lessons can be evaluated relatively efficiently. In short, we believe this evaluation method is practical, can provide the Army with valuable information about courseware quality, and points to directions for needed quality improvements. We recommend adoption of this approach for evaluating TADLP courseware.

Our approach to course evaluation does have several limitations. For example, some of the ratings are subjective; all criteria were weighted equally regardless of impact; the sample size of courses was relatively small; and most courses that were provided on CDs did not include some features of the courseware. These limitations can be effectively addressed in future efforts. Consequently, we recommend that TADLP do the following:

- Reassess and refine the criteria used in this study, particularly to reflect fully operational courseware.
- Where possible, establish objective standards for criteria such as the degree to which lessons provide sufficient examples, checks on learning, and practical exercises.
- Establish weights for the criteria according to their relevance to course objectives and the resource implications of remedying deficiencies.
- Evaluate fully functional versions of courses in order to assess features such as course tests, bookmarks, and "live" technical support.

To implement a program of courseware evaluation, we offer the following steps as elements of one potential approach:

- TADLP would assign responsibility for IMI courseware evaluation to a team of two or three individuals trained in principles of instructional design. This task would constitute a portion of each staff member's general responsibilities.
- The evaluation team would work together to refine the evaluation criteria, with input from proponent school staff, subject matter experts (SMEs), and staff from Army Training Support Center (ATSC).
- Proponent schools would provide access to completed courseware as well as to courses under development.
- The evaluation team would evaluate at least a sample of the courseware and provide the completed checklist to the proponent school and contractor, along with specific suggestions for improvement.
- The evaluation team would create quarterly reports and an annual report that summarizes results at the program level.
- These activities would be supported by a website to enable proponent school staff and contractors to download the checklist and supporting material.

We recommend that at the outset, the program focus on providing feedback to individual schools about their courses and on reporting aggregate IMI quality over time. Once the evaluation program becomes more established, the data might be used more aggressively to improve courseware quality. For example, results could be aggregated by school and by contractor as part of a larger evaluation program that fosters continuous quality improvement. Relevant evaluation criteria could be incorporated into the schools' courseware validation processes to standardize and improve these efforts or into the language of DL delivery orders. However, initiating these steps may require formulating new policies to support them, including policies relating to contracts, funding mechanisms, appropriate incentives, and development processes.

As noted earlier, the type of evaluation presented here is only one aspect of assessing training effectiveness. We recommend that TADLP

pursue a comprehensive approach to assessment of training quality at the program level. In addition, many of the methods and measures used to evaluate training can be facilitated by the use of information technology (IT). For example, a broad-based assessment would include the following components, many of which could be tested on a pilot basis using a sample of DL courses:

- **Learner reactions.** The Army could develop a core set of questions to administer to students across IMI courses, develop an IT platform to enable schools to create/customize surveys, and create automated scoring and reporting capabilities.
- **Learning (pre/post comparisons).** The Army could develop an IT platform to administer course pretests and posttests, along with automated procedures to analyze and report test scores across courses or schools.
- **Learning (knowledge retention).** The Army could administer and score follow-up tests relatively efficiently and inexpensively using IT. For example, the Army could administer follow-up tests to IMI learners via Army Knowledge Online (AKO) after students return to their units or when they register for or attend the residential portion of a course after completing the DL phase.
- **Behavior.** IT could be used to collect and analyze data to assess the association of individuals' performance in DL with performance in subsequent residential training or with ratings of subsequent job performance in the unit.
- **Test evaluation.** IT could be used to administer course tests and conduct statistical analyses of objective test items (i.e., multiple choice or true/false) to provide information such as whether items are at the appropriate level of difficulty and whether the tests discriminate between good and poor performers in the course.
- **Administrative data.** Automated systems can capture data such as enrollment and dropout rates, DL usage, and information pertaining to course development, such as cycle time. These types of metrics can be indicators of course quality and should be monitored to assess progress in meeting other TADLP objectives.

Conclusion

This research indicates that a program-level evaluation of IMI courseware is needed and demonstrates a feasible method that can identify areas for improvement in courseware quality. It also suggests other ways in which the quality of IMI training could be practically evaluated at the program level.

In a larger context, evaluation of IMI quality would be part of a more comprehensive assessment component for TADLP. In addition to evaluating quality, that program would include an examination of learning models, courseware usage and efficiency, cycle time of production, and identification of DL areas with the highest payoff for the Army. We are conducting continuing research related to these subjects. Taken together, these efforts will give the Army a basis for managing continuous improvement in the development and use of IMI and will provide important tools to help meet the goals of TADLP.

Acknowledgments

This study benefited from the assistance of many people in the Army and affiliated organizations. We are particularly indebted to COL James Markley and his staff at the Distributed Learning Directorate; to the DL staff at the Army Training Support Center, including James Tripp and Joseph Delizzio; and to additional personnel from Army proponent schools who participated in interviews about DL evaluation and provided access to courseware. Special thanks also go to Wanda Majors and Nathan Ashlock at the Armor School, who made courseware available to us online; to Richard Clark from the University of Southern California and D. E. "Mac" MacAllister from TRADOC for sharing their training evaluation criteria and other resources pertaining to DL; to the American Society for Training & Development for providing access to their E-Learning Courseware Certification standards; and to Dave Nilsen from Alion Science and Technology for sharing his extensive knowledge about DL. Our work also benefited from the input of a number of RAND colleagues: We especially want to acknowledge Henry Leonard and Kristin Leuschner for their contributions and Lawrence Hanser and Christopher Paul for their helpful reviews of this report.

Acronyms

AKO	Army Knowledge Online
ANCOC	Advanced Noncommissioned Officer Course
ARFORGEN	Army Force Generation
ARI	Army Research Institute
ASI	Additional Skill Identifier
ASTD	American Society for Training & Development
ASVAB	Armed Services Vocational Aptitude Battery
ATLDS	Army Training and Leader Development Strategy
ATRRS	Army Training Requirements and Resources System
ATSC	Army Training Support Center
AUTOGEN	Automated Survey Generator
BNCOC	Basic Noncommissioned Officer Course
CAT	computer-adaptive testing
CCC	Captains Career Course
CCT	computerized classification testing
COA	course of action
DL	distributed learning
DLKN	Distributed Learning Knowledge Network
FM	field manual

FY	fiscal year
IMI	interactive multimedia instruction
IT	information technology
KSA	knowledge, skills, and abilities
LMS	Learning Management System
MOS	Military Occupational Specialty
NA	not applicable
NCO	noncommissioned officer
NCOES	Noncommissioned Officer Education System
PME	Professional Military Education
RL	residential learning
SME	subject matter expert
SQI	Special Qualification Identifier
TADLP	The Army Distributed Learning Program
TCP	TRADOC Campaign Plan
TRADOC	Training and Doctrine Command

CHAPTER ONE

Introduction

Background

Since 1998, the Army's Training and Doctrine Command (TRADOC) has been engaged in establishing and fielding The Army Distributed Learning Program (TADLP) to enhance and extend traditional methods of learning within the Army's training strategy. Through distributed learning (DL), the Army aims to achieve a number of goals, including increased readiness, improved access to training and reach-back capabilities, increased effectiveness in training and education, reduced costs, and greater stability for families. The Army is in the process of converting 525 training courses to DL by 2010 and has an expansive vision for a greatly increased role for DL over time.[1] TADLP aims to achieve DL objectives for Army training by developing courseware that uses various (primarily digital) media, implementing an effective learning management system, and maintaining modern, state-of-the-art training facilities (both fixed and deployable). In fiscal year (FY) 2007, TRADOC asked RAND to assess how efficiently and effectively

[1] For example, as early as 2001, the TADLP Campaign Plan listed as a critical success indicator that "DL satisfy between 30–65 percent of the quota-managed training load" (Section 1.9.4). More recently, the TRADOC Campaign Plan (TCP) called for the creation of an exportable Noncommissioned Officer Education System (NCOES) capability (Section 2.8). The current Army Training and Leader Development Strategy (ATLDS) is moving toward such concepts as expanded lifelong learning, greater support of Army Force Generation (ARFORGEN) processes, and the development of more adaptable leaders and creative thinkers. The ATLDS makes clear that DL will play a major role in the implementation of the strategy.

TADLP has accomplished its objectives and to recommend improvements in the Army DL program. This report documents one component of this project, an assessment of the quality of TADLP courseware.[2]

TADLP courses consist primarily of asynchronous interactive multimedia instruction (IMI). An assessment of IMI quality is necessary to determine progress toward meeting TADLP goals and to respond with appropriate strategic planning efforts, i.e., to manage budgets devoted to increasing quality and to identify and implement needed improvements to processes that affect quality. In addition, ensuring and documenting the quality of IMI courseware is important to show the value of this approach to instruction, to gain the buy-in of DL stakeholders, and to secure the resources needed to achieve the program's goals.

When the research reported here began, TADLP had developed 217 DL courses and products representing 23 proponent schools. Our analysis shows that usage of "high-priority" courses[3] (n = 166) has been lower than expected.[4] The quality of the courseware may account for some of the deficit. Surveys have suggested that some schools, commanders, and students believe that TADLP courseware does not have the same quality as traditional resident instruction and that it does not

[2] The DL project consisted of a comprehensive assessment of the state of TADLP. In addition to assessing the quality of courseware, the project examined the impact of TADLP courseware utilization on Army training, the cycle time and cost-effectiveness of the courseware production process, and the responsiveness of TADLP content to changing requirements. The purpose of these analyses was to establish a baseline against which future improvements to TADLP could be measured. In addition, the DL project proposed and developed options that the Army could implement to improve DL performance.

[3] Courses considered high priority are those that make the greatest contribution to unit readiness. These include Military Occupational Specialty (MOS) reclassification courses, as well as selected Professional Military Education (PME) courses supported by TADLP, including Basic Noncommissioned Officer Courses (BNCOCs), Advanced Noncommissioned Officer Courses (ANCOCs), Captains Career Courses (CCCs) for reserve component soldiers, and key functional courses designed as Additional Skill Identifiers (ASIs) or Special Qualification Identifiers (SQIs).

[4] For example, fewer than 100 high-priority DL courses had enrollments in FY2006, and the median enrollment was about 250 per course. As implied by these numbers, DL training accounted for only a small percentage of all structured training conducted by the Army, typically less than 6 percent of the training load even in categories where DL was most concentrated, such as the BNCOCs and ANCOCs.

train to the same standard. For example, in the December 2005 *Status of Forces Survey of Active Duty Members* (Defense Manpower Data Center, 2006), 46 percent of respondents reported that online training was less effective than traditional classroom training for individual learning. This perception was particularly pronounced among officers. The Fall 2006 *Sample Survey of Military Personnel* (U.S. Army Research Institute, 2007) indicates that attitudes toward Internet-based training have become more positive over the past few years, but attitudes toward online training remain substantially less positive than perceptions of classroom instruction.[5]

Being able to document success in Army IMI could go a long way toward changing those views, reenergizing the program, and positioning TADLP to compete more successfully for resources in the programming process. However, as discussed in more detail below, there is no systematic program-level assessment of course quality. Course evaluation is generally left to individual proponent schools, and although some schools have conducted course assessments, their efforts have tended to vary greatly in scope, purpose, and content, thereby precluding aggregation at the program level. We believe that a program-level approach to evaluation is needed to achieve TADLP goals.

Purpose and Organization of This Report

This report describes the development and application of a program-level approach to evaluating IMI courseware. We designed an approach

[5] Many researchers have debated the relative value of DL and residential learning (RL). Reviews of the literature typically have found no clear advantage for one or the other (e.g., Phipps and Merisotis, 1999). However, this cannot be interpreted to mean that these methods of instruction are equally effective. In addition to the problem of attempting to "confirm the null hypothesis," the evidence base is too limited to support such a conclusion, and many of the studies comparing DL and RL instruction are subject to methodological limitations (Straus et al., 2006). We have argued previously (Straus et al., 2006) that research focused on enhancing the quality of DL, rather than contrasting DL with RL, is more likely to yield results that are of practical value to the Army. This is both because the Army is moving to DL due to the benefits DL offers (e.g., anytime/anyplace learning) and because it is often impractical to conduct rigorous studies that compare results for the same course conducted in both RL and DL formats.

based on several existing methods and used it to evaluate lessons from a sample of IMI courses. The courses were selected from all high-priority DL, i.e., courses devoted to structured individual training required for promotion under the Army's professional military education (PME) program and reclassification training for qualification in a military occupation. The evaluation allowed us to (1) test our method and (2) show how it can be used to identify areas for improvement in IMI courseware and to monitor quality at the program level.

