INSTITUTE FOR DEFENSE ANALYSES

Accelerating Development of Expertise: A Digital Tutor for Navy Technical Training

J. D. Fletcher
John E. Morrison

November 2014
Approved for public release;
distribution is unlimited.
IDA Document D-5358
Log: H 14-001221

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Project BE-2-3831, "Assessment of Revised Digital Tutor," for the Under Secretary of Defense (Personnel and Readiness), Training Readiness and Strategy Directorate. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Copyright Notice

© 2015 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-5358

Accelerating Development of Expertise: A Digital Tutor for Navy Technical Training

J. D. Fletcher
John E. Morrison

Executive Summary

Background

The value of technical expertise is as evident from empirical research as it is from casual observation. However, the years of experience and practice needed to develop technical expertise increase its cost and limit its supply. Empirical demonstrations that the time to develop technical expertise can be compressed from years into months are few, but extant. These demonstrations have relied on computer technology for their delivery. This report summarizes design, development, implementation, and assessments of a recent Defense Advanced Research Projects Agency (DARPA) effort to accelerate ab-initio (from the beginning) development of authentic expertise in information systems technology (IT) by newly recruited Navy sailors. This training was intended to compress years of knowledge, skill, and experience into 16 weeks, the time allocated by the Navy to prepare and initially qualify sailors for the IT rating.

The fundamental strategy for development was first to capture for computer delivery the practice of individuals who were expert both in a relevant area of IT and in one-on-one tutoring and then to focus on instructional objectives leading to IT expertise—well beyond the entry-level or journeyman-level abilities ordinarily targeted by introductory courses. The goal was to accelerate the development of expertise without increasing requirements to qualify for admission to the training or the time allocated to complete it.

As training (in contrast to education), the Digital Tutor program was able to focus on preparing learners to perform a comprehensive range of IT activities that graduates of the program were likely to encounter on the job. IT knowledge was assessed as essential for both retention and transfer of IT competencies, but assessment went beyond the acquisition of declarative knowledge. It emphasized the ability to solve a full range of the most difficult or unique job-sample problems that individuals with years of tacit knowledge, experience, and comprehensive understanding of IT might encounter and solve.

Design and development of the Digital Tutor was eclectic and pragmatic, based on an iterative, formative evaluation approach—building the Digital Tutor segment by segment, testing each empirically, and then revising until a satisfactory level of training effectiveness was achieved. Its instructional approach was similarly spiral. It presents conceptual material followed by authentic problems that apply the concepts. The course was intended to be as authentic, comprehensive, and epiphanic as experience obtained from long experience in the Fleet.

Using the Digital Tutor, learners interact directly with IT systems found in the Fleet. These systems were programmed to communicate and share data with each other while the Digital Tutor's information structures dynamically observed, tracked, and modeled learner progress while helping learners find and assess their own solution paths to IT problems.

Assessment

Five major formative assessments were performed during the Digital Tutor's development. The first four are briefly summarized in this report. The fifth assessment, designated IWAR 2, was the capstone assessment for the DARPA program. It provided summative evaluation of the first complete 16-week version of the Digital Tutor.

Three groups of learners were included in the assessment:

- 12 graduates of the 16-week Digital Tutor.
- 12 graduates of a 35-week classroom lecture and laboratory-oriented Information Technology Training Continuum (ITTC) course.
- 12 senior ITs averaging 9.6 years of experience in the Fleet who were selected for their superior levels of IT performance and competency.

In addition to 4 hours of knowledge testing, IWAR 2 consisted of three types of practical exercises:

- Troubleshooting by six three-member Digital Tutor, Fleet, and ITTC teams over a period of 2–1/2 days.
- A Security exercise performed by the same teams for about 4 hours.
- A System Design and Development exercise conducted for 6 hours by all six members of each group **(**Digital Tutor, Fleet, and ITTC) for the week participating in self-organized teams.

The table shows overall results of IWAR 2 testing. At least four patterns were repeated across the different performance measures:

- With the exception of the Security exercise, Digital Tutor participants outperformed the Fleet and ITTC participants on all other tests.
- Differences between Fleet and ITTC participants were generally smaller and neither consistently positive nor negative.
- On the Troubleshooting exercises, which closely resemble Navy duty station work, Digital Tutor teams substantially outscored Fleet ITs and ITTC graduates, with higher scores at every difficulty level, less harm to the system, and fewer unnecessary steps.

- In individual tests of IT knowledge, Digital Tutor graduates also substantially outscored Fleet ITs and ITTC graduates.
- Summary of Results from Assessment 5 (IWAR 2)

Summary of Results from Assessment 5 (IWAR 2)

Performance Measure	Direction	Significance[a]	Effect Size[b]
DT versus Fleet			
Problem Solving (PS) Total Score	DT > Fleet	<.0001	4.19
PS Harmful Actions	DT > Fleet	<.0001	–1.85
PS Unnecessary Steps	DT > Fleet	<.0001	–2.26
Review Board	DT > Fleet	<0.01	1.07
Security Exercise	DT > Fleet	N.S.	–0.97
Network Design and Development	DT > Fleet	N.S.	0.74
Knowledge Test Total Score	DT > Fleet	< 0.0001	3.11
DT versus ITTC			
Problem Solving (PS) Total Score	DT > ITTC	<.0001	7.98
PS Harmful Actions	DT > ITTC	<0.01	–1.63
PS Unnecessary Steps	DT > ITTC	<.0001	–2.10
Review Board	DT > ITTC	<0.05	0.89
Security Exercise	DT > ITTC	N.S.	–0.03
Network Design and Development	DT > ITTC	<0.01	1.52
Knowledge Test Total Score	DT > ITTC	<0.0001	3.54
ITTC versus Fleet			
Problem Solving (PS) Total Score	ITTC > Fleet	N.S.	–1.33
PS Harmful Actions	ITTC > Fleet	N.S.	0.06
PS Unnecessary Steps	ITTC > Fleet	N.S.	0.60
Review Board	ITTC > Fleet	N.S.	0.32
Security Exercise	ITTC > Fleet	<0.05	–1.92
Network Design and Development	ITTC > Fleet	N.S.	–1.04
Knowledge Test Total Score	ITTC > Fleet	N.S.	0.77

[a] Two-tailed probability from *t*-test for independent means.

[b] Negative Effect Sizes are opposite of the indicated direction

DT = Digital Tutor

ITTC = Information Technology Training Continuum

An appropriate human tutor was not available for the Security component of the Digital Tutor design and development. Much of its training was presented in lecture mode.

This situation may explain the poor performance of Digital Tutor participants in this exercise. It may also reinforce the importance of modeling tutoring with individuals who are expert in both the subject matter and one-on-one tutoring in designing and developing any digital tutor.

Overall, if Digital Tutor graduates had matched Fleet IT performance in the practical exercises, the goals of the program to accelerate acquisition of expertise would have been met. Instead, the Digital Tutor students outscored Fleet participants with years of on-job training by substantial margins in performing job-sample practical exercises. It is also notable that the Digital Tutor graduates substantively outscored ITTC graduates who had spent more than twice the time in training.

Ancillary Issues

Three ancillary issues were examined. Results from a comparison of human and digitally tutored participants were mixed, showing superior performance for digitally tutored participants on one part of the Knowledge test, but not on the other, nor on both parts combined.

Gini coefficients were used to assess the fairness or equality of learning provided to learners across the spectrum. They found the digital tutoring to be more equitably distributed than that provided by classroom instruction.

Finally, reading ability measured by a standard test of reading vocabulary and reading comprehension was found to account for little of the variance in IT knowledge among Tutor graduates, but about 25% of the variance among the ITTC classroom graduates.

In summary, it seems reasonable to conclude that development of expertise can be substantially compressed and accelerated in technical training, that this capability is of considerable monetary and operational value, and that it should be vigorously pursued.

Contents

1. Introduction 1
   A. Expertise in Information Technology 1
   B. Expertise and Digital Tutoring 2
2. The DARPA Digital Tutor 7
   A. Why Information Technology? 7
   B. Identifying Learning Objectives 8
   C. Identifying Human Tutors 8
   D. Increasing Effectiveness Versus Reducing Time 9
   E. Tutor Design and Development 9
3. Assessments 15
   A. Schedule 15
   B. Comparison Groups 16
   C. Statistical and Practical Significance 18
4. Assessments 1–4 21
   A. Assessment 1—April 2009 21
      1. Measures 21
      2. Results 21
   B. Assessment 2 21
      1. Measures 22
      2. Results 22
   C. Assessment 3—April 2010 23
      1. Measures 23
      2. Results 23
   D. Assessment 4—November 2010 24
      1. Measures 24
      2. Results: Troubleshooting 24
      3. Results: Packet Tracing 25
      4. Review Board Interviews 25
      5. Knowledge Test 26
5. Assessment 5 27
   A. Participants 27
   B. Support Teams 28
   C. Facilities 28
   D. Schedule 28
   E. Measures 29
      1. Practical Exercises 29
      2. Troubleshooting Exercise 29

3. Security Exercise 31
4. System Design and Development 31
5. Review Board Interviews 32
6. IT Knowledge Test 33
6. Results 35
A. Troubleshooting Exercises 35
1. Troubleshooting Scores 36
2. Harmful Errors 38
3. Unnecessary Solution Steps 38
4. Review Board Interviews 39
5. Security Exercise 40
6. Network Design and Development 41
7. Knowledge Test 42
B. Assessment 5 Summary 43
C. Additional Analyses 45
1. Is Digital Tutoring as Effective as Human Tutoring? 45
2. Is Digital Tutoring Equally Beneficial for all Students? 46
3. Does Digital Tutoring Effectiveness Depend on Reading Ability? 48
D. Discussion 49
1. Corroboration from the Veteran's Project 50
2. Comparison Groups 50
3. Focus—Learning or Theory 51
4. Blending with Human Monitors and Mentors 51
5. Discovery and Guidance 52
6. Abductive Reasoning 52
7. Mixed-Initiative Dialogue 52
8. Minimize Cost or Maximize Effect 53
9. Cost 53
10. Novice, Journeyman, and Expert 54
11. Return on Investment 54
12. Authoring Systems 54
7. Final Word 57
References A-1

1. Introduction

The value of technical expertise is as evident from empirical research as it is from casual observation (Ericsson et al. 2006; Hoffman, et al. 2014). However, the years of experience and practice needed to develop technical expertise increase its cost and limit its supply. Empirical demonstrations that the time to develop technical expertise can be compressed from years into months are few, but extant. These demonstrations have relied on computer technology for their delivery.

For instance, the Sherlock project (Lesgold et al. 1988) prepared technicians to solve complex problems occurring in a test stand used to troubleshoot components of Air Force avionics systems. Assessments found that 20–25 hours of Sherlock training produced about the same improvement in performing difficult and rarely occurring diagnostic tasks as 4 years of on-job experience (Gott and Lesgold 2000; Lesgold and Nahemow 2001).

Additional evidence was provided by IMAT, the Navy's Interactive Multi-Sensor Analysis Trainer (Wetzel-Smith and Wulfeck 2010). An at-sea trial found that 2 days of training with a laptop version of IMAT increased submarine effective search area by a factor of 10.5 (Chatham and Braddock 2001). In effect, a submarine with IMAT-trained sonar operators could provide the sonar surveillance of 10 submarines lacking operators with IMAT training. The operational and monetary value of this capability is substantial.

This report concerns Navy technical training. It describes a recent effort to accelerate ab-initio (from the beginning) development of authentic expertise in Information Systems Technology (IT)[1] for newly recruited Navy sailors.