This report is organized as follows. In Chapter Two, we describe different approaches for evaluating training quality and discuss the extent to which these approaches are currently being applied by the schools participating in TADLP. In Chapter Three, we describe the approach to courseware evaluation we designed for this study. In Chapter Four, we present detailed findings from our evaluation. Finally, in Chapter Five, we discuss the conclusions and implications of our findings as well as potential directions for future program-level evaluations of DL quality. Our findings and conclusions concern both areas for improvement in TADLP courseware and directions for improvement in the evaluation of IMI within TADLP.

CHAPTER TWO

Quality-Evaluation Approaches and Their Usage in TADLP

In this chapter, we describe different approaches for evaluating training quality and discuss the extent to which they are currently being applied by the schools participating in TADLP.

Approaches to Evaluating Training Quality

There are a number of approaches to evaluating the quality of training, including evaluating training outcomes, assessing the quality of course tests, analyzing administrative data, and evaluating courseware content and design. A comprehensive evaluation of training quality will include several types of measures and methods.

Training Outcomes

Kirkpatrick (1959–1960, 1994) identified four levels of training outcomes: reactions, learning, behavior, and results.

Reactions. Learner reactions typically are assessed through post-course surveys of student satisfaction. This is the most common method of evaluating training quality. Reaction measures may not be related to learning or to performance on the job, but they can be an important factor in determining whether to continue offering a particular course (Goldstein and Ford, 2002) and in identifying areas for improvement in courses.

Learning. Learning refers to acquisition of knowledge and skills during training. It can be assessed by measures such as knowledge tests, skills tests, or peer ratings (Goldstein, 1991). Later assessments of learning (some time after the completion of training) can be used to assess knowledge retention.

Behavior. Behavior refers to whether students apply what they learned in training on the job. It can be measured through job performance or proxy measures of performance such as rates of promotion in an organization, or through performance in subsequent training (Goldstein, 1991). Ideally, job performance is measured by objective criteria such as production quality or accuracy, time to complete tasks, or use of raw materials. Frequently, however, job performance is assessed by subjective supervisory performance ratings.

Results. Results provide information about the effect of a training course or program on organizational outcomes. For example, unit readiness could be used as a measure of results. However, it is difficult to identify appropriate, concrete measures of organizational performance or impact and link them with training. We do not address this type of measure further, as it is beyond the scope of this study.

It is important to note that most of these outcome measures are not indicators of training quality in and of themselves. In order for outcomes such as changes in learning or behavior to be attributable to a training program, they must be assessed using an appropriate evaluation design. For example, the effect of a training program on learning can be assessed by comparing students' scores on pretests and posttests. Generally, a control group is also needed to determine whether the training program (versus other factors) accounts for differences in pretest and posttest scores (although some study designs allow such inferences to be made without a control group; see Goldstein and Ford, 2002). The effect of training on behavior can be assessed by comparing job performance of individuals who participate in training with a no-training control group. The predictive validity of training can be assessed by examining the association (correlation) between measures of learning in training (e.g., test scores) and ratings of subsequent relevant job performance or performance in later, more advanced training.

Test Evaluation

Although not a measure of training quality itself, good tests are an essential component of several methods of evaluating training. Content validity is one facet of test quality; it is defined as the extent to which a measure consists of a representative sample of tasks, behaviors, or knowledge drawn from a particular domain (see, e.g., Society for Industrial and Organizational Psychology, 1987). A test with high content validity would include questions representing the range of concepts and skills taught in a course. Content validity is typically assessed by subject matter experts (SMEs). Good tests also have appropriate levels of difficulty that discriminate between good and poor performers in a course. These characteristics of tests can be assessed using statistical analyses of students' responses to objective test items (i.e., multiple choice or true/false questions). Statistical analyses can also be used to examine patterns of responses to questions to identify possible instances of cheating.

Administrative Data

Administrative data obtained from archival sources provide information about the degree to which a training program is meeting objectives that are related to quality. For example, metrics showing few enrollments, low completion rates, or low graduation rates might suggest that there are problems with the quality of the course. Similarly, a long cycle time for courseware production might indicate that the course does not include the most relevant, up-to-date training content reflecting best practices and the needs of the operational Army.

Courseware Evaluation

The quality of training can also be assessed by evaluating the quality of courseware in terms of what is taught and how it is conveyed. SMEs typically evaluate what is taught by examining course materials and determining if the course content adequately and accurately represents the knowledge, skills, and abilities (KSAs) required for the job. Evaluation of how content is conveyed consists of assessing the quality of training delivery or the learning experience (ASTD Certification Institute, 2001–2003). This type of evaluation focuses on

characteristics of courseware such as ease of navigation and tracking, lesson structure, clarity of presentation, quality and value of audio-visual information, frequency and quality of examples, opportunities for practice, and frequency and quality of performance feedback. Whereas many of these criteria are important for all types of training, they are especially important in evaluating DL, because less is known about how to deliver technology-mediated instruction than about delivery of traditional in-person training. The courseware evaluation described in this report uses such criteria.

Current Evaluation Efforts in Army Training

As part of its overall assessment of DL for TRADOC, RAND conducted structured telephone interviews with representatives from TRADOC proponent schools to collect data on a variety of topics pertaining to 20 of the Army's DL programs with high-priority courses developed under TADLP. Participants typically included contracting representatives, course managers, team leads, training division or branch chiefs responsible for the production of DL, and, in some cases, the school's director of training. A complete description of this undertaking is beyond the scope of this report, but relevant here are responses to a question that addressed the schools' efforts to assess their DL programs:[1] Participants were asked to describe the methods their school uses or plans to use to assess the quality of IMI training.

Responses showed that, in general, evaluations of Army DL courses are not comprehensive or systematic in terms of either the range of training-evaluation measures collected or standardization of measures and synthesis of results across courses or schools. The nature of current efforts is summarized in Table 2.1 and described below.

[1] Other topics in the interview included the role of DL in the school's larger training strategy, how training content to be converted to DL was selected by the school, the amount of resources dedicated to the DL program at the school, and obstacles to and suggested improvements in the implementation of TADLP.

Table 2.1
Summary of Current Evaluation Efforts in Army Training

Type of Evaluation	Nature of Current Efforts
Reactions	Some local end-of-course attitude surveys are administered, but they are not systematic or standardized. Several broad-based efforts have also been made.
Learning	Many courses have posttests, but there are no reports of pretest/posttest comparisons across courses or schools and no known efforts to measure knowledge retention.
Behavior (performance)	AUTOGEN[a] attempts to capture a component, but most schools do not view it as useful for DL.
Courseware evaluation	Each school validates substantive content during development. For automated courseware, the validation process includes criteria about the quality of the learning experience. Validation efforts are not standardized or systematic.
Test evaluation	Validation is conducted at the end of development. A few efforts are made after fielding, but they typically do not use the most informative methods.
Administrative data	Some efforts are planned, but they are not comprehensive.

[a]AUTOGEN (Automated Survey Generator) is a survey and analysis platform licensed by the U.S. Army Research Institute (ARI). It is designed to perform job analysis and to develop and conduct training evaluation surveys. AUTOGEN is explained in more detail later in this section.

Measures

Reactions. Learner reactions are measured in some local end-of-course surveys, but questions are not standardized and efforts are not systematic. There are, however, some broad-based efforts to measure student satisfaction, such as the U.S. Army Research Institute's (ARI's) *Sample Survey of Military Personnel*, which is conducted every other fall and includes questions about satisfaction with Internet-based training. The Defense Manpower Data Center conducts numerous surveys, including *Status of Forces Survey of Active Duty Members*, which periodically has included questions about perceptions of the effectiveness of Internet-based training. These surveys are not specific to the Army, however.

Learning: Pre/Post Comparisons. Most IMI courses have posttests, but aggregate results from multiple courses (e.g., the percentage of students that pass DL courses) are not collected or reported at the program level. Moreover, there appear to be no systematic efforts to measure gains in learning by comparing performance on pretests and posttests, and we are not aware of any efforts to measure knowledge retention.

Behavior. Currently, no efforts are being made to assess the effect of training on job performance. Such evaluations are particularly difficult to perform. Studies that use a no-training control group are impractical to conduct on a routine basis in the Army. Assessing the predictive validity of training, i.e., the correlation or association of performance in training with performance on the job, requires large samples. Obtaining sufficient samples may be challenging in the Army, both because some soldiers do not perform the job for which they were trained (at least in Advanced Individual Training), and because it might be difficult to get unit leaders to provide performance-evaluation information in a timely way. Obtaining sufficient data should be feasible, however, if performance in DL courses is used to predict performance in subsequent, more advanced training.[2]

The AUTOGEN program collects unit leaders' perceptions of training effectiveness and therefore may have the potential to capture ratings of individual students' job performance for use in predictive validity studies. AUTOGEN is an automated system that enables proponent schools to develop their own computer-assisted surveys using standardized as well as customized questions. Evaluation questions are tied to specific training courses. AUTOGEN appears to provide the foundation for collecting data needed to analyze predictive validity;

[2] Assessing the association of performance in training with job performance or with subsequent training performance also is difficult due to "restriction of range" in training scores, job performance, or both. Restriction of range occurs when most students pass training courses, for example. Restriction of range in job performance also can occur if individuals with performance problems drop out of courses or are assigned to other duties, thereby eliminating lower scores from the range. Restriction of range can limit the observable correlation between performance in training and the outcome of interest (job performance or subsequent training performance).

however, it currently does not distinguish responses for DL and residential training. In addition, because the outcome in AUTOGEN is performance in the unit, it is useful only for standalone DL courses, not for DL that is used to prepare learners for the residence portion of a course or other types of blended learning.

Courseware Evaluation. Courseware evaluation occurs during courseware validation, a required step in the Army's training development process (TRADOC PAM 350-70-10). Courseware validation is a localized effort. It tends to emphasize accuracy and completeness of course content, although the process also is intended to assess the quality of the learning experience. However, it appears that validation efforts are not systematic or standardized in terms of evaluation criteria and processes, results are not quantifiable, and there are no efforts to aggregate findings beyond individual courses.

Test Evaluation. Test validation is conducted at the end of DL course development. We are aware of only one proponent school that evaluates the quality of its tests after courses are fielded by analyzing learners' responses to test items.

Administrative Data. We found no initiatives at the school level that use archival data to monitor quality, although informal assessments are likely to be conducted in some cases. The Army Training Support Center (ATSC) does maintain data regarding other characteristics of courseware (e.g., type of course, proponent school, hours of DL) funded under TADLP, some of which may be related to quality (e.g., cycle time to develop courseware). ATSC also collects data on development status (e.g., still in development, completed, fielded) of all courses funded under TADLP and is initiating a new information system[3] that will help personnel keep better track of completion of the phases of development. The Army Training Requirements and Resources System (ATRRS),[4] which is managed by the G-1 of

[3] The Distributed Learning Management Information System is a recently developed, automated management information tool designed to provide real-time DL development data to ATSC and TRADOC managers. Once fully implemented, this system should provide the capability for more detailed tracking of cycle times.

[4] ATRRS is the Department of the Army management information system of record for managing student inputs to training.

the Department of the Army, tracks the usage of all quota-managed courseware, including DL courseware. However, ATRRS records may not be as complete for DL phases of courses as for residential courses, because in some cases, completion of only the final, residential phase is entered into ATRRS. Furthermore, TRADOC does not currently use any ATRRS data in its TADLP management processes, partly because there are no unique course identifiers, making it difficult to match courses found in ATRRS with development efforts under TADLP.

Conclusion

In summary, TADLP does not have sufficient data or data integration capabilities to gauge overall program effectiveness. Recommendations for a comprehensive approach to evaluating program effectiveness are presented in Chapter Five of this report.

CHAPTER THREE

RAND's Approach to IMI Evaluation

In this chapter, we describe the approach used to conduct our evaluation of IMI quality.

Evaluations of DL courseware are recommended by TRADOC, DL industry groups, and others, yet no systematic assessment has been attempted within TADLP at the program level. Therefore, we evaluated a sample of IMI courses produced under TADLP. Like other approaches, our approach focuses on the quality of the learning experience (see, e.g., ASTD Certification Institute, 2001–2003) in terms of instructional design and technical features of courseware, rather than on the accuracy or comprehensiveness of substantive course content. The goals of this effort were to develop and test a method of evaluation for use on a broader scale in the future.