A. Expertise in Information Technology

Designing, maintaining, and troubleshooting our increasingly ubiquitous IT systems require multidimensional abilities in solving complex problems. Problems can occur with local workstations, their connection to a network, the network itself, receiving workstations, security and administration practices, the source and version of all software involved, user error or intentional misbehavior, and even the actions of IT technicians themselves. The background that these technicians must master ranges from fundamentals of computer and network hardware, through ever-changing system, network, and

[1] In common with Navy usage, "IT" in this report refers to Information Systems Technology, as in "IT rating," and to Information Systems Technician(s), as in "an IT."

application software, to a multitude of administrative factors concerning user services, permissions, groupings, and subgroupings.

Expertise in IT appears to meet Wetzel-Smith and Wulfeck's (2010) definition of incredibly complex tasks. They describe these as broad, multifaceted, abstract, co-dependent, nonlinear tasks that require a large repertoire of patterns and pattern-recognition capabilities. Clark and Wittrock (2000) suggested that experts employ "X-ray vision" to see through the surface features and symptoms of problems to identify causal structures and principles underlying them. Sternberg and Hedlund (2002) describe the ability to do so as tacit knowledge—the latent knowledge acquired by experts from years of experience that they can only partly articulate but that enables them to solve complex, real-world problems.

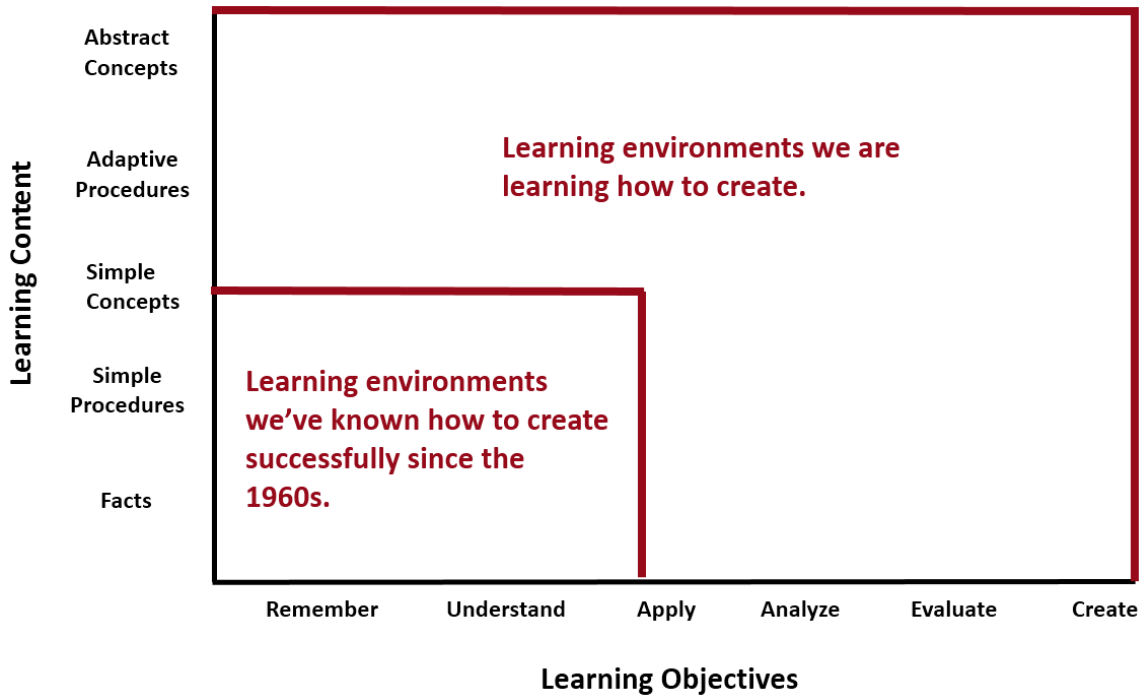
Individuals who develop expertise purposively abstract basic principles from their experiences and seek out novel situations and challenging problems as they become available (Ericsson 2006). A key to accelerating expertise, then, is to provide opportunities for authentic, deliberate practice that compresses years of relevant, insight-producing experiences into weeks or, at most, months.

B. Expertise and Digital Tutoring

Learning in most subjects begins with basic facts, nomenclature, and simple procedures. Objectives for such essentials are found at the low end of Bloom's often-referenced hierarchy of instructional objectives (Bloom et al. 1956) and the lower left-hand corner of its two-dimension extension by Anderson and Krathwohl (2001), which is adapted and displayed in Figure 1.

Considerable data from the 1960s, 1970s, and onward have found that these rudiments can be learned efficiently and effectively through computer-assisted drill and practice (e.g., Jamison, Suppes, and Wells 1974; Kulik 1994; Niemiec and Walberg 1987; Suppes and Morningstar 1972; Suppes, Fletcher, and Zanotti 1975, 1976; Vinsonhaler and Bass 1972). Most successful drill-and-practice programs focus on a large body of discrete associations such as arithmetic facts, vocabulary words, dates in history, technical nomenclature, etc. Some of these programs applied sophisticated optimization routines to accelerate learning. They selected and presented items intended to optimize the number of items students learn given constraints such as time available, learners' prior knowledge, and learners' progress (Atkinson 1968, 1972; Atkinson and Paulson 1972; Chant and Atkinson 1978; Groen and Atkinson 1966; Suppes 1964). Comparisons of these early programs with conventional classroom instruction generally found effect sizes averaging around 0.40 (e.g., Fletcher 2004; Kulik 1994).[2]

[2] See Table 2 for suggestions on interpreting effect sizes.



Source: Adapted from Anderson and Krathwohl (2001).

Figure 1. Overview of Learning Objectives

However successful these drill-and-practice programs were, few aimed to develop the conceptual understanding that is characteristic of expertise. Enabling Clark and Wittrock's (2000) X-ray vision requires a level of conceptual abstraction beyond the straightforward associations built up by drill and practice, however efficiently and effectively that may be done.

Much instruction intended to develop expertise relies on guided, authentic, and situated environments such as those being advocated by constructivists (Tobias and Duffy 2009). Many of these environments develop learners' repertoire of strategies, enabling them to rapidly shift their approaches to problems as needed (Feltovich, Prietula, and Ericsson 2006). These environments also prompt learners to seek higher levels of knowledge by encouraging reflection on principles abstracted from specific experiences. Such reflection allows learners to attain levels of knowledge that are persistent and transferable (Craik and Lockhart 1972; Bourne, Raymond, and Healy 2010; Gick and Holyoak 1980; Healy, Kole, and Bourne 2014; Mayer 2002; McDaniel, Cahill, Robbins, and Wiener 2014; Tobias 1989).

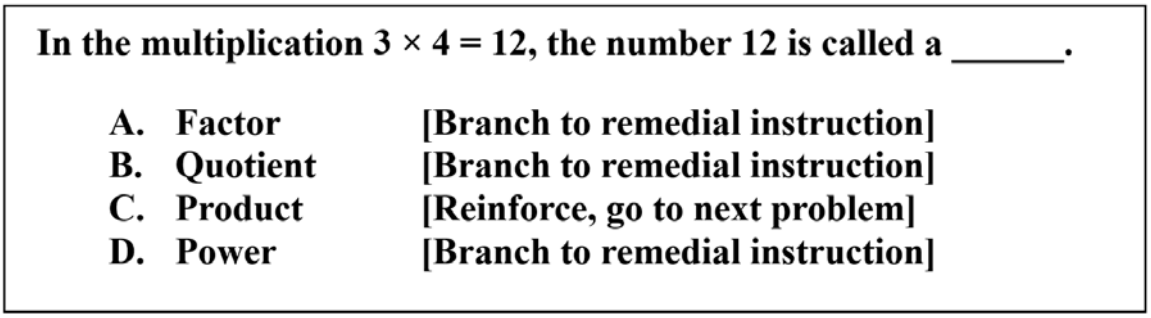
Instructional environments of this sort can be readily provided by one-on-one human tutoring, which has been shown to be significantly more effective than many-on-one classroom instruction (Bloom 1984; Graesser, D'Mello, and Cade 2011; VanLehn 2011;

Kulik and Fletcher, in press). But the economic argument against providing an individual human tutor for every student is, with few exceptions, obvious and decisive.

Over the last 50 years, this understanding has motivated the development of computer systems, which are readily affordable and available, with sufficient intelligence to generate in real time the tutorial interactions that are characteristic of expert human tutoring (Carbonell 1970; Feurzeig 1969; Fletcher 2009; Fletcher, Tobias, and Wisher 2007; Woolf and Regian 2000). The goal of these systems is to operate much in the way that expert human tutors do, but to do so affordably.

Based on initial development of MENTOR by Feurzeig (1969), semantic networks by Quillian (1969), and his own work on SCHOLAR, Carbonell (1970) identified two basic features that distinguish this instructional approach from computer-assisted drill and practice:

- Use of *information structures* in place of pre-programmed, frame-oriented exercises such as Crowder's (1959) ubiquitous Intrinsic Programming approach, which was adapted for computer presentation from paper-based programmed learning. Figure 2 is an example of an intrinsic programming item. Information structures, such as those based on ontologies, concept maps, natural language understanding, and one-on-one tutorial strategies, relieve developers from the need to anticipate every state that might exist for individual learners and the instructional system.



Source: Adapted from Crowder (1959).

Figure 2. Intrinsic Programming Example

- Mixed-initiative *tutorial dialogue*. These systems allow either the computer-tutor or the learner to initiate relatively open-ended questions during an instructional dialogue. Either may take the initiative in asking questions, posing problems, and providing explanations. The tutor should be prepared to provide guidance and assist the learner before the learner knows what questions to ask.

These capabilities, which allow computers themselves to "author" instructional material in real time, were a major, early incentive for applying machine intelligence in

this area (Brown, Burton, and Zdybel 1973; Brown, Burton, DeKleer 1982; Feurzeig 1969; Fletcher 2009; Fletcher and Rockway 1986).

2. The DARPA Digital Tutor

Development of the Digital Tutor was undertaken by the Defense Advanced Research Projects Agency (DARPA) in cooperation with the United States Navy. The project was to serve two broad purposes—advancing the technology of computers used in education and training and meeting the needs of the operational Navy.

The Tutor was to capture in computer technology the capabilities of individuals who were recognized experts in a specific area of IT and proficient in one-on-one tutoring. Its training objectives were intended to substantially exceed the entry and journeyman levels targeted by existing IT training without increasing training time. Unlike many technology-based systems where the technology is intended to support human-delivered classroom instruction, the opposite was the case for the Tutor. Human mentors supported the instruction being delivered by the Tutor. A 16-week training period was chosen to match that of existing initial qualification training for the Navy's IT rating.

The Tutor established a problem-based learning environment for each learner that was managed and controlled by information structures. It used these structures, which are functionally similar to those of human tutors, to establish a problem-solving environment that:

- Presents the subject matter.
- Establishes training objectives and requisite competencies.
- Generates models of learners that evolve with their learning progress.
- Adapts and assigns problems that optimize individuals' learning subject to constraints of time and learning progress.
- Shadows and assesses learners' problem-solving efforts.
- Ensures learners' understanding of issues and concepts underlying the problems presented.

A. Why Information Technology?

In preparing for this project, DARPA reviewed technical training schools (i.e., courses) across the Department of Defense. The review assessed (1) criticality of the human performance the training was intended to produce; (2) recognition at all echelons of operation and training commands of a need to improve and revise the existing training; and (3) technical opportunities, including digital tutoring, for meeting this need.

DARPA identified 42 technical domains as targets for investment. Navy training for the IT rating was among the most prominent of these. Navy operations, ships, and systems are increasingly dependent on IT. When IT fails, naval operational capability suffers, sometimes catastrophically.

Each year about 2,000 sailors, newly graduated from recruit training, attend an "A" school course to qualify for the Navy's IT rating (job classification) and civilian IT certifications. Despite this effort, about 5,000 trouble tickets that cannot be solved by uniformed IT technicians in the Fleet are referred each year to shore-based civilian experts. The situation has produced a rapidly growing operational demand in the Navy for IT problem-solving expertise.

B. Identifying Learning Objectives

No activity is more essential to the success of a training program than determining its objectives—the knowledge and skills that learners must acquire. Training differs from education in that it is a means to an identifiable end—performance of a specific job or task. The difference is a matter of emphasis, but training effectiveness depends heavily on accurate and comprehensive analysis to identify the knowledge, skills, and standards required by the identifiable activities it is preparing people to perform. Most training contains elements characteristic of education, and most education contains element characteristic of training. Preparation for transfer and retention is essential in both, but training objectives—the ability to perform authentic practical exercises that integrate knowledge and skills—are more focused than those of education and are central in evaluations of training.

The design of the Digital Tutor was notable for its commitment to detailed and thorough analysis of the knowledge, skills, and standards required for high-level IT performance. This analysis included a careful review of reference materials and existing IT courses. However, it also involved numerous observations and interviews with IT technicians in Fleet assignments who were identified by managers, peers, and subordinates for their expertise. Design of the Tutor focused on the key experiences, problems, and insights that had contributed to their expertise over the years so that they could be compressed into weeks.

C. Identifying Human Tutors

A related strategic commitment was the care used to collect, observe, and replicate the best practices of highly skilled human tutors. Candidate civilian experts were identified by their peer-acknowledged expertise, publications, and contributions to IT technology, covering one or more of the sub-domains (e.g., routers, networks, operating systems, group policy) that had been identified as IT training objectives. These candidates were then examined for their ability to perform one-on-one tutorial dialogues with individual

learners—a setting that differs appreciably from one-on-many classroom settings as Graesser, Person, and Magliano (1995) and Graesser, D'Mello, and Cade (2011) have found. Candidates were auditioned in half-hour sessions tutoring a representative student in a topic of the candidate's choice. Twenty-four individuals were chosen by this process to provide (human) tutorial instruction in their IT sub-domain for capture and replication by computer.

D. Increasing Effectiveness Versus Reducing Time

Trade-offs between costs and effectiveness are commonly considered in designing programs of instruction—although they are not always well informed by data. The Digital Tutor held learners to a scheduled time (about a week) for each of its subtopics (e.g., Group Policy, Internet Protocol, Active Directory, Windows Operating System, User Account Management, Network Topology). Fast-paced learners who reached targeted levels of learning early were given more difficult problems, problems that dealt with related subtopics that were not otherwise presented in the time available, problems calling for higher levels of understanding and abstraction, or challenge problems with minimal (if any) tutorial assistance. The basic approach was to hold segment time (a proxy for costs) constant for all learners while enabling individual learners to achieve all they could (maximizing effectiveness) in the time available. This approach differs from self-paced approaches that minimize time (costs) to achieve baseline training objectives and standards (holding effectiveness constant) for all graduates.

This approach required the Tutor to maintain continuous, interactive learner involvement by selecting and tailoring problems for each learner as an individual. Observers noted that the tutor established the same kind of concentration, involvement, and flow that is characteristic of interactive computer game playing (Csikszentmihalyi 1990).

E. Tutor Design and Development

The design and development of the Digital Tutor began with human tutoring. About half its funding was used to identify and recruit individuals who were expert both in specifically identified IT topics and in one-on-one tutoring and then to create and run a comprehensive, human-tutored, 16-week IT course. This course became the basis for designing the Digital Tutor and demonstrating economic scalability for accelerating the acquisition of expertise.

All human tutor and student interactions were recorded in audio. These interactions were reviewed by a Content Author—generally one of the human tutors who presented the material in the first place. The Content Author then worked closely with a Content Engineer—an individual with a deep understanding of the Digital Tutor and its software architecture—to develop its operating characteristics and instructional presentations. All

work was performed under the direction of a Course Architect who was responsible for the overall direction and flow of the instruction during the full course of its operation.

The Tutor relies on a back-end that involves classic knowledge engineering—feature extraction, ontologies, and inferencing. The Inference Engine captures problem-solving processes—the strategy and path each learner is using to solve problems based on what the learner understands or misunderstands. To enhance the scalability of the system, allowing it to train thousands of students at remote locations cost-effectively, the Tutor employs a server-based architecture with its own "cloud"—server racks in a central location. This architecture required the Inference Engine to operate at speeds beyond those typically found elsewhere. These speeds were needed to rapidly process hundreds of thousands of data points used to convert low-level data about student actions and states into semantic constructs.

The Inference Engine passes its findings to an Instruction Engine that decides what instructional element(s) each learner should address next. These decisions are intended to maximize the learner's progress by assigning problems that fill in information the learner may have missed or misunderstood, or understood at an insufficiently deep level. The Instruction Engine also decides when to ask the learner to reflect on what's been learned and incorporates that in its dynamic model of the learner's knowledge, skills, and progress. This Engine then uses the information to decide what to specify next, what instructional technique to use, and how to present it. In keeping with the Tutor's tactics, it must distinguish between typing errors made by the learner and real misconceptions.

A Conversation Module then applies results from the Inference and Teaching Engines to engage learners in natural language tutorial dialogues that avoid leading and telling but, in a Socratic fashion, emphasize asking. It uses a Recommender to determine when to call assistance from a human monitor/mentor. It uses natural language processing, but it does not attempt full natural language understanding. It relies on the Recommender to decide if and when to summon a human monitor/mentor.

The Conversation Module administers frequent, substantive tests of learners' factual and conceptual understanding to assess their learning progress. In addition to short answer, factual responses, the tests frequently dive deeper using both Inference and Teaching Engine findings to require explanatory sentences and short paragraphs to determine learners' understanding, integration, and use of the concepts underlying problems and solutions. In debriefs, learners often reported using specific basic concepts to solve problems.

In solving problems, which are central to the Tutor's instruction, learners interact with the real systems used in the Fleet and a real network hosted on two standard UNIX workstations—a client and a server—and a tutor. These systems communicate and share data with each other while the Tutor's information structures observe, track, and model

learner progress and solution paths. Rescue and restart of the Fleet systems are not unusual during problem-solving exercises. Exercises in each IT subarea evolve from a few minutes and a few steps to open-ended 30–40 minute problems.

In this sense the Tutor's instruction is not based on simulation, but in another sense it is. Learners are not pitching, rolling, and yawing on a duty station ship at sea with all the additional responsibilities, concerns, and assignments that Fleet duty requires. As with combat, if it is not the real thing, it is simulation.

The Tutor uses different training approaches for problems of different complexity. It presents the authentic and contextualized knowledge needed to solve simple problems. For complicated problems that concatenate multiple simple problems, the Tutor provides prompting, as discussed by Anzai and Simon (1979), to help individuals analyze problems into their component parts and monitor for their solutions. For complex problems that require unique and potentially epiphanic ("aha") solutions, the Digital Tutor prompts for reflective explanations to reveal and emphasize their deep structure and analogies with other problems, in accord with research findings by Healy et al. (2014), Mayer (2002), McDaniel et al. (2014), among others.

The Digital Tutor has specific features intended to promote this learning:

- Frequent and substantive dialogue interaction with learners.
- Authentic, situated problem-solving.
- Continual diagnostic assessment of individual learning and progress.
- Required reflection on concepts illustrated by problem content and processes.
- Integration of human monitors and mentoring.

The DARPA Digital Tutor relies considerably on the use of information structures to provide pedagogical support for the learner and to parse free-form learner responses, but it does not attempt full, free-form tutorial conversations. Because of its natural language parsing abilities and the specialized language of IT, the Digital Tutor can understand most natural language responses from the students and provide appropriate responses for most states of the system and students. The Recommender, however, can recognize when consultation with a human monitor is needed and advise the Conversation Module of this situation.

The classroom monitors were uniformed Navy personnel assigned to the school as instructors. They performed proctor duties for the course and conducted the end-of-day study halls. Study halls included exercises with IT hardware—a tactile aspect of IT that must be taught directly and can be readily presented by Navy IT instructors. Most important, the instructors provided a Navy presence and orientation into the culture, traditions, and practices of the U.S. Navy for young sailors. "Sea stories" are a welcome

and necessary component of training for novice sailors, even, perhaps especially, for those engaged in technical training, however advanced it may be. It is as hard to deny the value of a Navy presence and sea stories for young sailors as it is to quantify it.

Overall, the technical design and development of the Tutor was eclectic and pragmatic, based on an iterative, formative evaluation approach—building the Tutor segment by segment, testing each empirically, and then revising until a satisfactory level of effectiveness was achieved. Basically, the Tutor provides guided, authentic, and practical problem-solving experience with Navy IT systems, workstations, networks, and administrative policies.

The instructional strategy is similarly spiral, mirroring its iterative development strategy. It presents conceptual material followed by problems that apply the concepts and are intended to be as authentic, comprehensive, and epiphanic as those obtained from years of IT experience in the Fleet. Once the learner demonstrates sufficient understanding of the material presented and can explain and apply it successfully, the Digital Tutor advances either vertically, to the next higher level of conceptual abstraction in the topic area, or horizontally, to new but related topic areas.

Instructional tactics and procedures embodied in the Digital Tutor include the following:

- Promoting learner reflection and abstraction by:
  - Prompting for antecedents, explanations, consequences, or implications of answers.
  - Questioning answers, both right and wrong, and asking learners what they have learned so far in their problem-solving.
  - Probing vague or incomplete explanations and other responses by the learner.
- Reviewing knowledge and skills already acquired when the learner reaches an impasse or displays a misconception by asking why something did or did not happen.
- Rarely providing a correct answer or a direct hint.
- Never articulating a misconception.
- Sequencing instruction to pose problems that are selected and tailored to optimize each learner’s progress.
- Requiring logical, causal, or goal-oriented reasoning in reviewing or querying steps taken by the learner to solve problems.

- Refocusing dialogue if the learner's responses suggest absent or misunderstood concepts that should have been mastered.
- Continuing problem-solving until the learner discovers a careless error in applying a concept already mastered,
- Verifying learner understanding of any didactic material before proceeding.

These tactics reflect practices of the expert tutors chosen to present IT topics covered in the human-tutored, preliminary version of the course and findings from the empirical development of the Tutor. Those familiar with research on tutoring will note they address issues commonly discussed in instructional theory, even though they arose empirically in developing the Tutor.

3. Assessments

A. Schedule

Five summative assessments, as shown in Table 1, were performed during development of the Tutor. Assessments 1–4 are summarized in this section. Assessment 5 (designated by DARPA as IWAR 2), is the capstone of the program and described in more detail in the next section.

Assessment 1 was undertaken by the Navy school that provides IT training at Corry Naval Technical Training Station, Florida. The remaining four were performed by the Institute for Defense Analyses (IDA), a Federally Funded Research and Development Center, acting as an independent third-party evaluator.

The first week of the Digital Tutor was available on computer and used by the sailors who were examined in Assessments 1 and 2. The remaining 15 weeks of the course were then taught in one-on-one sessions using human tutors covering the subarea of IT in which they were particularly expert. Their tutoring served as the basis for subsequent design and development of the Digital Tutor, which was the instruction for Assessments 3–5.

Assessment 1 compared IT knowledge obtained by sailors who had finished 10 weeks of the 16-week, mostly human-tutored course with that of graduates from the self-paced IT "A" school Integrated Learning Environment (ILE) course, which sailors finished in an average of 10 weeks.

Assessment 2 (IWAR 1) provided a more summative evaluation, comparing the knowledge and troubleshooting skill of the mostly (15 of 16 weeks) human-tutored learners with those of senior ITs who averaged 7.2 years of Fleet experience.

Assessment 3 compared the IT knowledge produced by the partly completed (4 weeks) Digital Tutor with that of ILE graduates and that of their instructors. The Digital Tutor students returned to the ILE "A" school after the assessment.

Assessment 4 compared the IT knowledge and problem-solving skills provided by the partly completed (7 weeks) Digital Tutor with three groups—ILE graduates, graduates of an intense 19-week Information Technology of the Future (IToF) classroom course, and Navy instructors assigned to the IT school. After Assessment 4, Digital Tutor students returned to the IT "A" school.