Courses Evaluated

We evaluated a sample of lessons from ten DL courses produced between 2005 and 2007. To select courses for evaluation, we started with a database of all DL products (n = 217). We focused on high-priority courses that are most directly connected to readiness (and therefore are also the longest and most demanding); these include MOS-producing courses and PME courses. Therefore, we eliminated unit training products (n = 3),[1] self-development courses (n = 33), and obsolete courses

[1] Training products are meant for training conducted by units, whereas courses are meant for training conducted by institutions.

or courses under maintenance (n = 52). We narrowed our sample further to recently completed (from 2005 to 2007) high-priority courses (n = 86) that were fielded (n = 74 out of 86) and had an active course number in ATRRS, meaning that students could register for them (n = 50 out of 74). These 50 courses were found in 14 proponent schools.

We selected a modified random sample of these courses, including more courses from larger proponent schools and courses representing a range of topics and levels, e.g., reclassification courses, Basic and Advanced Noncommissioned Officer courses (BNCOCs and ANCOCs), and Captains Career Courses (CCCs). In four cases, we could not get access to the course we selected, so we replaced these courses with a convenience sample of courses that were similar in topic and/or level and from the same proponent schools, where possible. The courses we evaluated are shown in Table 3.1.

We had online access to two of the courses, and the remaining eight courses were provided on CDs. Three evaluators, or "raters," went through the courseware in much the same way a student would

Table 3.1
Courses Included in RAND's Evaluation

Course	Proponent School	MOS	Lessons Coded/Total Lessons
Medical Logistics Specialist	Army Medical Department (AMEDD)	91J10	3/6
M1A2 Abrams Crew BNCOC	Armor	19K30	10/24
Cavalry Scout	Armor	19D10	15/32
Maneuver C3, Phase II	Armor	—	5/16
Chemical Operations Specialist	Chemical	74D10	6/21
General Construction Equipment Operator	Engineer	21J10	3/7
Bradley Fighting Vehicle System Maintainer	Ordnance	63M10	3/20
Food Service Specialist ANCOC	Quartermaster	92G40	10/17
Battle Staff NCO	Sergeants Major Academy	—	15/64
Watercraft Operator	Transportation	88K20	7/14

(although the raters also intentionally made errors, attempted to proceed out of order, and so forth, to learn how the courseware responded).

Typically, we assessed lessons from the first two or three modules of each course, where we define a module as a general topic area comprising one or more lessons.[2] We focused on the beginning modules of the courses rather than a random sample of modules, because in most courses, material in later modules builds on material in earlier ones. By going through the material in order, raters were better able to evaluate criteria such as whether the sequencing of lessons was logical and whether concepts and practice opportunities progressed from simple to complex and from part-task to whole-task. In four courses, however, we selected lessons from both beginning and later modules to ensure that we evaluated varied content, including instruction of concepts, processes, and procedures, if available.

We coded a total of 79 lessons, which comprised 39 percent of the lessons per course, on average, or approximately 200 hours of instruction.

Evaluation Criteria

We referred to three existing checklists to develop a set of evaluation criteria: (1) the American Society for Training & Development (ASTD) Certification Institute E-Learning Courseware Certification Standards (ASTD Certification Institute, 2001–2003); (2) a checklist from the Center for Cognitive Technology at the University of Southern California (provided by Richard Clark, February 5, 2007, in the context of his work on military training); and (3) a checklist developed by TRADOC (provided by D. E. MacAllister, Instructional Systems Specialist, May 16, 2007). Each of these checklists has advantages and disadvantages, as outlined in Table 3.2 and discussed below.

The ASTD standards reflect four elements of course design: interface, compatibility, production quality, and instructional design.

[2] Terms such as *module*, *lesson*, and *topic* were used differently across courses by the proponent schools.

Table 3.2
Sources Used to Develop Criteria for RAND's Evaluation

Source	Advantages	Disadvantages
ASTD E-Learning Courseware Certification Standards	Based on consensus among training and development experts. Comprehensive and detailed technical and production criteria. Weighted scores. Informative user manual.	Less emphasis on pedagogical criteria. Criteria fairly generic; some require revision for military training.
Center for Cognitive Technology, University of Southern California	Based on research. Comprehensive and detailed production quality and pedagogical criteria.	Lacks criteria for technical features.
TRADOC	Comprehensive and detailed technical, production quality, and pedagogical criteria.	Criteria somewhat too specific. Uses overlapping/redundant criteria.

Rating options include "yes," "no," and for some criteria, "not applicable." The ASTD's approach has several strengths: The standards are based on consensus among training-development experts, and criteria are weighted to reflect their relative importance or impact. The ASTD also has established interrater reliability, demonstrating that the standards can be applied consistently. In contrast to the other checklists, the ASTD also provides an informative user manual that gives clear explanations and examples of the criteria. However, standards for pedagogical features of the courseware are somewhat less detailed and comprehensive than those in the other checklists. The criteria are fairly generic in nature (by design) and therefore require revision for evaluating some aspects of military training courseware (e.g., by including references to the "mission environment" as the context; adding criteria for checks on learning; specifying criteria for learning objectives as action, standard, and condition; and so forth).

The checklist developed by the Center for Cognitive Technology reflects four aspects of course design: course and lesson introductions; instruction of concepts, processes, and procedures; practice, feedback, and assessment; and multimedia design. It also includes criteria

for evaluating contractors' qualifications. Rating options include "go" and "no go," along with space to provide comments for each criterion. Key advantages of the checklist include comprehensive and detailed pedagogical criteria that are linked to empirical research findings (see O'Neil, 2003). Most of the criteria in the checklist are relevant to all training, not just DL. The checklist is also appropriate for evaluating Army courseware. Unlike the other checklists, however, it lacks criteria for technical features of courseware, such as navigation functions.

The TRADOC checklist includes eight categories encompassing a wide range of elements.[3] Rating options include "go," "no go," and "not applicable." This checklist has a number of strengths. It is very comprehensive, encompassing a large number of detailed criteria. It is designed to apply to all courses, not just DL. It is tailored to military training, and unlike the other checklists, it includes criteria for course administration issues and strategies for learner self-development and study guidance. The primary disadvantage of this checklist is that applying it may be very resource-intensive given the number of criteria and their level of detail. It also has quite a bit of redundancy both across and within categories. For example, we identified eight criteria that concern maintaining learner interest by using varied instructional techniques.[4] In addition, clarification is needed regarding the basis for judging some of the criteria, e.g., "Colors are appropriate for their use."

According to ASTD's website, as of October 30, 2007, its certification standards and process had been used to assess 212 courses for 31

---

[3] General categories are evaluation of the course introduction; presentation (which includes learning objectives, use of audiovisual materials, checks on learning, feedback, and other aspects of pedagogy); learner study guidance; performance measurement/tests; remediation; course management; technical matters; and instruction of concepts, processes, and procedures.

[4] The eight criteria are "Includes varied methods of instruction"; "Involves learners in activities by using visual . . . auditory . . . and physical senses"; "Maintains learner attention (e.g., uses humor, novelty, 3-D graphics, music, storytelling, etc.)"; "Includes various types of interactivity to maintain learner interest and promote learning"; "Uses scenarios to stimulate thinking and discussion"; "Triggers concrete imagery through stories, examples, analogies"; "Uses graphics, pictures, animation, or video when concrete examples are needed rather than relying solely on printed text or audio"; and "Excludes the talking head approach to presenting material."

companies.[5] Clark and MacAllister reported that at the time RAND began its evaluation, their respective checklists had not been used in any systematic assessments of courseware.

The RAND team used an iterative process to develop criteria for this evaluation, drawing from the three sources described above. We selected or modified criteria from the checklists, applied them to a subset of lessons, revised the criteria as needed, and so on. The rationale for eliminating or revising topics and criteria was to yield a set of standards that

- **Focused on what we viewed as the most important issues.** We emphasized pedagogical aspects of the courses, although we also assessed other features of the courseware, particularly for topics that were common to two or more checklists. We also eliminated some criteria that were redundant with others or seemed less central to evaluating courseware quality.
- **Struck a balance between level of detail and efficiency.** We attempted to select or write criteria that were comprehensive but not overly time-consuming to use. For example, the ASTD checklist includes seven criteria regarding navigation, with separate items to evaluate functions such as "start," "forward," "backward," "save," and so forth. TRADOC's checklist includes 11 criteria to evaluate navigation. We combined the ideas of both lists into one criterion, "A learner with modest computer skills could easily navigate through the course." If the rating was "no," the rater commented on the specific navigational features that were problematic. For example, the rater could note whether navigation features or buttons were missing, were too small, did not function predictably, and/or were not labeled.
- **Could be evaluated from available materials.** Because we had CD versions of most courses, we did not have access to course tests, nor could we evaluate features that are specific to online

[5] A more recent review of ASTD's website (on January 9, 2009) showed that the ASTD Certification Institute retired its E-Learning Courseware Certification program as of December 31, 2008.

courses, such as registration, bookmarks, and "live" technical support.

The resulting topics that we evaluated are shown in Table 3.3. They comprise four general categories reflecting technical, production quality (ASTD Certification Institute, 2001–2003), pedagogical features of the courses (with an emphasis on generic pedagogical criteria), and IMI levels. The complete checklist, which includes criteria for each topic, is shown in Appendix A. Although we group criteria into these categories, there is some overlap between them; for example, production-quality criteria (such as the use of animation to demonstrate procedures) influence the quality of pedagogical criteria (such as providing clear demonstrations of processes and procedures).

Table 3.3
RAND's Evaluation Topics

Technical
Launching
Navigation functions
Supplementary instructional resources
Technical support
Production quality
Legibility of graphics and text
Audiovisual material
Pedagogical
Lesson learning objectives
Course sequencing, pacing, and learner control
Feedback
Instruction of concepts
Instruction of processes
Instruction of procedures
Checks on learning
Practical exercises
IMI levels

NOTE: Criteria in italics were coded for each lesson; criteria in roman font were coded across lessons.

Technical criteria concern characteristics of the computer interface and related support. Examples include the ease of launching a course and navigating through it, the accessibility and usability of supplementary instructional materials (e.g., glossaries, field manuals (FMs), and other reference material), and the availability and quality of technical support.

Production-quality criteria concern the physical presentation of material. Examples include legibility of graphics and text and characteristics of audiovisual material, such as the relevance of the audiovisual content, the absence of sensory conflicts, and the use of techniques to maintain learner interest.

Pedagogical criteria concern the quality of the instructional content and processes. General topics include the following:[6]

- specificity of learning objectives
- degree to which the courseware prevented learner control of the course sequence, pacing, and activities, as well as the effects of such control on efficiency
- frequency and quality of feedback
- characteristics of instruction of concepts, processes, and procedures
- number and quality of checks on learning
- number and quality of opportunities for practice.

In addition to using these criteria, we assessed the level of IMI in each lesson. Contractually, all of the training content was required to have Level 3 interactivity (explained below), although some of the developed content typically has a lower level of interactivity.[7] We used the ATSC definitions of interactivity levels (Wardell, 2006). In brief, in Level 1 lessons, the learner is a passive recipient of information; i.e., the lessons are "page-turners" with no learner interaction. Level 2 lessons are those

[6] Specific criteria for each topic are shown in Appendix A and discussed in Chapter Four.

[7] A more recent contract calls for the school and contractors to specify the amount of content at each level of interactivity. The new contract allows for the possibility that levels lower than Level 3 are appropriate for some content.

in which learners have more control and interaction; for instance, they are asked to click on icons to reveal information or activities, move objects on the screen, fill in forms, and complete checks on learning and practical exercises. Level 3 lessons include more involved participation, such as the use of scenarios for testing, the need for the learner to make decisions, and complex branching based on the learner's responses.[8] We rated each lesson in terms of the highest level of IMI. Therefore, lessons with ratings of Level 2 IMI also included some Level 1, and lessons with Level 3 ratings included some Level 1 and Level 2.

As shown in Table 3.3, some criteria (depicted in italics) were rated for each lesson. The other criteria were rated once per course, because either there was only one "event" to rate (e.g., launching) or there was substantial consistency in these features across lessons (e.g., placement and functioning of navigational buttons, and the type of feedback provided following checks on learning and practical exercises). We did not rate the introductory lessons on the pedagogical criteria, because most did not apply (e.g., course introductions typically do not have checks on learning or practical exercises). Thus, all lessons were evaluated for IMI level, and 74 lessons were rated on the remaining criteria when applicable.

We determined whether each of the specific criteria shown in Appendix A was present ("yes" or "no"). We also identified the need for additional options for some criteria, including "sometimes," "not applicable," and "don't know." "Don't know" was used for only one question, which pertained to whether students had the option of testing out of course materials (which could not be assessed from CD versions of the courses). The "sometimes" option was used for only 5 percent of the ratings of individual lessons for which it was an option. "Not applicable" (NA) was used for 46 percent of the individual lesson criteria where it was an option. Typically, if a particular course feature did not exist (e.g., instruction on processes), the item was rated "no" and subsequent criteria regarding the same item (e.g., quality of instruction on

[8] None of the lessons sampled used Level 4 IMI, which consists of high-fidelity, immersive, simulation-based training.

processes) were rated "NA." In addition to providing a rating for each criterion, the rationale for negative ratings was documented in qualitative comments. Raters also provided comments about positive features of the courses.

Coding Process

The evaluation was conducted by three researchers: two master's-degree-level research assistants and the lead author of this report, who has a Ph.D. in industrial/organizational psychology. One research assistant evaluated seven of the courses, another evaluated three courses, and the lead author evaluated one course. The research assistants were trained by the author on lessons from three of the courses until satisfactory interrater reliability was achieved. For research results to be valid, they must be replicable, and interrater reliability demonstrates the degree to which different judges evaluate content the same way. It is particularly important to establish reliability when conducting qualitative research, because judgments are subjective.

Interrater reliability was assessed using Cohen's kappa statistic. Kappa can range from 0 to 1.00; values above 0.60 are considered "substantial" (Landis and Koch, 1977). After training, values of kappa on three lessons ranged from 0.74 to 0.90 among all three rater pairs. The two primary raters began evaluating courses independently (i.e., each course was rated by one rater). After evaluating lessons from two to three courses, they rechecked interrater reliability on three lessons. Kappa dropped substantially, to from 0.23 to 0.34. The disagreements that occurred were largely due to differences in the use of "sometimes" versus "yes" or "no." Previously coded lessons were recoded, and after a brief retraining, values of kappa on two additional lessons were satisfactory at 0.69 and 0.90.

CHAPTER FOUR

IMI Evaluation Findings

In this chapter, we present the findings from our evaluation of Army IMI. We begin with some general observations and then present specific findings about technical criteria, production-quality criteria, and pedagogical criteria. We also include recommendations for addressing deficiencies in courseware quality.

Interactivity and Course Length

The majority of the lessons we evaluated (75 out of 79) included instruction of concepts; 52 lessons included procedures; only three lessons included instruction concerning processes. More than half of the lessons (48) included more than one type of instruction; the vast majority (45 lessons) included concepts and procedures.

We found that 4 percent of the lessons were Level 1 IMI (most of these were introductory lessons), 76 percent were Level 2, and 20 percent were Level 3. Thus, the preponderance of the lessons we evaluated consisted of Level 2 IMI. Instruction for some of the basic principles in these courses may not lend itself to Level 3 IMI. However, we found that instruction for the majority of higher-level content consisted largely of Level 2 IMI, even though the delivery orders for this courseware specified Level 3 as the standard. Moreover, the lack of Level 3 interactivity may account for shortcomings in pedagogical features of the courseware, discussed later in this chapter.

Although time to completion was not a focus of our evaluation, we noticed that in some instances, it took much less time to complete

lessons than was stated in the lesson introduction. For example, the introduction to a lesson on basic tactical concepts estimated that the lesson would take 2.5 hours to complete, but the rater completed it in a little over 1 hour. An eight-page (i.e., eight-screen) lesson about computing back azimuths was estimated to take 2.5 hours to complete, but it took the rater approximately 10 minutes. Even allowing for variations in cognitive ability or academic experience, we expect that most learners would require far less than 2.5 hours to complete this lesson.

Remaining results are presented in Tables 4.1 through 4.7 below. For each criterion, ratings of "yes" were scored as 1.0, and ratings of "sometimes" were scored as 0.5. Each criterion was scored by summing the positive (i.e., "yes" and "sometimes") ratings and dividing by the number of total ratings across all lessons or courses. In the tables, criteria with 85 percent or greater positive ratings have a white background; criteria with 70 to 84 percent positive ratings have a gray background; and those with less than 70 percent positive ratings have a black background. These ratings correspond to the need for improvement in features of the courseware. As in Table 3.3, criteria in italics were coded for each lesson; criteria in roman font were coded across lessons.

Technical Criteria for Courseware

Table 4.1 shows ratings for the technical aspects of the courseware.

Table 4.1
Technical Criteria for Courseware

Criterion	Rating
Launching[a]	0.38
Navigation	
Ease of navigation	1.00
Cueing to position in course	0.90
Cueing to position in lesson	0.70
Online technical support	0.80
Supplementary instructional resources	
Easily accessible	0.70
Usable	0.30

[a]Evaluated only for CD versions of courses.

The technical features of the courseware were generally strong, with scores for most criteria ranging from 0.70 to 1.0. Examples of positive and negative comments about technical features include:

- "Very helpful instructions for adjusting computer settings on front page."
- "There is an introduction module that explains all of these things; very clear, very helpful."
- "Learner does not know where he/she is in lesson, which is rather frustrating."

We identified two technical features of courseware needing substantial improvement. One was the ability to launch courseware without assistance. We had no difficulty "launching" online courses, but we required assistance from RAND technical support to launch five of the eight courses provided on CDs. Technical support personnel assisted us with changing our computer settings, and even then, some courses and lessons did not launch smoothly, requiring some users to disable ActiveX controls for each lesson or for different segments within a lesson. These problems may have been due to the particular format in which we received copies of the courseware. However, if Army learners experience similar problems without access to technical support, it is reasonable to expect that many of them will give up, which would be reflected in low completion rates. Surveys of students' reactions could be used to determine whether technical features of courseware, such as difficulty in launching, are related to other outcomes such as dropout rates.

The second technical issue was the usability of supplementary instructional resources (i.e., reference materials), which was rated 0.30. Most courses had a menu through which the learner could easily access PDF files of FMs. However, the concepts on any particular page of instruction were not linked directly to the relevant FMs. Therefore, the learner had to go to the list of FMs, open the PDF (assuming he or she could determine which one was appropriate), and then search for the relevant material. Raters' comments about these types of instructional resources include:

- "Reference section contains manuals but does not point to specific content. Difficult to navigate."
- "Relevant FMs are listed; no links to connect to relevant content within instructional segments."

Some courses also had glossaries. The glossaries were generally accessible, but they were not always as informative as they could be. Comments about the glossaries include:

- "Glossary does not seem very extensive or comprehensive."
- "Glossary does not include definitions." [This glossary explained acronyms used in the course but did not define them.]

We recommend exploring the possibility of providing links throughout the courseware that take the learner directly to relevant sections of appropriate FMs and glossaries.

Production-Quality Criteria for Courseware

Table 4.2 presents the ratings for production-quality criteria, including legibility of text and graphics and use of audiovisual media. In general, the ratings indicate that these features have some strong components but need some improvement.

Table 4.2
Production-Quality Criteria for Courseware

Criterion	Rating
Legibility of text and graphics	0.80
Audiovisuals	
Narration easy to understand	1.00
Minimal irrelevant content	0.85
Use of animation/video to demonstrate processes	0.75
Techniques to maintain learner interest	0.50
Few sensory conflicts	0.40

The ratings for legibility of text and graphics were generally favorable, with an overall score of 0.80, but comments suggest some need for improvement:

- "Discrepancies between table of contents lesson titles and titles on screen during lesson."
- "Overall rating is good but there are a few grammatical errors and a few cases where the graphic does not match the concept being presented."
- "Some typos, grammatical errors; some graphics are illegible."

The use of audio and visual media showed high variability across the five features we assessed, with scores ranging from 0.4 to 1.0.

Narration and minimal irrelevant content received positive ratings. Narration was not available in all courses, but where it was an option, it was easy to understand. The courseware generally had very little irrelevant content.

Courses received a moderate rating (0.75) for the use of animation/video to demonstrate processes and procedures. Some lessons contained engaging and informative animation and graphics, whereas others were less effective. Comments regarding the use of animation and video include:

- "Most of material is relatively straightforward and animation may not be necessary."
- "Excellent graphics explaining the hull and its components."
- "Nice examples of audiovisuals for determining grid coordinates and measuring distance on a map that could be played to demonstrate procedure. Use of movable piece of paper to measure distance on a map and movable protractor to determine azimuths are nice features."
- "Some video clips are unclear; some are not necessary to demonstrate the concept."

There was a great deal of variability in ratings of the use of techniques to engage the learner. Raters identified both positive and nega-

tive examples, as noted in the following comments. The overall score of 0.50, however, indicates the need for significant improvement in engaging learners and maintaining their interest.

- "Cartoons/pictures may facilitate user interest."
- "Comic book style thought bubbles and cartoons add an element of humor/entertainment."
- "This course has an appealing style, the graphics are well done and pleasant to look at, but there are not many special techniques used to maintain interest."
- "A better effort to maintain user interest could be made. More graphics, interesting scenarios, etc. . . . there are some screens that require interaction, but not many."
- "Only pictures and text . . . no techniques to maintain user interest, but frequent checks on learning do provide variety and breaks between the text and pictures . . . this is not a very engaging course."

Finally, the score of 0.40 indicates that substantial improvement is needed to eliminate sensory conflicts, which included the presentation of text and audio of different content simultaneously or presentation of text and audio of the same content without the ability to disable one of the sources.

Pedagogical Criteria for Courseware

The pedagogical aspects of the courseware, reported in Tables 4.3 through 4.7, are the most important criteria in our quality assessment. This was also the area in which we found the most reasons for concern. Pedagogical features of courses include lesson objectives and sequencing, instruction of concepts and procedures, checks on learning, practical exercises, and feedback.

As shown in Table 4.3, ratings for lesson objectives and course sequencing were generally favorable, with scores ranging from 0.70 to 1.0. The vast majority of lessons included objectives that clearly stated

Table 4.3
Pedagogical Criteria for Courseware: Lesson Objectives and Sequencing

Criterion	Rating
Lesson objectives	
Clearly describe knowledge or skills	0.90
Have an observable, measurable performance standard	0.70
Course sequencing, order of lessons	1.00
Efficiency of restricting learner control	0.42

the knowledge and skills to be learned and typically did so in observable and measurable terms (e.g., by stating the action, condition, and standard). Improvement is needed in objectives for some lessons, however, either because formal objectives were not provided or because objectives were rather lengthy and/or not uniform across lessons.

Although it is not reflected in our rating system, we question the value of objectives that specify the condition as performance in the DL environment. Given that the purpose of DL training is to foster learning in residential training or performance on the job (or both), the expectation of mastery only within the context of the DL course warrants further scrutiny. The issue of job context also comes up in the ratings of instruction of concepts and procedures and in the quality of checks on learning and practical exercises, discussed later in this chapter.

Ratings of the sequence of lessons were uniformly favorable—lessons were presented either in the order in which job- or mission-relevant tasks are accomplished or in order of task difficulty.

We also examined the degree to which learners had control over the sequence and pace of the lessons. According to Clark, substantial research shows that as learner control over various aspects of instruction increases, learning decreases, except for "the most advanced expert learners" (Clark, 2003, p. 14). Clark recommends that DL courseware direct sequencing, contingencies, and learning strategies for all but expert learners and permit only minimal learner control over pacing.

In the CD versions of the courses, there was virtually no courseware control over the order of lessons; learners could go through the lessons in any sequence and they could skip most course activities.

In the majority of the courses we examined, it was not clear whether "expert" learners could be identified and given the chance to test out of lessons. However, the lack of control may be due to the format (CD) in which we received the courses. The two courses for which we had online access controlled the sequence of lessons and required learners to complete checks on learning and practical exercises. Therefore, we did not consider control over sequence in our evaluation.

However, the courseware controlled the pace in 60 percent of the lessons we reviewed. Learner control over pacing was largely dependent on whether the courses had narration. If there was no narration, the learner controlled the pacing. Similarly, if narration could be turned off and the corresponding narrative text could be turned on, the learner controlled pacing without losing the content. However, in many of the courses, the corresponding text of the narrative was not available. As a result, the learner could disable the audio and/or click "next" to proceed to the next screen, but that meant that he or she did not hear or read all of the content.