Assessment 5 (IWAR 2) compared the knowledge and practical skills obtained by graduates of the fully competed 16-week Digital Tutor with those of a 35-week classroom Information Technology Training Continuum (ITTC) course and senior ITs who averaged

9.6 years of IT experience in the Fleet. Tutor graduates went directly to Fleet assignments after Assessment 5.

B. Comparison Groups

Comparison groups in one or more of the assessments were the following:

- ILE graduates who participated in Assessments 1, 3, and 4. The ILE course was the standard "A" School training that qualified recent graduates of recruit training for the IT rating. It was designed in accord with the Navy's standard ILE procedures and practices (U.S. Naval Education and Training Command 2010). It consisted of self-paced, frame-oriented, computer-assisted instruction. Sailors were given 16 weeks to finish the course and averaged about 10 weeks to do so. After that they were sent directly to Fleet assignments. The limited success of ILE graduates in Assessments 1 and 3 indicated that they did not constitute a robust control group. In Assessment 4 they were only required to complete the written test, and they were excused from participation in Assessment 5.
- Fleet ITs with an average of 7.2 years (Assessment 2) and 9.6 years (Assessment 5) experience of Navy IT duty. Considerable effort was made to ensure that Fleet ITs selected for these assessments were, in the judgment of their supervisors and peers, the most proficient in port at the time of the assessments.
- Navy instructors. Assessments 3 and 4 included administration of a written knowledge test to Navy IT instructors drawn from those assigned to teach IT courses at the school.
- Information Technology of the Future (IToF) graduates. IToF graduates included in Assessment 4 had received 19 weeks of classroom and laboratory training provided by highly motivated Navy instructors who were chosen for this course.
- Information Technology Training Continuum (ITTC) graduates. ITTC graduates included in Assessment 5 received the 19-week IToF course followed by 16 weeks of additional training presented by instructors who were qualified to prepare IToF graduates for civilian IT certifications. Much of the training was based on current versions of Microsoft and CISCO instructional material.

There were numerous similarities in the objectives assumed by the courses. For example, all the objectives included in the ILE and 19-week courses were included in the Digital Tutor course, but the Digital Tutor covered these objectives in more depth in terms of concepts and IT principles than the other courses did.

Table 1. Assessments and Comparison Groups in the Digital Tutor Program

	Assessments				
	1 April 2009	**2 (IWAR 1)** July-August 2009	**3** April 2010	**4** November 2010	**5 (IWAR 2)** March-April 2012
Treatment	Human Tutoring	Human Tutoring	Digital Tutoring	Digital Tutoring	Digital Tutoring
Comparison Groups (Number/ Duration of training or experience)	Tutored (15/10 weeks)[a] ILE (20/10 weeks)[b]	Tutored (12/16 weeks)[a] Fleet (12/7.2 avg. years of experience)	DT (20/4 weeks) ILE (31/10 weeks)[b] Instructors (10)	DT (20/7 weeks) ILE (17/10 weeks)[b] IToF (20/19 weeks) Instructors (10)	DT (12/16 weeks) ITTC (12/35 weeks) Fleet (12/9.6 avg. years of experience)
Testing	IT Knowledge (written)	IT Troubleshooting (practical) IT Security (practical) System Design (practical) IT Knowledge (written)	IT Knowledge (written)	IT Troubleshooting (practical) IT Packet Tracing (practical) Board Interviews (oral) IT Knowledge (written)	IT Troubleshooting (practical) IT Security (practical) Board Interviews (oral) System Design (practical) IT Knowledge (written)

[a] Includes the first week of Digital Tutoring that was available, human tutoring thereafter.

[b] Self-paced, average of 10 weeks to complete.

DT = Digital Tutor

ILE = Integrated Learning Environment

IToF = Information Technology of the Future

ITTC = Information Technology Training Continuum

Armed Forces Qualification Test (AFQT) scores, a measure of general mental ability, did not differ statistically across the participants. In all these assessments, however, ILE, IToF, and ITTC students averaged about 4 percentile points higher on the AFQT than Digital Tutor students. Fleet ITs averaged about 1 percentile point higher than the Digital Tutor students.

Human instructors were monitors and mentors for the ILE and the Digital Tutor course. ILE instruction presented 8 hours of computer-based training daily. The daily schedule for IToF and ITTC training was 8 hours of classroom lecture, supplemented by laboratory exercises and use of a ship's radio room simulator. Digital Tutor instruction was typically used each day for 6 hours of IT training followed by a 2-hour Navy instructor–led study hall with about five students. Learning-to-learn skills, such as note-taking, reading for content, and studying were included in early study halls for the Digital Tutor learners. Aside from that, the content, quality, and structure of these study halls varied at the discretion of the instructor in charge. An IT expert was often brought in for the last DT study hall of each week to address particularly difficult issues and questions.

C. Statistical and Practical Significance

Tests of statistical significance were included in all five assessments.

Effect sizes are descriptive statistics commonly used to estimate the practical significance and magnitude of different treatments. They were calculated for this report using Hedges' g (1981). Calculation of effect size remains a matter of discussion. The means reported here are followed by their standard deviations enclosed in parentheses to allow alternative calculations.

Interpreting effect sizes is also a matter of discussion. The interpretations used in this report are shown in Table 2. The What Works Clearinghouse (U.S. Department of Education 2009) considers effect sizes of 0.25 or higher to be of substantive importance for instruction. Beyond that, Cohen (1988) offered rough guidelines for interpreting effect-size values. He tentatively recommended characterizing effect sizes as small, medium, and large, roughly as shown in Table 2. We added very large to characterize effect sizes in excess of 0.80. Bloom (1984) famously challenged instructional researchers to develop instructional capabilities that reliably produced effect-size improvements of 2.00 standard deviations over current practice.

Table 2. Effect Size Interpretation

Effect Size	Suggested Interpretation[a]	50th Percentile (Roughly) Raised To …
ES < 0.25	Negligible[b]	59th percentile
0.25 < ES < 0.40	Small	60th–65th percentile
0.40 < ES < 0.60	Moderate	66th–72nd percentile
0.60 < ES < 0.80	Large	73rd–78th percentile
ES > 0.80	Very Large	79th percentile and up
ES > 2.00	Bloom's challenge[c]	98th percentile and up

[a] Extended from suggestions by Cohen (1988).

[b] U.S. Department of Education, *What Works Clearinghouse,* 2012.

[c] Bloom (1984).

4. Assessments 1–4

A. Assessment 1—April 2009

The human-tutored IT course began with 15 sailors who had completed recruit training and were selected by the Navy IT school for assignment to the Defense Language Institute in Monterey, California, where the necessary spaces—cubicles for one-on-one tutoring and rooms for computers and other equipment—were available. Although described here as human-tutored, the first week of Digital Tutor instruction was available and used for this training. Aside from a few sessions when the students were taught in groups of three, the remaining 15 weeks of the course were conducted as one-on-one human-tutored sessions. Notably, students often received tutoring from different tutors on the same subtopic to give them different perspectives on what they were learning. The comparison group for this assessment consisted of 20 sailors at Corry Station, Florida, who had completed the standard ILE training described above and were chosen by the Navy school for this assessment.

1. Measures

The assessment was performed by the Navy IT school. It used a written paper-and-pencil test prepared by the school to compare the IT knowledge of the 15 tutored Monterey students after their first 10 weeks of tutoring with that of 20 graduates of the ILE school at Corry Station. The written test consisted of 100 multiple-choice, network-diagram, and essay questions.

2. Results

The Monterey and ILE students averaged scores (and standard deviations) of 77.7 (11.8) and 39.7 (18.7) points, respectively. This difference is statistically significant—$t(33) = 6.90$, $p < 0.001$—with a very large effect size of 2.30. Separate analyses showed no significant or practical differences between scores of the two groups on the AFQT or in the scores used to qualify individuals for IT training.

B. Assessment 2

Assessment 2 (IWAR 1) assessed IT knowledge and skills acquired by the 12 human-tutored students who completed the 16-week course (Fletcher 2010). Three students were dropped from the course for nonacademic reasons—two involved Navy discipline and one for health reasons. The 12 graduates were compared with 12 Navy ITs with an average of 7.2 years of Fleet experience. Space and computer equipment limitations required the

assessment to be conducted over 2 weeks, in two 5-day sessions, with six tutored students and six Fleet ITs tested each week.

Troubleshooting problems used in this assessment, as well as in Assessments 4 and 5, were derived from over 20,000 trouble reports sent from Fleet ITs requesting technical assistance from the shore-based Fleet Systems Engineering Team (FSET)[3] for problems that shipboard ITs could not solve. One FSET and two senior ITs, specifically selected for their knowledge and experience, scored the performance of each team after agreeing on a single score value, ranging from 0 to 5 as described in Table 3. Participants in the Troubleshooting and Security practical exercises for Assessments 2 and 5 competed as teams of three, which is consistent with shipboard practice.

Table 3. Scoring for Troubleshooting Problems

Score	Description
5	Solved as described in the instructions or deemed equal in quality.
4	Solved, but omitted items such as documentation or full implementation.
3	Weak solution with explanation (e.g., work-around that requires later upgrading).
2	Solution that relieves the symptom but not the underlying problem.
1	Solution does not solve the problem.
0	No attempt.

1. Measures

Testing consisted of:

- 13.25 hours of hands-on troubleshooting virtual and physical IT systems that mirrored those found in the Fleet.
- 4 hours of security testing.
- 7 hours of IT system design and development.
- 4 hours of paper-and-pencil Knowledge Testing.

2. Results

The four Monterey teams scored a total of 120.63 (14.98) points on the troubleshooting problems compared with 89.12 (28.51) scored by the four Fleet IT teams. This difference is not statistically significant—$t(6) = 1.96$, $p > 0.05$—but it yielded a very

[3] As with IT, FSET may refer the shore-based Fleet Systems Engineering Team or to an individual member of the FSET team.

large effect size of 1.20, favoring the Monterey teams. In troubleshooting discipline and technique, the human-tutored teams left fewer uncorrected harmful changes after troubleshooting (8 versus 18), verified more problem solutions (97% versus 85%), and solved more problems of those attempted (95% versus 77%).

Findings were the opposite on the Security Test, where Monterey teams and Fleet teams averaged 23.75 (9.29) and 37.25 (8.77) points, respectively. These differences favoring the Fleet IT teams are not statistically significant—$t(6) = 2.11$ $p > 0.05$—but with a very large effect size of 1.30.

In one sense, the Security test results validated the approach used for the human tutoring. The principal security expert scheduled for the security section was recalled by his command shortly after the course began, requiring rapid selection of last-minute substitutes who were capable ITs, but not at the same level of expertise in either tutoring or subject matter as the original tutor. Much of the security portion of the course was taught by classroom lectures with limited opportunities for interactive work. This problem continued and was evident throughout the project and its assessments.

In System Design and Development, both groups participated in six-member, self-organized teams. The two Monterey teams successfully accomplished 32% of the objectives, and the two Fleet teams successfully accomplished 34%. The two Monterey teams averaged 42.2 (35.7) points out of 220, and the Fleet teams averaged 56.8 (11.0) points. The difference is not statistically significant—$t(2) = 0.55$, $p > 0.05$—with a small effect size of 0.31.

On the Knowledge Test, the Monterey students and Fleet ITs averaged 143.8 (17.5) and 85.75 (33.7) points, respectively. This difference is statistically significant—$t(22) = 5.30$, $p < 0.01$—with a very large effect size (2.09) favoring the Monterey students.

C. Assessment 3—April 2010

Assessment 3 compared the IT knowledge of 20 students who had completed 4 weeks of the (computerized) Digital Tutor training then available with the IT knowledge acquired by 31 students who had graduated from the ILE A school and with that of 10 Navy IT instructors at the school (Fletcher 2011).

1. Measures

All three groups completed a written knowledge test, which was administered in two 2-hour sessions.