Although narration in the courses was easy to understand, as noted previously, we found the pace of the courses frustrating because many of the narrators spoke very slowly, which inhibited the pace of learning. This resulted in a score of 0.42 for efficiency of learner control. The efficiency of instruction needs to be weighed against the potential for improved learning achieved by the combination of animation and narration, which produces better retention and transfer than does animation plus on-screen text (Mayer and Moreno, 1998; see also Mayer, 2003). However, some of the course content was presented without animation, i.e., with narration and still photographs or diagrams, or narration alone. For this content, it would be more efficient, and unlikely to hinder learning, to offer the option of disabling the audio and enabling the corresponding text.

Table 4.4 shows ratings for instruction of concepts. The score of 0.97 indicates that the topics covered in each lesson appear to reflect the knowledge addressed in the learning objectives. In addition, instructional content, with a rating of 0.86, generally provided clear definitions of concepts.

Table 4.4
Pedagogical Criteria for Courseware: Instruction of Concepts

Criterion	Rating
Instruction reflects learning objectives	0.97
Clear, comprehensive definitions	0.86
Exercises identify examples and non-examples	0.54
Informative, sufficient examples from job/mission environment	0.53

However, clearer, more comprehensive definitions were needed in some courses, as noted in the comments below:

- "Some definitions are unclear."
- "Fifteen different concepts are listed; some definitions are very brief and contain little context."
- "Breakdown of Army structure is unclear."
- "Diagram doesn't match concept; diagram unclear."
- "Acronyms not defined; staff role statements are vague; the same words or phrases are used to describe multiple concepts—distinction is unclear."

Another criterion used in our evaluation was the presence of both examples and non-examples of concepts in practice opportunities. Examples are critical in training, and providing novel and varied examples promotes transfer of training to the job (for a review, see Clark, 2003). Non-examples are also important, as they help learners discriminate among situations in which the concepts being taught do or do not apply.

Virtually all the lessons we evaluated included examples, but considerable improvement is needed in the presentation of non-examples (which was rated 0.54). We chose a relatively low standard for the use of non-examples; for instance, checks on learning or practical exercises that used multiple choice or matching questions were considered sufficient, because the learner had to choose a correct answer (an example) from options that included incorrect answers (non-examples). A higher standard (which we did not use but which should be considered in future evaluations) would be to require that exercises explicitly ask questions about non-examples (e.g., a question that asks learners to

identify which among five items does not belong within a group or which step illustrating a procedure is out of sequence). The low rating on this criterion in the courses we evaluated came from two sources. In some lessons, the types of checks on learning or exercises did not include non-examples. In others, the lack of non-examples was due to a general failure to provide any checks on learning and/or practical exercises.

The final criterion regarding instruction of concepts is the provision of relevant context for the course material. Some lessons provided mission-relevant examples of concepts, including examples related to current or very recent military conflicts. For example, each lesson in the Maneuver CCC starts with a motivator that puts the skills to be learned into the context of a historical battle situation. However, a great deal of content in many courses was taught without any context in which the knowledge and skills are used, as indicated by the score of 0.53.

Representative comments from raters include:

- "More examples might be helpful for understanding when this form is used."
- "Example locations of unified forces are unclear. Few other examples are provided."

Specific examples that illustrate these comments include the following:

- A lesson on applying the course-of-action (COA) development process describes the elements of combat power and identifies six steps used to analyze relative combat power and develop an appropriate COA. The instruction includes an engaging narrated guided tour in which different officers talk about their roles; however, the commentary does not provide a specific scenario or mission context.
- A lesson on how to use international code flags to send and receive messages presents flags with their meanings and explains where to look up the meaning of flag combinations in the International

Code Flags document. The lesson includes some checks on learning in which the learner is asked to recall the meaning of a particular flag or flag combination, and there is a practical exercise that requires the learner to match flags with their meaning. However, references to the mission environment are almost completely absent from instruction in this lesson and are referenced only minimally in the practical exercises.

As with the instruction of concepts, the procedures taught in each lesson appeared to be consistent with those stated in the learning objectives (rating = 0.91). However, ratings on other aspects of instruction of procedures were low, with three of five criteria receiving scores ranging from 0.58 to 0.62 (see Table 4.5).

Raters felt that the courseware lacked clear, step-by-step demonstrations. According to Clark (2005), demonstration of procedures is one of the most critical parts of training design. Relevant comments include:

- "No overview of procedure; order is not clear."
- "Decision matrix is presented, but step-by-step demonstration is not provided."
- "Information is not presented in clear 'steps' . . . more of a narration . . . a step-by-step outline might be helpful."

Table 4.5
Pedagogical Criteria for Courseware: Instruction of Procedures

Criterion	Rating
Instruction reflects learning objectives	0.91
Where decisions are taught, includes alternatives and criteria[a]	0.72
Explains why procedure works in cases where learner must modify procedure[a]	0.62
Provides clear step-by-step demonstrations	0.61
Presents informative, sufficient examples from job/mission environment	0.58

[a]Small number of courses.

We note that there were some effective explanations of procedures that did provide clear step-by-step demonstrations. For instance, instruction on the steps for eye splicing, i.e., forming a permanent loop at the end of a rope by weaving the end into itself, provided clear descriptions and illustrations of each step and gave details about how far to measure, how many strands to count, where to mark each section, and how to remove bunching that may occur in the over layer. As noted earlier, the demonstration for using a strip of paper to measure distance on a map was also quite effective.

Instruction of procedures also needs improvement in that many of the lessons lacked sufficient mission-relevant examples. Comments include:

- "More examples illustrating use of the matrix would be helpful."
- "Explains how to do calculations but does not put into job/mission context."

Some specific examples illustrate these comments:

- Lessons on map reading lacked information about the circumstances in which a soldier would be expected to perform certain actions—for example, the situations in which a soldier would determine a magnetic azimuth using a compass or calculate back azimuths, or examples of the consequences of providing or failing to provide this information accurately.
- A lesson on how to solve mathematical problems involving degrees and time provided little context for these calculations. There was a brief reference to a situation in which the learner would need to add five degrees when calculating variance from a map, but no other significant scenarios were provided. In addition, the lesson did not communicate the source of the degrees being used in the calculations—for example, whether they would be provided in instructions from another person or whether the soldier would be measuring them on a map or taking readings from an instrument.

- In this same lesson, the courseware provided some context for calculating time, as shown in the following check on learning: "Your vessel is scheduled to arrive at its destination port at 0400. You estimate it will take seven hours to travel to the destination port. When must you depart to arrive at 0400 hours?" This offers some mission-related context, but it does not provide an engaging, mission-relevant scenario.

Finally, few courses provided instructions on how to make decisions in alternative scenarios or how to modify procedures in novel situations. In lessons in which learners were taught to make decisions, the rating of 0.72 indicates the need for some improvement in presenting alternatives and criteria for selecting among them. In lessons that explained how to conduct procedures, the rating of 0.62 indicates that the lessons did not explain why the procedures work the way they do or when to use alternatives. For example, one lesson discussed four methods of fratricide prevention but did not provide examples describing when each technique would be appropriate. Similarly, a lesson on direct fire planning defined two techniques for distributing fires but did not provide guidance regarding the conditions in which to use each method.

Table 4.6 presents ratings for checks on learning and practice. Checks on learning generally consist of brief questions staggered throughout a lesson to reinforce the material and enable the learner to verify that he or she understands the concepts being taught, whereas practical exercises occur at the end of a lesson and require greater synthesis and application of the material. The ratings for these criteria cut across all types of instruction, including concepts, processes, and procedures.

A number of lessons in two courses had no checks on learning. Where checks on learning were provided, the raters noted the need for some improvement in reinforcing key points from the lesson (rating = 0.70), as illustrated by their comments:

- "Checks on learning emphasize non-essential facts."
- "Focus on numbers/statistics rather than salient points from instruction."

Table 4.6
Pedagogical Criteria for Courseware: Checks on Learning, Practice

Criterion	Rating
Checks on learning	
Reinforce material	0.70
Sufficient in number	0.61
Practice	
Clear directions	0.95
Consistent with learning objectives	0.92
Adequate number of exercises	0.53
Requires learner to practice procedures	0.49
Whole-task practice reflects mission environment	0.48
Moves from simple to complex problems	0.31
Part-task practices followed by whole-task	0.24

Raters generally felt that there were insufficient numbers of checks on learning (rating = 0.61) (with the exception of one course, which the rater felt had too many checks on learning). Comments include:

- "More frequent checks on learning would help reinforce material learned."
- "Checks on learning would be more effective if there were more and if they were staggered throughout lesson rather than only at the end."
- "Only one concept is tested."
- "Considering [the] number of topics presented, checks on learning could be more frequent."

According to Clark (2005), opportunities for practice are another critical element of training design. We found that practical exercises tended to reflect the topics addressed in the lesson; hence, consistency with learning objectives was rated favorably (rating = 0.92). The directions also were very clear (rating = 0.95), but this was probably because many of the exercises were quite simple (e.g., consisted of multiple-choice knowledge questions).

Other aspects of practice opportunities need substantial improvement—the scores in the remaining features under "practice" range from 0.24 to 0.53. We found that learners were rarely required to integrate concepts or practice procedures. Fourteen lessons had no practice problems. Some of the lessons that did include practice had only one problem. Others used problems that were labeled "Practical Exercise" but consisted of multiple-choice items testing knowledge of the procedure rather than hands-on or simulated performance. For example, in the lesson on eye splicing, learners were presented with the ten major steps of the task and asked to put them in order. Another question presented a step in the process and asked which strands the learner would mark as part of the next step. However, learners were not required to recall all of the steps on their own or to practice the physical task. In addition, where practical exercises were included, they tended to test part-task practice only. Consequently, raters reported that many of the exercises failed to reflect the mission environment. Similarly, raters found that many practical exercises did not move from simple to complex problems.[1]

If questions on course tests are similar to those used in practical exercises, the value of the tests for measuring learner proficiency is likely to be at issue. Future courseware evaluations should include a review of course tests.

Most courses provided some form of feedback, but we found that improvements are needed in some key areas, as shown in Table 4.7. Feedback associated with checks on learning and practical exercises typically allowed the learner to correct his or her strategy, i.e., gave

[1] Some of the calculations for these ratings eliminate lessons from the denominator when criteria were not applicable. Lessons without any practical exercises were eliminated from the calculations for items about part- and whole-task practice, the progression from simple to complex problems, consistency with learning objective, and clear directions. Scores for the criterion "moves from simple to complex problems" eliminated lessons for which the rater judged the complexity of the material as constant across the lesson (i.e., the criterion was rated "NA"). The score for "whole-task practice reflects mission environment" eliminated lessons for which there was no whole-task practice.

Table 4.7
Pedagogical Criteria for Courseware: Feedback

Criterion	Rating
Learner can correct his/her strategy	0.90
Feedback is frequent enough that errors do not accumulate	0.70
Provides opportunity to review relevant material	0.65
Provides "why" explanations	0.30

the learner more than one try to answer the questions (rating = 0.90). Some improvement is needed in the frequency of feedback to prevent errors from accumulating (rating = 0.70). However, greater improvement is needed in giving students an opportunity to review material if they answer questions incorrectly (rating = 0.65). Some courses provided a link following an incorrect answer that took the learner back to the relevant material in the lesson, whereas others did not direct the learner to the relevant information, as noted in these comments:

- "User can go back in the lesson, but there are no specific directions or links to the relevant information."
- "Learner can navigate back; feedback screen does not provide link to relevant section of the instruction for review."

Finally, a score of 0.30 indicates the need for substantial improvement in providing more comprehensive feedback in response to incorrect answers. For the most part, when the learner answered a question incorrectly, the course presented only the correct answer and did not explain why the learner's response was incorrect. In addition, the correct answer often merely repeated the text in the instructional materials. For example, a check on learning about the Bradley fighting vehicle included the following multiple-choice question:

> The track and suspension system has six pairs of wheels that support the hull, and support rollers that _____.

The rater selected an incorrect response, "Reduce bounce over rough ground." After a second attempt, the rater was told, "Incorrect. The support rollers keep the track away from the wheels." This feedback is identical to the narration in the original lesson screen, i.e., "The

track and suspension system has six pairs of wheels that support the hull, while the support rollers keep the track away from the wheels." More instructive feedback would also identify the parts of the truck or suspension system that reduce bounce over rough ground.

An example of more comprehensive feedback was provided in response to a multiple-choice question about how to clear a building. The question asked, "Which of the following are the principles of precision room clearing?" The rater selected an incorrect answer, "Surprise, diversion, and uncontrolled violence," and received the following feedback: "A diversion can assist you in surprising the enemy and you must maintain control at all times. The principles of precision room clearing include moving quickly, and having the mind-set of complete domination." Here, the feedback is more informative in that it uses different words to describe the two items that are missing in the learner's answer choice.

In another course, the checks on learning asked the learner to type responses to questions about doctrinal policy in a free-text format. For example: "Explain in your own words those actions a commander must plan and implement in order to restore units to a desired level of combat effectiveness." After the learner submits the response, the courseware displays the correct answer, and the learner is instructed to compare his or her answer to it. In some cases, the correct response was identical to the text provided in the lesson. In addition, this feedback is not informative because the learner may not be able to judge the extent to which his or her answer captured key elements of the correct response, particularly if the answer is complex. The learner also does not receive feedback about why incorrect responses are wrong.

For this example, the question could be posed in other ways that would provide more feedback about the learner's response. The correct response should include five steps in reconstitution operations: remove the unit from combat; assess it with external assets; reestablish the chain of command; train the unit for future operations; and reestablish unit cohesion. An alternative question format would be to present a list of, say, ten steps and ask the learner to select the five correct actions and put them in order. This format would give learners feedback from the presence of non-examples (the five incorrect options)

and the sequence of operations. Providing specific feedback regarding why incorrect responses are wrong would be even more helpful.

CHAPTER FIVE

Conclusions and Implications for TADLP's Assessment of IMI

In this chapter, we first discuss how our findings can be used to identify improvements for IMI courseware. Next, we describe potential directions for improvement of program-level evaluations of DL quality in the future. Finally, we summarize the conclusions from our research and what they imply for TADLP.

How IMI Courseware Can Be Improved

Recommendations for Courses in Our Sample

Overall, we found that some features of the courses we evaluated were generally strong, whereas other areas, especially those involving pedagogy, require improvement. Below, we summarize our findings and show how these results can be used to identify needed improvements in IMI courseware. Although the focus of our evaluation was quality at the program level, results can be reported at the course level to identify specific needs for improvement.[1]

Our analysis revealed that technical characteristics were the strongest features of the courseware. All the courses we evaluated were easy to navigate, and cues to the learner's position in the course were easily accessible. Improved cueing to one's position in the lesson is desirable

[1] TADLP could provide proponent schools with the completed checklists, which report results at the criterion level, e.g., "yes" and "no" ratings for each item in the checklist by lesson or course.

and should be relatively straightforward to implement. The key areas for improvement in technical criteria involve linking course content with relevant supplementary instructional resources and ensuring that students can launch the courseware without professional assistance.

Production quality was also found to be generally strong. Narration was easy to understand, courses had minimal irrelevant content, and graphics and text were typically legible. Some courses also used animation effectively to demonstrate processes and procedures. Substantial improvement is needed, however, in eliminating sensory conflicts (which could be achieved by providing narrative text along with narration and enabling the learner to disable the audio or text) and in using multimedia effectively to make the courses more engaging. The use of animation for instruction of procedures and higher levels of IMI can accomplish this goal and also is likely to address the lack of demonstrations of procedures, which was one of the key needs we identified in our evaluation of pedagogical features of the courseware.

Modifications in technical and production-quality features together could help to attain the Army's overarching goal for DL of increasing the pace of learning while maintaining course quality. For example, directly linking course content to relevant information in FMs could provide students with a powerful tool for rapidly deepening (or sustaining) their knowledge in specific task areas. Allowing learners to replace narration with text could also increase the pace with which they proceed through course material. A summary of specific suggestions for improving these and other aspects of IMI courseware is presented in Appendix B.

Pedagogical strengths of the courses included descriptions of lesson objectives, the order of lessons, definitions of concepts, and opportunities for learners to correct their strategies in checks on learning and practical exercises. However, pedagogy was the area most in need of improvement. A pervasive problem in many of the courses was a lack of examples from the job or mission environments in instruction of both concepts and procedures. Courses also need to do a better job of demonstrating procedures and providing explanations of why procedures work the way they do, so that students can better understand the

concepts and skills taught and can adapt their behavior in nonstandard situations.

Most of the courses we evaluated provided insufficient opportunities to integrate concepts and practice procedures. The challenges in providing practice in asynchronous DL for hands-on skills, in particular, are not surprising. Whereas it is possible to provide practice opportunities in DL for tasks such as filling out forms, completing logs, or performing calculations (as some courses did), it is clearly much more difficult to provide opportunities to practice physical tasks, such as entering and clearing a building, distributing fires, using hand grenades, or eye splicing. A frequent comment from our raters about this criterion was, "Practical exercises allow user to identify correct steps in the procedure, but not perform it."

Some procedures could be taught effectively in DL with the use of job aids. For example, students could be provided with maps, compasses, protractors, or other simple devices to enable them to practice scouting procedures. In some situations, however, it may not be practical to provide job aids, and a potentially greater payoff may result from using higher levels of IMI to support many of the improvements suggested. For example, for eye splicing, it would be possible to use even Level 2 IMI to enable the learner to drag graphics in a way that would loosely resemble the procedure or to select sections of the rope to perform some action. Level 3 IMI can offer even more powerful techniques to improve the relevance and realism of the courseware and the fidelity of practical exercises, thereby enhancing opportunities for transfer of training. For example, tasks such as entering and clearing a building could use Level 3 IMI to create simple simulations (similar to video games) in which learners must select appropriate methods of entry or move team members into their correct positions; students would lose points by violating procedures (e.g., moving team members into the wrong positions or getting too close to the walls). Higher-level IMI can also increase learner engagement.

Findings regarding the use of lower-than-required levels of IMI, as well as the short time needed to complete some lessons, also indicate the need for implementation processes that ensure that courseware meets expected standards. Because the cost of IMI courseware

is a function of number of course hours and level of interactivity, it is important to verify that the Army gets what it pays for in contracting for IMI.

Before the Army undertakes new improvement initiatives, these findings and recommendations should be considered in light of changes in the courseware development process that are currently being implemented or that have been implemented since we began our research. Because the courseware production cycle is relatively long and we reviewed only fielded courses, some of those courses were funded and development began in FY2005 or before. ATSC has continually added improvements to TADLP processes since that time, including providing clearer definitions of IMI levels, requiring consultation between contractors and school staff regarding appropriate IMI levels for varying types of content, and developing a series of checklists for schools (e.g., a 2006 *Lesson Specification Worksheet*) that may have contributed to improved IMI quality.[2] ATSC also has published standards for graphical user interfaces that should result in improved screen elements and functions, which in turn should mitigate the need to evaluate some of the technical features of courseware (Army Training Support Center, 2007). Thus, courseware currently being developed may have fewer deficiencies than the courseware in our sample. Nonetheless, because ATSC has no way to document how effective its continuing efforts have been, the deficiencies noted in this report are worthy of further investigation.

Strategic Issues in Instructional Design

The preceding discussion suggests ways to modify IMI courseware in the sample that we assessed. A more strategic issue for the Army concerns determining the match between instructional technologies and the KSAs being taught. Given the importance of providing effective demonstrations of procedures and opportunities for practice in training (Clark, 2005), our findings raise questions about the value of IMI, as currently used, to provide instruction on procedures that ultimately

[2] The checklists were accessed on December 21, 2007, from http://www.atsc.army.mil/ITSD/IMI/CrsWareMgmtProponent_CkListForms.asp.

need to be practiced in hands-on training. Teaching procedures that involve physical performance requirements is especially problematic when those requirements are not practiced immediately after the cognitive portion of training. Under the current training plan for many Army courses, the DL portion of a course can be taken at any time up to a year before the student attends residential training. Furthermore, some students who have not completed the DL portion of a course are admitted to the residential portion. As a result, instructors frequently must review the DL material at the start of residential training. This is especially problematic because time is at a premium in residential training.

To the extent that the Army plans to continue using its current phased approach to training, we expect that an emphasis on training concepts and processes will have a higher payoff. Using IMI may be best reserved for training procedures when

- the procedures can be practiced within the context of IMI (e.g., completing forms or performing calculations) or with the addition of simple job aids;
- DL is used to train procedures that are not subject to rapid decay or are easily refreshed;
- DL is used as a supplement to residential training, e.g., when the IMI can be assigned as "homework" immediately preceding a practical exercise;
- exported training (i.e., institutional training executed at or near a unit location, away from the proponent school) can be supported by a high level of instructor-student interaction;
- the purpose of the instruction is to provide information where practice is not important, such as disseminating doctrinal and technique updates to the operational force.

Ultimately, however, a comprehensive research effort is needed to determine how to use IMI effectively for Army training. For example, empirical research is needed to address questions such as the following:

- To what extent do learners retain procedural knowledge taught in IMI by the time they get to residential training?
- What is an acceptable time lag between IMI and residential training to prevent knowledge or skill decay?
- What types of knowledge and skills are less subject to decay or are more easily refreshed?
- Are there differences in performance on tasks in residential training for learners who have completed the DL portion of the course and those who have not?

The Army might also benefit from examining courseware produced in other domains (other services, industry, and academia) to identify best practices in instructional design and in the use of IMI courseware. The larger project, of which this assessment is a part (see Chapter One), addresses this issue.

Below, we present additional suggestions for improving assessment of TADLP training.

Recommendations for Improving TADLP's Assessment Program

One of the conclusions from our research is that program-level assessments of courseware quality are needed on an ongoing basis. The method used in this study has a number of strengths and can effectively fill at least a part of that need for TADLP. To our knowledge, this is the first program-level evaluation of TADLP courseware. Our approach provides a systematic method of evaluation using multiple raters and a comprehensive set of criteria based on standards proposed by experts in training development and assessment. The method points to needed improvements in specific courses and yields quantifiable data that can be used to monitor progress at the program level—across courses, schools, or other units of analysis. The method also has relatively modest resource requirements; once criteria are developed and raters are trained, lessons can be evaluated relatively efficiently. In short, we believe this evaluation method is practical, could provide

the Army with valuable information about courseware quality, and points to directions for needed quality improvements. We recommend adoption of this approach at the program level for evaluating TADLP courseware on an ongoing basis.

Our approach does have has several limitations, many of which can be effectively addressed in future efforts.

First, our evaluators were not SMEs. Although we believe that SMEs are not needed to judge the criteria used in our evaluation (and that, in fact, it would be impractical for TADLP to find and train SMEs to conduct this type of evaluation for a wide range of courses), it would be useful to determine whether experts would judge the criteria differently or could conduct the evaluations more efficiently.

Second, many of the criteria in our checklist require subjective judgment. For example, there are no objective standards for the degree to which techniques engage the learner and the extent to which courses have sufficient numbers of examples, checks on learning, or practical exercises. However, a number of criteria are straightforward to judge (e.g., navigational cues either exist or do not), and even for more subjective criteria, satisfactory interrater reliability on a diverse sample of lessons provides confidence in the validity of the judgments. Nonetheless, future efforts should establish objective standards where possible, e.g., determining what constitutes a sufficient number of examples, checks on learning, or practice problems (of varying levels of difficulty) for training different KSAs. Standards could be established for tasks based on the complexity of the course material, the degree to which leaders are expected to show flexibility and adaptability, and the degree to which the skill is difficult to sustain.

Third, the criteria we used were not weighted. The impact of a rating might differ depending on the importance of the criterion and how easily it can be changed, which in turn could influence resource requirements or priorities for revision. For example, the impact of a low rating for the usability of FMs might be minimal given that improving usability may be relatively straightforward. In contrast, we expect that low ratings on criteria having to do with practice opportunities (e.g., whether exercises move from simple to complex problems or whether opportunities for whole-task practice reflect the mission environment)

will have a much greater impact, since these aspects of training are particularly important and require far more resources to fix. Therefore, future evaluation efforts should develop appropriate weights for the criteria.