2. Results

Knowledge Test scores averaged 126.0 (14.21) for the 20 Digital Tutor students, 62.58 (26.68) for the 31 ILE graduates, and 96.0 (34.27) for the 10 IT instructors. One-way

ANOVA found an overall statistically significant difference across the groups. All pairwise comparisons were statistically significant ($p < 0.01$). The differences were very large in all cases, 2.80 favoring Digital Tutor students over ILE graduates, 1.32 favoring Digital Tutor students over their instructors, and an effect size of 3.27 favoring instructors over the ILE graduates.

D. Assessment 4—November 2010

Assessment 4 (Fletcher 2011) assessed the IT knowledge and troubleshooting skills of four groups:

- 20 Digital Tutor students who had completed 7 weeks of then available Digital Tutor training. They performed all Assessment 4 tests.
- 20 IT of the Future (IToF) students who had graduated from a newly revised 19-week IT School consisting of classroom instruction and laboratory exercises. They performed all Assessment 4 tests.
- 17 graduates of the original ILE self-paced A School. They only received and completed the IT knowledge test.
- 10 Navy instructors, who had been trained to present IToF material. They only received and completed the IT knowledge test.

1. Measures

Testing consisted of a written knowledge test, troubleshooting exercises, packet-tracer exercises, and individual interviews conducted by a three-member review board composed of experienced Navy ITs led by a senior FSET.

2. Results: Troubleshooting

Unlike other troubleshooting assessments, which examined three-member team performance, Digital Tutor students and IToF graduates in Assessment 4 participated as individuals. Troubleshooting consisted of 15 trouble tickets presented on virtual systems. Pairs of senior ITs scored the performance of each individual after agreeing on a single score value, ranging from 0 to 5 as shown in Table 3. The scores assigned by members of each pair rarely deviated by more than one point. These trouble tickets were, as usual, derived from the database of 20,000 trouble tickets used for all assessments.

Digital Tutor students averaged 26.55 (14.09) points in these exercises, and IToF graduates averaged 5.65 (6.56) points. The variance for the IToF graduates exceeds the average because many IToF graduates could not solve the problems and scored zeros in the exercise. The difference favoring Digital Tutor students over IToF graduates is statistically significant—$t(38) = 6.02$, $p < 0.001$—with a very large effect size of 1.86.

3. Results: Packet Tracing

Understanding message traffic and the various paths taken by packets through a network is an essential IT skill. Packets are units of user and transmission data that may take different paths through a network, but ensure accurate reassembly at destination. The test consisted of 18 problems presented on virtual systems. The Digital Tutor students averaged 36.91 (16.2) unweighted points on these exercises, compared with 25.29 (15.3) for IToF graduates. This difference is statistically significant—$t(38) = 2.33$, $p < 0.05$—with a large effect size of 0.72. The Digital Tutor students averaged 30.39 (15.9) points weighted for problem difficulty compared with 15.85 (13.0) for IToF graduates. This difference is statistically significant—$t(38) = 3.17$, $p < 0.01$—with a very large effect size of 0.98.

4. Review Board Interviews

Review Board interviews were conducted with individual participants. Time permitted interviews with seven Digital Tutor students and six IToF graduates, drawn at random from the participants. Board members were not told from which group each participant was drawn. The Board rated each on a nonlinear scale, with 1 for participants who demonstrated less than 3 months of IT experience; 2 for evidence of 3 months of experience; 3 for evidence of 1–3 years of experience; 4 for evidence of 4–5 years of experience; and 5 for evidence of more than 5 years of experience. Each of the three Board members could award up to 30 points covering 6 topics, making a total of 90 points possible.

Digital Tutor students averaged 41.64 (12.93) points in the interviews, compared with 18.8 (20.90) for IToF graduates. This difference is statistically significant—$t(11) = 2.41$, $p < 0.05$—with a very large effect size of 1.25.

5. Knowledge Test

All four groups completed the knowledge test in two 2-hour sessions before beginning the practical exercises. Table 4 shows means, standard deviations, *t* scores, and effect sizes (*g*) for the four groups who took the written Knowledge Test. All differences except those between IToF graduates and instructors were statistically significant, with the Digital Tutor students statistically outscoring the other three groups with very large effect sizes.

With such large effect sizes favoring Digital Tutor and IToF (4.58 and 3.46, respectively) over ILE students, it appeared that the ILE graduates did not provide a robust comparison for the other groups. They were released from further participation in these assessments.

Table 4. Knowledge Test Means, Standard Deviations, and Number of Observations for 7-Week Digital Tutor Students, 19-Week IToF Graduates, ILE Graduates, and Navy IT Instructors.

Direction	*t*	*g*	*df*	Means (SDs)
DT > IToF	6.18[a]	1.91	38	207.90 (37.30) vs. 145.75 (25.17)
DT > ILE	14.20[a]	4.58	35	207.90 (37.30) vs. 64.53 (19.96)
DT > Instructors	3.49[b]	1.31	28	207.90 (37.30) vs. 149.30 (53.96)
IToF > ILE	10.64[a]	3.46	35	145.75 (25.17) vs. 70.00 (16.32)
IToF > Instructors	-0.25	-0.09	28	145.75 (25.17) vs. 149.30 (53.96)

[a] $p < 0.001$.

[b] $p < 0.01$.

DT = Digital Tutor

ILE = Integrated Learning Environment

IToF = Information Technology of the Future

5. Assessment 5

Assessment 5, designated as IWAR 2 by DARPA, was the capstone assessment for DARPA's Digital Tutor program. It provided summative evaluation of the first complete 16-week version of the Digital Tutor. Like Assessment 2 (IWAR 1), it was conducted in two successive 5-day sessions. Eighteen participants (6 participants from each of 3 groups) were examined in each session. As in Assessment 2, the scheduling was determined by the availability of computer systems and floor space.

Assessment 5 addressed four basic questions:

- Did the Digital Tutor program provide its graduates with relevant Fleet-required IT skills and knowledge?
- Were the skills and knowledge acquired by Digital Tutor graduates superior to those of experienced Fleet ITs?
- Were the skills and knowledge acquired by Digital Tutor graduates superior to those provided by classroom instruction?
- Did the Digital Tutor program capture in digital form the human tutoring effectiveness found in Assessment 2?

The first two questions concern Navy operational needs. The last two questions concern the use of computers in education and training. Overall, the goal was to determine if acquisition of relevant technical expertise could be substantively accelerated.

A. Participants

Assessment 5 participants were drawn from three groups:

- 12 graduates of the 16-week Digital Tutor course. Six individuals were drawn from each of two 20-student classes that completed Digital Tutor courses a week apart.
- 12 graduates of the 35-week ITTC course. These students were the first 12 who had passed a certification exam and graduated out of a full class of 30 students at about the same time as the Digital Tutor students.
- 12 Fleet ITs with 4–15 years (average of 9.6 years) experience as Fleet ITs. As in Assessment 2, individual ITs chosen were identified by their commands as especially capable Fleet ITs at this level of experience—they were the shipboard "go-to" ITs.

B. Support Teams

As in Assessments 2 and 4, a White Team made up of senior Navy ITs and Navy FSETs performed essential services in conducting this assessment. Members of the White Team interviewed participants in all review boards, organized participants for practical exercises, ensured that exercise protocols were observed, scored all performance in the practical exercises, and coordinated activities with the Technical Support Team.

Again as in Assessments 2 and 4, a Technical Support Team performed essential services for this assessment. The team prepared the IT systems for practical problems (troubleshooting, security, and system build), corrected any system problems that arose, and ensured that the systems used for testing were restored and restarted as needed. As before, there were no significant technical disruptions during either of the two 5-day sessions in Assessment 5.

C. Facilities

Participants were tested in three separate classrooms provided by the San Diego Naval Base, which hosted the assessment. Each classroom contained three IT systems—one physical system, with a full complement of servers and software (such as Microsoft Server, Windows XP, Microsoft Exchange, CISCO routers and switches, and the Navy's COMPOSE overlay to Windows) and two identical virtual systems running on virtualized hardware with the same software. The systems were designed to mirror those typically found on Navy vessels and duty stations. The software was not simulated. Participants interacted directly with software used in the Fleet. The availability of three systems in each room allowed one to be prepared for the next exercise while the two three-member teams used the others. Limited access to the Internet was provided as needed for specific problems. Wider Internet access was not available.

More time was required to prepare the physical systems than the virtual systems for troubleshooting problems. Having finished a problem on the physical system, a team would typically move to the next problem on a virtual system, freeing the physical system to be configured for the next problem. As a result, Assessments 2 and 5 presented more virtual than physical system troubleshooting problems.

Classrooms were instrumented with video cameras and microphones. Participant activity was available live and time stamped for later review using tools developed for further Digital Tutor development and training.

D. Schedule

Each of the two Assessment 5 sessions tested three six-member cohorts (18 participants in each week's session) chosen at random from the three participating groups

(Digital Tutor, ITTC, and Fleet). Table 5 shows the schedule followed for each week. The assessment consisted of five activities: a Written Knowledge Test, which was administered before the week began; interviews with a Review Board conducted throughout the first day; and practical exercises that consisted of 2–1/2 days of Trouble Ticket Troubleshooting, a 1/2-day Security exercise, and 1 day of System Design and Development.

Table 5. Schedule for Each of Two 5-Day Assessment 5 Sessions

Monday	Tuesday	Wednesday	Thursday	Friday
Review Board Interviews with IWAR participants.	Practical Troubleshooting exercises (six teams of three individuals—two teams at the same time in each room)	Practical Trouble-shooting exercises continued	Practical Troubleshooting exercises continued for a half day. A half day of Security exercises with the same three-member teams.	System Design and Development exercise (one six-member, self-organized team in each room)

E. Measures

1. Practical Exercises

As in IWAR 1 and as shown in Table 5, practical exercises were conducted each week: Troubleshooting by three-member teams over a period of 2–1/2 days; a Security exercise performed by the same teams for about 4 hours; and a System Design and Development exercise conducted for about 6 hours by all members of each group (Digital Tutor, ITTC, and Fleet) participating together in self-organized teams. Scoring for these exercises was provided by three White Team members, generally headed by an FSET. The White Teams rotated among the participants so that they scored participant teams from each group an equal number of times.

2. Troubleshooting Exercise

Troubleshooting problems were again drawn from the 20,000 trouble ticket database. The troubleshooting items provided the core assessment for Assessment 5. Differences in team performance were assessed based on four data points (four teams for each group, two per session).

Troubleshooting problems were presented as they are at Navy duty stations—as Trouble Tickets. Figure 3 shows a sample trouble ticket. Participating teams were required to solve the problem, describe the solution, and document the steps they had taken to correct it. Figure 4 shows the setup instructions for the Figure 3 problem.

TROUBLE TICKET

Day 3, July 29, Time (Start/End) _____ **Team:** _____

Problem Symptom: Lt Sulu complains he is not receiving email

Problem Solution:

Key Solution Steps:

Figure 3. Example Trouble Ticket Presented to IWAR 2 Participants

Scenario	TS-SV-GC-30	
Concept Tested	IP configuration and Troubleshooting	
General Description		
Add a static route for the 172.16.0.0/30 network to point to 172.16.1.254		
Injection Script		
1. Log on to EX01 as the proctor admin account (proctor) 2. Open a command prompt 3. In the command prompt, enter the following command: route add 172.16.0.0 mask 255.255.252.0 172.16.1.254 –p 4. To test, try to ping WKS01. If all is configured correctly, this will fail.		
Problem Symptoms		
• LT Sulu complains he is not receiving email.		
Preferred Solution(s)		
• Delete the static route on EX01		
Impact:	If Clients cannot connect to EX01, they will not be able to send or receive email	
Impact Rating: 7-high		**Difficulty:** Very Hard
Time to Resolve: 30 Minutes		

Figure 4. Example Troubleshooting Problem Description and Setup Instructions

Of the 210 troubleshooting problems developed for IWAR 2, 182 were scheduled for initial presentation, with the remaining 28 held in reserve for use as needed. A different set

of problems was presented in each week's session—92 in the first week's session and 90 in the second week's session. On each day, the same problems were presented to the Digital Tutor, ITTC, and Fleet teams in the same order. Fifteen minutes after a problem was presented, a team was free to move to the next problem when it chose to do so. The teams used their own notes along with IT reference materials on compact discs, just as they would at Navy duty stations.