Fourth, we were able to assess only a small sample of courses. In addition, we sampled lessons primarily from the first two or three modules of courses. It is possible that lessons in later modules differ systematically in production quality or pedagogy. Thus, these results should be viewed as an initial attempt to understand what is possible in terms of quality assessment and how results can be used to identify quality improvements. A larger number and wider range of courses should be assessed in efforts to further develop and validate this approach. In addition, by conducting initial evaluations of entire courses and comparing ratings across lessons, one could determine whether earlier and later lessons vary (which suggests the need to evaluate entire courses) or whether it is reasonable to evaluate a sample of lessons from each course in subsequent efforts. Clear definitions of terms such as *module*, *lesson*, and *topic* are also needed to ensure consistency in selecting segments of courses to assess in future evaluations.

Fifth, we limited our criteria, in part, to reflect information that was available in the courses to which we had access and to achieve a balance between level of detail and efficiency. In the process, we may have omitted or overlooked important aspects of the courseware or course development. Future attempts should reassess and refine the evaluation criteria. In addition, the rating options should be reviewed; for example, inclusion of a "sometimes" option may not be necessary given its low frequency of use, and standards should be developed to determine when lessons or courses merit "yes" or "no" ratings on the criteria.

Finally, most of the courses were available to us only on CDs, which do not contain course tests or support some functions of online courses, such as bookmarks and "live" online support. The CDs also presented some initial technical problems in launching the courses, which required intervention by technical support staff. For future evaluation efforts, we recommend access to fully functional courseware to permit evaluation of all course features.

Implementing a Program of Courseware Evaluation in TADLP

We recommend the following steps as elements of one possible approach to implementing a program of course evaluation:

- TADLP should assign responsibility for IMI courseware evaluation to a team of two or three persons trained in principles of instructional design. Multiple staff members are needed to establish interrater reliability. Having more than one person with this expertise also helps preserve institutional memory when individuals leave the organization. This task would constitute a portion of each staff member's general responsibilities.
- The evaluation team would refine the evaluation criteria and rating options, establishing objective standards and appropriate weights on criteria where desirable. Criteria would be disseminated to stakeholders such as proponent schools staff, SMEs, and ATSC for vetting. Use of a collaborative approach is recommended to encourage the kind of buy-in from schools and contractors that will lead to the greatest possible improvements in courseware.
- Once the checklist is finalized, the team would establish interrater reliability in use of the criteria.
- Proponent schools would provide access to online courses as well as to courses under development.
- The evaluation team would evaluate at least a sample[3] of the courseware and provide the completed checklist to the proponent school and contractor, along with specific suggestions for improvement.
- The evaluation team would create quarterly reports and an annual report that summarize the evaluations and implications for course-

[3] Sampling might be required, depending on the resource levels obtained to support the program. In addition, the sampling approach would differ if the intent is to focus on program-level rather than individual-course results. Sampling courses (rather than evaluating all courses) is a reasonable approach to monitoring quality at the program level. We suggest using a stratified random sample based on factors such as course priority (e.g., MOS-producing, PME courses, functional courses, self-development), course level, and contractor. If the focus is on individual courses, then reviewing all courses or a sample of lessons in each course would be appropriate.

ware quality at the program level. For example, the report would include the number and types of courses evaluated and a summary of scores for the evaluation criteria, similar to those found in this report. Results also can be reported by other units of analysis (e.g., school, type of course, contractor). Trends in ratings for criteria over time would indicate whether the quality of courseware is improving. Courses also can be monitored for effects of changes in training policy, development processes, or doctrine.

- These activities would be supported by a website such as the Distributed Learning Knowledge Network (DLKN) on Army Knowledge Online (AKO) to enable proponent school staff and contractors to download the checklist and access other resources, such as lessons from courses that exemplify the evaluation criteria. We also recommend providing an online mechanism for school staff to provide feedback to TADLP about the evaluation process.

We recommend that at the outset, the program focus on providing feedback to individual schools about their courses and on reporting aggregate IMI quality over time. At the school level, staff could use results to improve the quality of existing and evolving courseware. At the TADLP level, aggregate results could be used to help sell the program and obtain funding (assuming quality is high or increasing), to assess the effect of ongoing improvement initiatives on courseware quality, and to help formulate new improvement initiatives for the future.

Once the evaluation program becomes more established and accepted, the data it produces might be used more aggressively to improve courseware quality. For example, results could be aggregated by school and by contractor as part of a larger evaluation program that fosters continuous quality improvement for TADLP courseware. Relevant evaluation criteria could be incorporated into the schools' courseware validation processes to standardize and quantify these efforts or into the language of DL delivery orders. However, initiating these steps may require formulating new policies to support them, including poli-

cies relating to contracts, funding mechanisms, appropriate incentives, and development processes.

Potential Directions for Other Program-Level Evaluations of IMI Quality

As described in Chapter One of this report, courseware evaluation is only one facet of assessing the quality of training. A comprehensive DL assessment strategy would include a variety of methods. Information technology (IT) can facilitate a number of these assessments both within and across proponent schools. For example, IT can be used to develop, administer, and score tests or attitude surveys within particular courses and to analyze results across courses or proponent schools.

Table 5.1 summarizes some evaluations of IMI quality that should be conducted and describes how IT could support these efforts. Many of these evaluations could be tested on a pilot basis using a sample of courses. These recommendations are discussed further below.

Learner Reactions. End-of-course attitude surveys measuring student satisfaction can provide insights into what does or does not work in a course or why certain outcomes occur. For example, learner reactions to different aspects of courses might provide information about why students do not graduate from a course or why they choose to bypass a DL course in favor of residential training. It is reasonable to expect that there is a core set of measures that are relevant across most, if not all, DL courses. IT can be used to enable schools to access such measures, as well as add customized items and administer surveys to students at the end of courses via the Internet. An automated system could also score the surveys, create reports for individual courses, and aggregate results across courses within a school or across schools. Data could be aggregated in numerous ways, e.g., by course type, level, length, type of IMI, or characteristics of learners. We recommend, in addition to surveying graduates of DL courses, administering surveys to students who failed to complete DL courses or who were eligible for DL courses but did not enroll. Data from these students may help determine whether issues of courseware quality explain these outcomes.

Table 5.1
Using IT to Support Future Evaluations of IMI Quality

Type of Evaluation	What Could Be Done
Reactions	• Develop a core set of questions to administer to learners across IMI courses. • Develop a platform to enable schools to create/customize surveys. • Create automated scoring and reporting capabilities; start with a sample of IMI products.
Learning: pre/post comparisons	• Move to computer-based administration of course tests. • Develop platform and automated scoring/reporting procedures to support systematic analysis within and across courses; start with a sample of IMI products.
Learning: knowledge retention	• Administer computer-based tests of concepts taught in DL at the start of residential training in the phased approach to training; start with one or two courses. • Administer follow-up tests to IMI learners via AKO after they return to their units; pilot with one or two IMI courses.
Behavior (performance)	• Explore facilitators and barriers to conducting predictive validity studies.
Test evaluation	• Make item analysis of end-of-course tests an integral part of IMI via the Learning Management System (LMS); start with a sample of IMI products.
Administrative data	Expand and standardize RAND's initial analyses.

Learning: Pre/Post Comparisons. Differences in scores on pretests and posttests give some indication of learning. Posttests already exist in DL courses, and some courses have pretests as well. IT can be used here in that tests can be administered, scored, and analyzed via computer. It is important to emphasize, however, that pre/post differences can be affected by factors other than training, and determining the effect of training on changes in test scores typically requires the use of control groups. In addition, meaningful pre/post comparisons require having good tests, a point that we address in more detail below.

Learning: Knowledge Retention. Measuring knowledge retention poses administrative challenges, because tests typically need to be administered after students have completed training and have returned

to their units. Consequently, knowledge retention tends to be assessed infrequently. However, given the design of most IMI courses in the Army (as knowledge preparation for residential learning), the Army has an opportunity to begin collecting relevant data fairly efficiently and inexpensively, and such tests can be administered and scored using IT. Knowledge retention tests could be administered when students register for the residential portion of a course or at the beginning of residential training. It is possible that applicable follow-up tests already exist on a local level and that these tests could be used to assess knowledge retention.[4] To broaden the concept within the Army, an IT platform such as AKO could eventually be used to administer follow-up tests to learners once they have returned to their units.

Behavior. IT can also be used to support predictive validity studies by collecting data to measure the association of performance scores in DL with performance scores in subsequent training or with ratings of job performance. It appears that AUTOGEN (or a system like it) could provide the foundation for collecting the necessary data. The system would need to collect data regarding individuals' performance in training (e.g., in Phase 1 and Phase 2) or their performance in training and their performance on the job (e.g., leaders' ratings of individuals' job performance or other measures of performance on relevant tasks). In addition, to establish the predictive validity of DL for job performance, assessment of DL courses or tasks taught in DL would need to be differentiated from assessments of residential training or tasks. Future work should examine the strengths and weaknesses of AUTOGEN as a platform for conducting predictive validity studies by reviewing AUTOGEN's processes and reports and soliciting input about the system from proponent school staff.

Test Evaluation. IT can be used to analyze students' responses to objective test items (i.e., multiple-choice or true/false), using statistical methods such as item-response theory (Lord, 1980). Such analyses can provide information such as whether items are at the appropriate

[4] Administering tests of physical performance tasks at the start of residential training could also be used to investigate the question raised earlier about the value of providing DL training on tasks that ultimately must be practiced in hands-on training.

level of difficulty and whether the tests discriminate between good and poor performers in the course. If students complete tests online, statistical software can be used to conduct item analysis and create reports. Results can then be used to determine which test items to include, eliminate, or revise.[5]

Administrative Data. Automated systems (e.g., ATRRS) can capture data such as enrollment and dropout rates, DL usage, and information pertaining to course development, such as cycle time. As discussed previously in this report, these types of metrics can be indicators of course quality. A primary task of RAND's overall project is the analysis of existing administrative data to establish a set of baseline metrics for DL. Such systems can also be used to monitor relevant data on a regular basis and help keep the program on track with regard to TADLP objectives.

Conclusion

TADLP does not have a program-level effort to assess course quality. This report describes how the Army could benefit from such an effort and demonstrates a method of evaluating IMI courseware that is practical and can identify areas where the quality of courseware should be improved. The report also suggests other ways in which the quality of IMI training could be addressed feasibly at the program level.

In a larger context, evaluation of IMI quality should be part of a more comprehensive assessment component for TADLP. In addition

[5] Methods such as item-response theory are also used as a basis for computer-adaptive testing (CAT) or computerized classification testing (CCT) (e.g., Drasgow and Olson-Buchanan, 1999; Eggen and Straetmans, 2000). CAT, or tailored testing, is a method of administering tests that selects items based on the learner's ability level. CAT generally provides precise estimates of an examinee's ability with far fewer items than standard, fixed tests. By using much shorter tests, CAT supports the goal of speeding the pace of DL. It also enhances test security in that each learner takes a test composed of different items. CCT is similar to CAT but is appropriate for tests in which the goal is to classify learners into categories such as "pass" and "fail." CAT has been applied to the Armed Services Vocational Aptitude Battery (ASVAB) for personnel selection (e.g., Sands, Waters, and McBride, 1997). It may be worthwhile to consider using CAT or CCT for tests in DL courses with large enrollments.

to evaluating quality, that program should include an examination of learning models, courseware usage and efficiency, cycle time of production, and identification of DL areas with the highest payoff for the Army. We are currently conducting research related to these subjects. Taken together, these efforts will give the Army a basis for managing continuous improvement in the development and use of IMI and will provide tools for managing the strategic use of DL.

APPENDIX A

Courseware Evaluation Criteria

Courseware Evaluation Criteria Worksheet

Criterion	Rating					Comments (If rating is "no," describe problems here. Also provide comments on positive features of the course.)
Technical						
1. Launching[a]						
a. Learners can conduct initial launch of courseware without assistance	☐ Y	☐ N	■ S	■ NA	■ DK	
2. Navigation Functions[a]						
a. A learner with modest computer skills could easily navigate through the course (using start or enter, forward, back, exit, etc.)	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Navigation features cue the student to his/her position within the lesson (e.g., screen numbers)	☐ Y	☐ N	■ S	■ NA	■ DK	
c. Navigation features cue the student to his/her position in the course (e.g., menu of modules and lessons)	☐ Y	☐ N	■ S	■ NA	■ DK	
3. Supplementary Instructional Resources (e.g., links to glossaries, FMs, or other source materials)[a]						
a. Supplementary instructional resources are easily accessible within each lesson	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Supplementary instructional resources are usable	☐ Y	☐ N	■ S	■ NA	■ DK	

NOTES: Y = yes; N = no; S = sometimes; NA = not applicable; DK = don't know. Black checkbox indicates that the rating choice was not an option for a particular criterion.