Scoring was determined by consensus among the three White Team members, who awarded 0–5 points for each problem. These points were anchored as shown in Table 3.

3. Security Exercise

The Security exercise was performed by participants in their three-member troubleshooting teams. Each team was presented with a virtual system containing the security violations that it was to identify and correct. The exercise covered seven different areas of security involving problems such as those arising from viral software, compromised passwords, and unauthorized displays. The teams were assisted by documentation and patches provided for the exercise. The White Team awarded 0–5 points for each security violation found, depending on the difficulties it presented, its severity, and the team's success in identifying and correcting it.

4. System Design and Development

This exercise was performed by assembling all six of the week's participants from each Digital Tutor, ITTC, or Fleet group into a single self-organized team. The teams were given hardware, including servers, routers, cables, and switches; operating system and application software (e.g., Windows XP and Microsoft Office); and a block of 128 IP addresses. These materials were sufficient to design and implement a system specified by the critical and secondary objectives of the exercise. Figure 5 shows example objectives of both sorts and their scoring. The task was to assemble an IT system that correctly met as many of the objectives as possible.

Teams were awarded 0–5 points for each objective. Different objectives were presented each week. Twenty-four objectives (5 Critical and 19 Secondary) were required for systems developed in the first week's session, making a total of 120 points to be awarded. Twenty-one objectives (5 Critical and 16 Secondary) were required for systems developed in the second week's session, making a total of 105 points possible.

Example Critical Objectives	Scoring
Establish a fault-tolerant Windows domain called SOTF.navy.mil to support the Operation.	0—Domain not created 3—Domain created and working correctly, but not fault tolerant 5—Domain created correctly and is fault tolerant
Install and configure an Exchange server for SOTF.navy.mil.	0—Exchange not installed 3—Exchange installed but configured incorrectly 5—Exchange installed and configured correctly
Establish Internet access for all internal client machines and servers.	0—Design not functional or complete 1—Only one system with Internet access 3—Some systems with Internet access, some without 5—All systems with Internet access
Example Secondary Objectives	**Scoring**
All client machine TCP/IP settings must be configured automatically.	0—Design not functional or complete 1—Clients using APIPA addressing 3—Some (not all) clients have DHCP IP addresses 5—All clients have DHCP IP addresses
Create domain user accounts for three inbound junior ITs: - ITSR Bert Dillard - ITSR Roscoe Burr - ITSA Randall Durham	0—Accounts not created 1—One account created 4—Two accounts created 5—All accounts created
DNS servers must be able to resolve internal names to IP addresses and IP addresses to names.	0—DNS not functional 3—Forward lookup configured properly, reverse hookup not functional 5—Forward and reverse lookups configured correctly

APIPA = Automatic Private IP Addressing
DHCP = Dynamic Host Configuration Protocol
DNS = Domain Name System
IP = Internet Protocol
TCP =Transmission Control Protocol

Figure 5. Sample Objectives and Scoring for System Design and Development

5. Review Board Interviews

Review boards are commonly used by the military for awards, certifications, and promotions. Assessment 5 included 20–30 minute interviews by a three-member White Team review board. Three boards operated in parallel and interviewed each participant in random order on the first day of each week's session. As in Assessment 4, each board was led by an FSET, assisted by two senior Navy ITs who had been identified and selected for their IT knowledge and expertise. The examinations were partly blind in that members of the board did not know from which group, Digital Tutor or ITTC, interviewees, who were all of about the same age and military rank, were drawn. Fleet participants were readily identified by their age and more senior rank.

Each participant was examined on a 0–5 scale with regard to six core topics: Networking, Workstations, Domain Controllers, Domain Name System, Disk

Management, and Exchange. The interview began with a common question, but after that was free to proceed as the board chose. Participants who demonstrated effectively no knowledge of a topic were assigned 0 points; any participant who demonstrated as much knowledge as that possessed by members of the board was awarded 5 points. One participant, a Digital Tutor student with no prior IT experience, received three 5s. Each of the three board members scored each participant so that a total of 90 points could be awarded for the six topics.

6. IT Knowledge Test

The Knowledge Test consisted of 3 parts totaling 272 1- and 2-point items worth 351 points in all. As in the earlier Knowledge Tests, about half the items were factual and half were conceptual. Most test items were short-answer questions, but multiple-choice and paragraph-length items were also included. As in all knowledge testing for this work, IDA professional staff, IDA technical staff, and the Navy Network Warfare Command had previously vetted all items for their central relevance to Navy IT duty assignments.

All participants completed the Knowledge Test before beginning other elements of Assessment 5. All three parts were administered under closed-book, closed-notes conditions. The test was intended to be sufficiently difficult to avoid ceiling or floor effects. Participants were given 75 minutes to finish Part 1, 75 minutes to finish Part 2, and 90 minutes to finish Part 3. Nearly all participants finished each part in less than an hour. All finished the test in the time available.

6. Results

A. Troubleshooting Exercises

Because the trouble tickets were drawn from problems that had been submitted from the Fleet for shore-based FSET assistance, their difficulty could be estimated from e-mail traffic. Problems were assigned one of five levels of difficulty, ranging from very easy to very hard. Table 6 gives brief descriptions of these difficulty levels.

Table 6. Rating of Troubleshooting Problems at Five Difficulty Levels

Level of Difficulty	Description
1-Very Easy	Solved by the average power user.
2-Easy	Solved by the average IT technician.
3-Average	Solved by an average network administrator.
4-Hard	Solved only by experienced network administrators.
5-Very Hard	Solved only by seasoned IT professionals.

Data reported here are based on efforts to solve 140 troubleshooting problems, the maximum attempted in the exercise. The troubleshooting assessment addressed three issues: quality of problem solution, unnecessary steps taken to solve the problem, and harmful changes made during troubleshooting and left in the system.

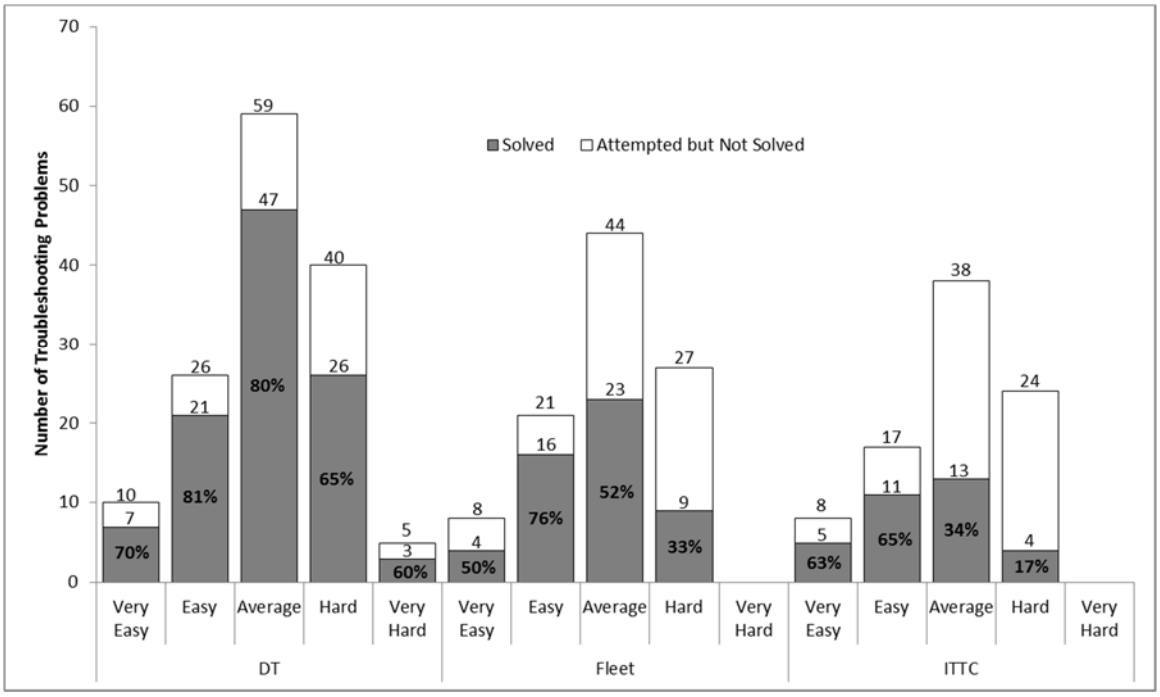
Solution quality was determined by discussion and consensus among the three White Team examiners. Scoring was performed as if each team were given a 140-item test, with each item worth 0–5 points. The sum of all points provided the final score. As in Assessments 2 and 4, ratings by the three examiners rarely differed by more than one point.

Harmful changes are especially pernicious. About 20% of Fleet trouble tickets arise from problems introduced by IT technicians themselves. The White Team tallied harmful changes left uncorrected or unrestored during troubleshooting and scored their severity based on the skill level needed to find and correct them. As described in Table 6, these scores ranged from 1 (least severe) to 5 (most severe).

Unnecessary steps provide an indirect measure of proficiency but a direct measure of efficiency in problem-solving. Their cost may be monetary at dockside, but they can be operationally critical during naval maneuvers and armed engagements. Unnecessary step scores were tallies of these steps taken during a solution attempt. The maximum score for a single problem was held to 5, even if more than five steps were taken—problem approaches described by the Navy as “Easter egging” are not uncommon.

1. Troubleshooting Scores

A problem was considered solved if the team working on it received a score of 4 or 5. Figure 6 shows troubleshooting problems attempted and solved. The figure arranges problems by difficulty. Digital Tutor teams attempted a total of 140 problems and successfully solved 104 of them (74%), with an average score of 3.78 (1.91). Fleet teams attempted 100 problems and successfully solved 52 (52%) of them, with an average score of 2.00 (2.26). ITTC teams attempted 87 problems and successfully solved 33 (38%) of them, with an average score of 1.41 (2.09).



DT = Digital Tutor
ITTC = Information Technology Training Continuum

Figure 6. Troubleshooting Problems Attempted and Solved by Difficulty Level

One-way ANOVA found the differences across the three groups to be statistically significant—$F(2, 9) = 52.03$, $p < .0001$. Table 7 shows means and standard deviations of total points awarded to the four teams from each group, along with *t*-scores, and effect sizes for group comparisons. All pairwise differences are statistically significant. The Digital Tutor teams outscored both the Fleet and ITTC teams with very large effect sizes of 4.19 and 7.98, respectively. The Fleet teams outscored the ITTC teams with a very large effect size of 1.33.

Table 7. Pairwise Comparisons for Mean Total Scores of IWAR 2 Troubleshooting Teams

Direction	*t*	*g*	*df*	Means (SDs)
DT > Fleet	6.81[a]	4.19	6	132.38 (8.29) vs. 70.00 (16.32)
DT > ITTC	12.98[a]	7.98	6	132.38(8.29) vs. 49.50 (9.72)
ITTC > Fleet	–2.16	–1.33	6	49.50 (9.72) vs. 70.00 (16.32)

[a] $p < 0.001$

DT = Digital Tutor

ITTC = Information Technology Training Continuum

Figure 6 also shows that the Digital Tutor teams attempted and correctly solved more difficult problems than did the Fleet or the ITTC teams and that they solved larger proportions of these problems. Only the Digital Tutor teams attempted the "very hard" problems, solving three of the five problems in this category. Digital Tutor teams correctly solved 65% of the "hard" problems they attempted, compared with 33% for Fleet teams and 17% for ITTC teams. Similar results were obtained for problems of average, easy, and very easy difficulty, with the Digital Tutor teams solving larger proportions of problems at each level of difficulty.