[a]Rated once per course.

[b]Rated for each lesson.

Courseware Evaluation Criteria Worksheet (continued)

Criterion	Rating					Comments
4. Technical Support[a]						
a. Learner can resolve most navigational and technical problems with the courseware using online technical support	☐ Y	☐ N	■ S	■ NA	■ DK	
5. Legibility of Graphics and Text[a]						
a. Text and graphics are legible and accurate	☐ Y	☐ N	■ S	■ NA	■ DK	
6. Audiovisuals[a]						
a. Contains no sensory conflicts (e.g., audio and text present the same information) except when conflict is desired to create confusion as during battle	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Excludes words, pictures, and sounds that are not directly relevant to required learning	☐ Y	☐ N	☐ S	■ NA	■ DK	
c. Uses animation or video clips to demonstrate processes or concepts that are difficult to visualize from verbal descriptions or graphics	☐ Y	☐ N	☐ S	☐ NA	■ DK	
d. Uses various types of techniques to maintain learner interest and promote learning, including scenarios, examples, analogies, humor, 3-D graphics, music, etc.	☐ Y	☐ N	☐ S	■ NA	■ DK	
e. Narration is easy to understand	☐ Y	☐ N	■ S	☐ NA	■ DK	
Pedagogical						
7. Lesson Learning Objectives[a]						
a. Clearly describe what knowledge and/or skills to be learned	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Have an observable, measurable performance standard	☐ Y	☐ N	■ S	■ NA	■ DK	

Courseware Evaluation Criteria Worksheet (continued)

Criterion	Rating					Comments
8. Course Sequencing, Pacing, and Learner Control[a]						
a. Presents outline of lessons in which sequence is either the order in which job- or mission-relevant tasks are accomplished or in order of less difficult tasks progressing to more difficult tasks	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Allows students with previous knowledge/skill to test out of modules as appropriate	☐ Y	☐ N	■ S	■ NA	☐ DK	
c. Course prevents learner control of lesson sequence, activities, or pacing	☐ Y	☐ N	■ S	■ NA	■ DK	
d. If c is "yes" or "sometimes": control over sequence, activities, or pacing is inefficient for the learner	☐ Y	☐ N	☐ S	☐ NA	■ DK	
9. Teaching Concepts: Instruction[b]						
a. Provides clear and comprehensive definitions of each concept	☐ Y	☐ N	☐ S	☐ NA	■ DK	
b. Provides informative and sufficient examples from job or mission environment	☐ Y	☐ N	☐ S	☐ NA	■ DK	
c. Provides exercises requiring learners to identify examples and non-examples of each concept	☐ Y	☐ N	☐ S	☐ NA	■ DK	
d. Instruction of concepts reflects learning objectives	☐ Y	☐ N	☐ S	☐ NA	■ DK	
10. Teaching Processes: Instruction[b]						
a. Provides an informative visual model with a narrated description stating sequence of events	☐ Y	☐ N	☐ S	☐ NA	■ DK	
b. Clearly explains how actions at each phase lead to subsequent phase and final outcome of the process	☐ Y	☐ N	☐ S	☐ NA	■ DK	
c. Provides informative and sufficient examples of processes from job or mission environment	☐ Y	☐ N	☐ S	☐ NA	■ DK	

Courseware Evaluation Criteria Worksheet (continued)

Criterion	Rating					Comments
d. Provides practical exercises requiring learners to identify a list of phases, the actions that occur at each phase, and how outcomes at each phase contribute to next phase and final outcome	☐ Y	☐ N	☐ S	☐ NA	■ DK	
e. Instruction of processes reflects learning objectives	☐ Y	☐ N	☐ S	☐ NA	■ DK	
11. Teaching Procedures: Instruction[b]						
a. Provides clear step-by-step demonstration of all decisions and actions needed to complete the task based on authentic job- or mission-relevant scenarios	☐ Y	☐ N	☐ S	☐ NA	■ DK	
b. Provides informative and sufficient examples of procedures from job or mission environment	☐ Y	☐ N	☐ S	☐ NA	■ DK	
c. Where decisions are taught, includes alternatives that must be considered and criteria that should be used to choose the best alternative in routine situations	☐ Y	☐ N	☐ S	☐ NA	■ DK	
d. In cases when learners are expected to be able to modify the procedure in novel situations, explains why the procedure works using concepts, processes, or principles	☐ Y	☐ N	☐ S	☐ NA	■ DK	
e. Provides practical exercises requiring learner to perform the procedure	☐ Y	☐ N	☐ S	☐ NA	■ DK	
f. Instruction of procedures reflects learning objectives	☐ Y	☐ N	☐ S	☐ NA	■ DK	
12. Checks on Learning[b]						
a. Reinforce lesson/material	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Are sufficient in number	☐ Y	☐ N	■ S	■ NA	■ DK	

Courseware Evaluation Criteria Worksheet (continued)

Criterion	Rating					Comments
13. Practice[b]						
a. There are an adequate number of practice exercises to master each skill or concept	☐ Y	☐ N	■ S	■ NA	■ DK	
b. Part-task practice is followed by whole-task practice	☐ Y	☐ N	☐ S	☐ NA	■ DK	
c. Whole-task practice reflects the performance of the learning objectives in the mission environment as closely as media will permit	☐ Y	☐ N	☐ S	☐ NA	■ DK	
d. Practice begins with simple problems and progresses to more-complex problems	☐ Y	☐ N	☐ S	☐ NA	■ DK	
e. There is consistency between the practice exercises and the learning objectives	☐ Y	☐ N	☐ S	☐ NA	■ DK	
f. Directions for practice exercises are clear	☐ Y	☐ N	☐ S	☐ NA	■ DK	
14. Feedback[a]						
a. Feedback is frequent, so that errors do not accumulate	☐ Y	☐ N	☐ S	■ NA	■ DK	
b. Provides learner with the opportunity to correct his or her strategy	☐ Y	☐ N	☐ S	■ NA	■ DK	
c. Provides the opportunity to review the relevant parts of the demonstration	☐ Y	☐ N	☐ S	■ NA	■ DK	
d. Corrective feedback provides explanations for why the answer was wrong or why the right answer is correct	☐ Y	☐ N	☐ S	■ NA	■ DK	

General comments:

APPENDIX B

Summary of Specific Changes for Improving the Quality of IMI Courseware

This appendix summarizes recommendations for improving IMI courseware presented in Chapter Four of this report.

- Develop procedures to ensure that courseware meets expected standards for levels of IMI and lesson completion times.
- Enhance learner support by providing links throughout the courseware that take the learner directly to relevant sections of appropriate FMs and glossaries.
- Eliminate sensory conflicts by (1) enabling the learner to disable text or audio if the same content is presented in both modes simultaneously, and (2) presenting content sequentially rather than simultaneously when multiple modes (e.g., text and audio) are used for different content.
- Increase learner efficiency by providing the option of disabling the audio and enabling the corresponding text when using narration with no animation (i.e., with still graphics or no graphics).
- Control course sequence in the CD versions of courses by providing lesson menus that give free access only to content that the learner has already viewed, but disable access to "future" content that is out of sequence. Also, require learners to complete course activities before they can proceed to subsequent material. Ensure that these controls are in place for the online versions of the courses.

- Enhance context by giving examples of situations in which course concepts, processes, and procedures are used (e.g., the motivators in the Maneuver CCC).
- Improve instruction of procedures by providing clear, step-by-step demonstrations, using animations or other graphics.
- Increase soldier adaptiveness by including instructions or criteria for making decisions or modifying procedures in alternative or novel situations.
- Reinforce concepts and procedures by
  - providing more checks on learning that reflect the range of topics in the lesson and staggering the checks on learning throughout the lesson;
  - including exercises that test the learner's ability to discriminate among examples and non-examples of the course content;
  - providing practical exercises that move from simple to complex problems and include both part- and whole-task practice;
  - directing learners (via a link) to the appropriate lessons for review if they answer questions incorrectly;
  - providing more specific and comprehensive feedback in response to incorrect answers.

References

Army Training Support Center (ATSC), *Graphical User Interface for Task/Lesson-Based Development: Content Developer's Guide*, 2007. As of April 9, 2009: http://www.atsc.army.mil/itsd/imi/GUI.asp

ASTD Certification Institute, *E-Learning Courseware Certification (ECC) Standards (1.5)*, Alexandria, Va.: American Society for Training & Development, 2001–2003.

Clark, R., "What Works in Distance Learning: Instructional Strategies," in H. F. O'Neil (ed.), *What Works in Distance Learning,* Los Angeles, Calif.: University of California, Los Angeles, Center for the Study of Evaluation, 2003, pp. 13-31.

Clark, R. E., *Guided Experiential Learning: Training Design and Evaluation, A PowerPoint Presentation*, 2005. As of April 21, 2009: http://projects.ict.usc.edu/itw/gel/

Defense Manpower Data Center, *December 2005 Status of Forces Survey of Active Duty Members: Tabulations of Responses*, Arlington, Va.: Defense Manpower Data Center, 2006.

Drasgow, F., and J. B. Olson-Buchanan, *Innovations in Computerized Assessment*, Mahwah, N.J.: Lawrence Erlbaum, 1999.

Eggen, T.J.H.M, and G.J.J.M. Straetmans, "Computerized Adaptive Testing for Classifying Examinees into Three Categories," *Educational and Psychological Measurement*, Vol. 60, 2000, pp. 713–734.

Goldstein, I. L., "Training in Organizations," in M. D. Dunnette and L. M. Hough (eds.), *Handbook of Industrial and Organizational Psychology*, Vol. 2, 2nd ed., Palo Alto, Calif.: Consulting Psychologists Press, 1991, pp. 507–619.

Goldstein, I. L., and J. K. Ford, *Training in Organizations: Needs Assessment, Development and Evaluation*, 4th ed., Belmont, Calif.: Wadsworth, 2002.

Kirkpatrick, D. L., "Techniques for Evaluating Training Programs," *Journal of the American Society of Training Directors*, Vols. 13–14, 1959–1960.

———, *Evaluating Training Programs: The Four Levels*, San Francisco, Calif.: Berrett-Koehler, 1994.

Landis, J. R., and G. G. Koch, "Measurement of Observer Agreement for Categorical Data," *Biometrics*, Vol. 33, 1977, pp. 159–175.

Lord, F. M., *Applications of Item Response Theory to Practical Testing Problems*, Mahwah, N.J.: Lawrence Erlbaum, 1980.

Mayer, R. E., "What Works in Distance Learning: Multimedia," in H. F. O'Neil (ed.), *What Works in Distance Learning*, Los Angeles, Calif.: University of California, Los Angeles, Center for the Study of Evaluation, 2003, pp. 32–54.

Mayer, R. E., and R. Moreno, "A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory," *Journal of Educational Psychology*, Vol. 90, 1998, pp. 312–320.

O'Neil, H. F., *What Works in Distance Learning*, Los Angeles, Calif.: University of California, Los Angeles, Center for the Study of Evaluation, 2003.

Phipps, R. A., and J. Merisotis, *What's the Difference? A Review of Contemporary Research on the Effectiveness of Distance Learning in Higher Education*, Washington, D.C.: The Institute for Higher Education Policy, 1999.

Sands, W. A., B. K. Waters, and J. R. McBride, *Computerized Adaptive Testing: From Inquiry to Operation*, Washington, D.C.: American Psychological Association, 1997.

Society for Industrial and Organizational Psychology, Inc., *Principles for the Validation and Use of Personnel Selection Procedures*, 3rd ed., College Park, Md.: Society for Industrial and Organizational Psychology, Inc., 1987.

Straus, S. G., J. R. Galegher, M. S. Shanley, and J. S. Moini, *Improving the Effectiveness of Distributed Learning: A Research and Policy Agenda for the U.S. Army*, Santa Monica, Calif.: RAND Corporation, OP-156-A, 2006. As of April 9, 2009:
http://www.rand.org/pubs/occasional_papers/OP156/

U.S. Army Research Institute, *Distance Learning: Findings from the Fall 2006 Sample Survey of Military Personnel*, Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 2007.

Wardell, C., "Levels of Interactivity," paper presented at the 2006 Army Training Support Center DL Workshop, Williamsburg, Va., March 2006.