Overall, Digital Tutor teams attempted and solved more troubleshooting problems with a higher probability of success than either Fleet or ITTC teams, and they were more likely to attempt and correctly solve very difficult problems.

2. Harmful Errors

As in all IT training, a goal of the Digital Tutor was to reduce the frequency of harmful errors—changes made to troubleshoot a system and left behind when the troubleshooting is done. The number of errors, shown in Table 8, was based on the number of problems attempted, which differed among the three groups. To compensate for these differences, the number of harmful errors was divided by the number of problems attempted for each team, yielding a rate of harmful actions, rather than a sum. The average rate of harmful errors per attempt by a Digital Tutor team was less than one-third that of either the Fleet or ITTC teams: 0.30 versus 0.98 and 1.01, respectively. ANOVA found the difference in error rates among the three groups was not statistically significant—$F(2, 9) = 4.06$, $p > 0.05$. Table 8 shows *t*-scores for these differences only as a matter of interest. The effect sizes were very large for comparisons between Digital Tutor and either Fleet or ITTC: 1.85 and 1.63, respectively. Rates of harmful errors left behind by ITTC teams were slightly greater than those for Fleet teams—a difference that is not statistically significant and with a negligible effect size of 0.06.

Table 8. Results from Pairwise Comparisons of IWAR 2 Harmful Action Rates

Direction	*t*	*g*	*df*	Means (SDs)
DT > Fleet	–3.02[a]	–1.85	6	0.30 (0.13) vs. 0.98 (0.43)
DT > ITTC	–2.65[a]	–1.63	6	0.30 (0.13) vs. 1.01 (0.52)
ITTC > Fleet	0.09	0.06	6	1.01 (0.52) vs. 0.98 (0.43)

[a] $p < 0.05$
DT = Digital Tutor
ITTC = Information Technology Training Continuum

The severity of harmful changes based on the level of IT ability needed to find and correct the change (per Table 6) was assessed. The percent of solution attempts with a severely harmful action (rated 4 or 5) was 1.5% for the Digital Tutor teams compared to 7.1% for the Fleet teams and 12.5% for the ITTC teams.

3. Unnecessary Solution Steps

Efficiency in IT troubleshooting and problem-solving is indicated by the number of unnecessary steps taken during solution attempts. The number of unnecessary steps was divided by the number of problems attempted for each team, yielding a rate of unnecessary

actions. The rate was 0.48 for the Digital Tutor teams, which was less than half that for Fleet (1.13) and ITTC (1.40) teams. ANOVA for the comparisons in Table 9 found the difference among the three groups to be statistically significant—$F(2,9) = 7.46$, $p < 0.05$. The rate of unnecessary steps taken by Digital Tutor teams was significantly lower ($p < 0.05$) than those for either Fleet or ITTC teams. Effect sizes comparing Tutor with Fleet and ITTC teams were very large –2.26 and –2.10, respectively. The difference in rate of unnecessary steps taken by ITTC versus Fleet teams was not statistically significant, but with a large effect size of 0.60.

Table 9. Results from Pairwise Comparisons of IWAR 2 Unnecessary Step Rates

Direction	*t*	*g*	*df*	Means (SDs)
DT > Fleet	–3.67[a]	–2.26	6	0.48 (0.24) vs 1.13 (0.26)
DT > ITTC	–3.37 [a]	–2.07	6	0.48 (0.24) vs 1.40 (0.49)
ITTC > Fleet	0.97	0.60	6	1.40 (0.49) vs 1.13 (0.26)

[a] $p < 0.05$
DT = Digital Tutor
ITTC = Information Technology Training Continuum

In sum, the frequency with which Digital Tutor teams left harmful changes in the system or took unnecessary steps in troubleshooting was lower than that observed for Fleet or ITTC teams. Moreover, the severity of harmful changes left behind by Digital Tutor teams was found to be about half that of either the Fleet or ITTC teams. The monetary and operational consequences of these differences have not been quantified, but they may be considerable.

4. Review Board Interviews

All participants were interviewed by the Review Board. As in Assessment 4, performance in the Review Board interview was rated on a 5-point scale (1 for no knowledge of the topic and 5 for as much knowledge of the topic as the Board members).

ANOVA for the ratings reported in Table 10, found the difference between groups to be statistically significant—$F(2,33) = 4.71$, $p < 0.05$. The differences between Digital Tutor and the Fleet and ITTC participants are both statistically significant ($p < 0.05$). Differences between ITTC and Fleet participants were not significant. Effect sizes were very large for comparisons of Digital Tutor with Fleet participants (1.07) and with ITTC participants (0.89). ITTC participants were rated higher than Fleet participants, with a small effect size (0.32). This difference was not statistically significant.

Table 10. Results from Pairwise Contrasts from IWAR 2 Review Board Ratings

Direction	*t*	*g*	*df*	Means (SDs)
DT > Fleet	2.71[a]	1.07	22	2.65 (0.67) vs. 1.93 (0.63)
DT > ITTC	2.29[a]	0.89	22	2.65 (0.67) vs. 2.11 (0.47)
ITTC > Fleet	0.79	0.32	22	2.11 (0.47) vs. 1.93 (0.63)

[a] $p < 0.05$

DT = Digital Tutor

ITTC = Information Technology Training Continuum

5. Security Exercise

As in Assessment 2, the security exercise findings, shown in Table 11, were opposite those obtained in other IWAR 2 tests. Participants in this exercise were organized into the same three-person teams used in the troubleshooting exercise. Six teams were tested each week. Teams could score a total of 87 points in the first week's session and 85 total points in the second week's session. Each team's score was expressed as a percentage of total points possible. A one-way ANOVA of the groups indicated no statistically significant difference among their means: $F(2, 9) = 2.18$, $p > .10$. Table 11 shows *t*-scores for these differences only as a matter of interest. The Fleet teams outperformed the other teams, with very large effect sizes of 1.03 in the case of Digital Tutor teams and 2.03 in the case of ITTC teams. Scores of the Digital Tutor and ITTC teams on the exercise show a negligible effect size of 0.03.

Table 11. Results from Pairwise Contrasts on IWAR 2 Security Exercise Scores

Direction	*t*	*g*	*df*	Means (SDs)
DT > Fleet	–1.58	–0.97	6	44.4 (29.4) vs. 69.2 (11.0)
DT > ITTC	–0.04	–0.03	6	44.4 (29.4) vs. 45.1 (10.9)
ITTC > Fleet	–3.11[a]	–1.91	6	45.1 (10.9) vs. 69.2 (11.0)

[a] $p < 0.05$

DT = Digital Tutor

ITTC = Information Technology Training Continuum

6. Network Design and Development

As in Assessment 2, six-person, self-organized teams performed the network and design exercise. Three teams were tested in each week's session. The White Team rated performance on each critical and secondary design objective from 0 to 5 using the scale guidelines shown earlier in Figure 5. There were 24 objectives (5 critical and 19 secondary) in the first week's session and 21 objectives (5 critical and 16 secondary) in the second week's session. Table 12 shows the mean total scores for critical, secondary, and overall objectives awarded to the three groups (two teams for each of the three groups).

Effect sizes for mean total score over all objectives were small for Digital Tutor versus Fleet (0.33) and large for the Digital Tutor versus ITTC (0.63). ANOVA indicated no significant difference across the three groups, $F(2,3) = 2.23$, $p > .10$.

Digital Tutor training was of more advantage in meeting critical objectives than secondary objectives. ANOVA for critical objectives found significant differences across groups—$F(2,3) = 9.55$, $p < .05$. Effect sizes for Digital Tutor versus Fleet were very large (0.97) and moderate (0.58) for Digital Tutor versus ITTC, but all pairwise comparisons were non-significant.

Table 12. Results from Pairwise Comparisons of Scores on Design and Development

Critical Objectives	*t*	*g*	*df*	Means (SDs)
DT > Fleet	1.70	0.97	2	4.80 (0.63) vs. 1.90 (2.33)
DT > ITTC	1.02	0.58	2	4.80 (0.63) vs. 3.10 (2.28)
ITTC > Fleet	0.53	0.30	2	3.10 (2.28) vs. 1.90 (2.33)
Secondary Objectives	***t***	***g***	***df***	**Means (SDs)**
DT > Fleet	0.32	0.18	2	3.63 (1.99) vs. 2.97 (2.18)
DT > ITTC	1.20	0.68	2	3.63 (1.99) vs. 1.31 (1.89)
ITTC > Fleet	–0.82	–0.47	2	1.31 (1.89) vs. 2.97 (2.13)
All Objectives	***t***	***g***	***df***	**Means (SDs)**
DT > Fleet	0.57	0.33	2	3.89 (1.84) vs. 2.73 (2.20)
DT > ITTC	1.10	0.63	2	3.89 (1.84) vs. 1.71 (2.10)
ITTC > Fleet	–0.47	–0.27	2	1.71 (2.10) vs. 2.73 (2.20)

DT = Digital Tutor

ITTC = Information Technology Training Continuum

7. Knowledge Test

There were 272 items worth 349 points across all three parts of the Knowledge Test. As shown in Table 13, ANOVA found that the overall difference between groups was significant: $F(2,33) = 61.59$, $p < 0.001$. The *t*-scores and effect sizes found all pairwise differences to be statistically significant. Effect sizes were all very large or large, with 3.11 for Digital Tutor scores compared with Fleet scores, 3.54 for Digital Tutor scores compared with ITTC scores, and 0.77 for ITTC compared with Fleet participants.

Table 13. Results from Pairwise Comparisons of Scores on the IWAR 2 Knowledge Test

Test	*t*	*g*	*df*	Means (SDs)
DT > Fleet	7.88[a]	3.11	22	271.50 (33.75) vs. 132.08 (51.16)
DT > ITTC	8.97[a]	3.54	22	271.50 (33.75) vs. 164.04 (24.12)
ITTC > Fleet	1.96	0.77	22	164.04 (24.12) vs. 132.08 (51.16)

[a] $p < 0.01$

DT = Digital Tutor

ITTC = Information Technology Training Continuum

B. Assessment 5 Summary

Fleet ITs were probably at a disadvantage on the Knowledge Test, which assesses up-to-date knowledge at considerable breadth and depth. Although the Fleet ITs receive follow-on certification training, sustainment training, and technical updates, this training is likely to vary in content, quality, and currency. Also, Fleet duties can limit the range, of experience—ITs on large ships must often specialize in a particular area. The ITTC graduates, who had just completed 35 weeks of up-to-date IT training, were far less subject to these disadvantages.

AFQT scores accounted for about 37% of variance ($r = 0.61$) in the Knowledge Test scores of the Digital Tutor graduates and 59% of variance ($r = 0.77$) in Knowledge Test scores of ITTC graduates. It appears that general ability as measured by the AFQT helped both groups of participants answer Knowledge Test questions, but to a substantially greater degree for ITTC graduates than for Digital Tutor graduates.

Even though knowledge is not the core outcome in training, it is not inconsequential. Scores on the Knowledge Test accounted for about 41% ($r = 0.64$) of troubleshooting variance among Digital Tutor students in Assessment 4, where sailors participated as individuals rather than as part of teams.

Assessment 5 findings, summarized in Table 14, suggest at least three patterns that were repeated across the different performance measures:

- With the exception of the Security exercise, Digital Tutor participants outperformed both the Fleet and ITTC teams.
- Differences between Fleet and ITTC participants were generally smaller and neither consistently positive nor negative.

- On the Troubleshooting exercises, which closely resemble Navy duty station work, Digital Tutor students substantially outscored Fleet ITs and ITTCs graduates, with higher ratings at every difficulty level, less harm to the system, and fewer unnecessary steps.

Table 14. Summary of Results from Assessment 5 (IWAR 2)

Performance Measure	Direction	Significance[a]	Effect Size[b]
DT versus Fleet			
Troubleshooting Total Score	DT > Fleet	<.0001	4.19
PS Harmful Actions	DT > Fleet	<.0001	–1.85
PS Unnecessary Steps	DT > Fleet	<.0001	–2.26
Review Board	DT > Fleet	<0.01	1.07
Security Exercise	DT > Fleet	N.S.	–0.97
Network Design and Development	DT > Fleet	N.S.	0.33
Knowledge Test Total Score	DT > Fleet	< 0.0001	3.11
DT versus ITTC			
Troubleshooting Total Score	DT > ITTC	<.0001	7.98
PS Harmful Actions	DT > ITTC	<0.01	–1.63
PS Unnecessary Steps	DT > ITTC	<.0001	–2.10
Review Board	DT > ITTC	<0.05	0.89
Security Exercise	DT > ITTC	N.S.	–0.03
Network Design and Development	DT > ITTC	N.S.	0.63
Knowledge Test Total Score	DT > ITTC	<0.0001	3.54
ITTC versus Fleet			
Troubleshooting Total Score	ITTC > Fleet	N.S.	–1.33
PS Harmful Actions	ITTC > Fleet	N.S.	0.06
PS Unnecessary Steps	ITTC > Fleet	N.S.	0.60
Review Board	ITTC > Fleet	N.S.	0.32
Security Exercise	ITTC > Fleet	<0.05	–1.92
Network Design and Development	ITTC > Fleet	N.S.	–0.27
Knowledge Test Total Score	ITTC > Fleet	N.S.	0.77

[a] Two-tailed probability from *t*-test for independent means.

[b] Negative Effect Sizes are opposite of the indicated direction

DT = Digital Tutor

ITTC = Information Technology Training Continuum

The exception to this pattern is the Security exercise where Fleet teams outperformed both Digital Tutor and ITTC teams with very large effect sizes of 0.97 and 1.92, respectively. This finding seems notable even though only the ITTC and Fleet difference was statistically significant. As noted earlier, appropriate human tutors were not available for the Security portion of the Digital Tutor design and development. This finding emphasizes the importance of selecting individuals who are expert in both the subject matter and one-on-one tutoring when designing instruction modeled on human tutoring.

Overall, if Digital Tutor graduates had simply matched Fleet IT performance in the practical exercises, the goals of the program to accelerate acquisition of expertise would have been met. Instead, the Digital Tutor students outscored the Fleet participants by substantial margins. From a monetary standpoint it is notable that they also outscored ITTC graduates, who spent more than twice the time in training as Digital Tutor students.

C. Additional Analyses

Three additional analyses were performed to determine what data from Assessment 5 might have to say about digital tutoring in general. They concern its effectiveness compared with human tutoring, equity in providing learning, and dependence on reading ability.

1. Is Digital Tutoring as Effective as Human Tutoring?

Early work reported by Meehl (1954), Goldberg (1970), Dawes (1971), and others compared the effectiveness of "clinical" with "statistical" decision-making. In these comparisons, researchers used regression analysis to capture as accurately as possible processes used to determine graduate school admissions and medical diagnoses. In Dawes' terms, these techniques were expected to provide a floor to be further refined by human judgment. Instead, and surprisingly, they turned out to be a ceiling. In nearly all cases, decisions based on statistical processes were found to be superior to those made by the humans on whose processes the statistical processes were based. These findings suggest that the algorithmic processes used in digital tutoring may be superior to the human tutoring on which they were based.

The Troubleshooting data did not permit assessment of this possibility. The Knowledge Test, however, provides an initial, tentative test. The human-tutored participants in Assessment 2 and the digitally tutored participants in Assessment 5 completed parts 1 and 2 of the Knowledge Test. Table 15 shows the results. They indicate statistically and practically superior performance by the digitally tutored participants on Part 1 of the test, but not on Part 2, and not when scores from both parts are combined.

Table 15. Part 1, Part 2, and Combined Knowledge Test Scores for Human versus Digital Tutoring (Assuming Human > Digital Tutoring)

Direction	*t*	*g*	*df*	Means (SDs)
Human > DT Knowledge Test Part 1	–3.10[a]	–1.22	22	80.50 (10.98) vs. 95.54 (12.75)
Human > DT Knowledge Test Part 2	0.19	0.07	22	63.33 (8.98) vs. 62.23 (17.96)
Human > DT Knowledge Test Parts 1 and 2	–1.54	–0.61	22	143.83 (17.52) vs. 157.77 (25.97)

[a] $p < 0.01$.

These findings are not conclusive, but like Meehl's regression equations, they suggest that the issue of algorithmic versus human tutoring deserves further research and investigation.

2. Is Digital Tutoring Equally Beneficial for all Students?

Does the rising tide of an instructional opportunity equally raise all learners? A measure that provides both a quantitative measure and comprehensive consideration of all learners is the Gini coefficient.

Gini coefficients are commonly used in econometrics to assess distribution of income. They might similarly be used to assess equitable distribution of learning—as early suggested by Jamison et al. (1976). The idea behind a Gini coefficient is that in a perfectly equitable system the total value accumulated by a given percent of some population should equal the same percent of total value available (Atkinson 1970; Shalit 1985). For instance, if the learning (however it is measured) accumulated by the bottom 20% of learners equals 20% of the total learning accumulated by all learners, they will have received their fair share. If they receive less, the area between the curve (a Lorenz curve) plotted for learners under the 45-degree straight line will increase, indicating increased inequality.

One Lorenz curve may cross another in a comparison, but a requirement for applying this analysis is that the Lorenz curve be concave. For instance the lower 20% of learners

must receive 20% or less of the total learning as determined by whatever measure of learning is used.

A Gini coefficient is keyed to the area between the two curves. It is expressed as the ratio of the area between the 45-degree line and the Lorenz curve over the total area under the 45-degree line. Perfect equality has a Gini coefficient of 0.00; perfect inequality has a Gini coefficient of 1.00—the larger the Gini coefficient, the greater the inequality.

Assessment 4 lends itself to such analysis for digital tutoring because it provides troubleshooting scores for individual learners from two different instructional treatments—7 weeks of Digital Tutor instruction and 19 weeks of classroom instruction for IToF graduates. There were 20 learners in each group.

In discrete terms, Gini coefficients for each of these instructional treatments may be calculated as:

$$\mathrm{G} = \frac{1}{n}\left(n + 1 - 2\left(\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{n=1}^{n} y_i}\right)\right)$$

where:

G = Gini coefficient,

y_i = Score on Assessment 4 Troubleshooting for individual *i*.

Figure 7 shows Lorenz curves for Assessment 4 troubleshooting by Digital Tutor and IToF learners. Gini coefficients, with correction for small sample sizes, were 0.47 for the Digital Tutor graduates and 0.69 for the ITTC graduates, with a difference between the two of 0.22. This finding raises a question of interpretation: How substantial or significant is this difference in Gini coefficients?

The sampling distribution for the difference between Gini coefficients is not known, so the sampling distribution was estimated using bootstrapping methods. Fifty thousand random samples of size 20 were drawn from the two samples, and differences were calculated. The upper and lower limits of the 99.99% confidence interval for the difference (0.53 and 0.10, respectively) did not include zero, implying that the result of no differences is extremely unlikely and suggesting that the difference between the obtained Gini coefficients may be significant at the .0001 level.

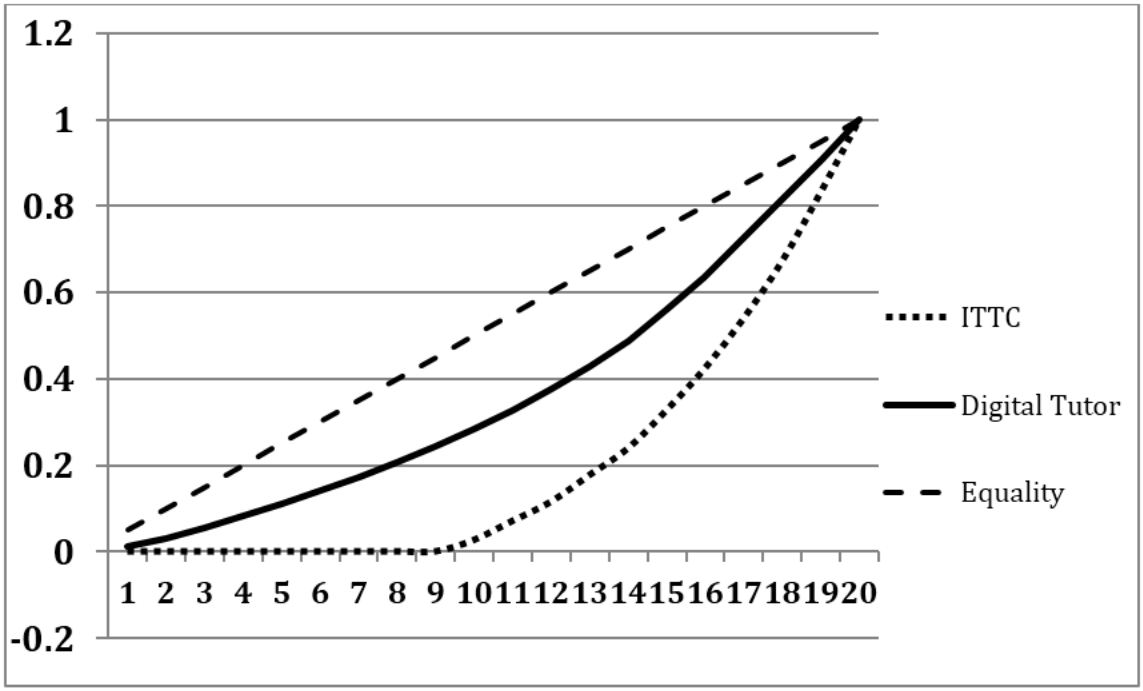


Figure 7. Lorenz Curves for Digital Tutor and ITTC Knowledge Test Scores

This difference in equality between digital tutoring and classroom instruction echoes that found earlier by Jamison et al. (1976). It suggests greater learning equality for digital over classroom instruction. This finding is not conclusive, with its statistical significance based on a Monte Carlo estimate of the standard error of estimate, but it corroborates the earlier results and suggests, as do the findings from digital compared with human tutoring, value in favor of digital instruction.

3. Does Digital Tutoring Effectiveness Depend on Reading Ability?

Because much information and instruction in the Digital Tutor is conducted through reading, the extent to which reading ability affected learning was of interest. The Armed Forces Vocational Ability Battery (ASVAB) and the Gates-MacGinitie Reading Test (GMRT) were used to assess this possibility. ASVAB scores were available for all participants, and the GMRT was administered to all Digital Tutor and ITTC participants before they began training.

Four scores from the ASVAB concern reading ability to some degree: AFQT, Paragraph Comprehension (PC), Verbal Comprehension (VC), and Word Knowledge (WK). Scores included from the GMRT were Extended Scale Scores (ESS), which provide equal interval units normalized from GMRT raw scores to a mean of 500 (MacGintie et al. 2007). Three GMRT scores were available: Total Reading (T-ESS), Reading Comprehension (C-ESS), and Reading Vocabulary (V-ESS). ITTC participants in

Assessment 5 scored higher than Digital Tutor participants on all seven of these measures, but none of the differences were statistically significant.

Table 16 shows correlations between these ASVAB and GMRT scores and Knowledge Test Scores from Assessment 5. ASVAB AFQT, Paragraph Comprehension, Verbal Comprehension, and Word Knowledge scores all accounted for substantially more variance in Knowledge Test scores of ITTC than they did for Digital Tutor graduates. GMRT Total reading accounted for about 25% of the variance in ITTC Knowledge Test scores, but none of the variance in Digital Tutor Knowledge Test scores. GMRT Reading Vocabulary accounted for about 35% of the variance in ITTC Knowledge Test scores, and about 11% of the variance in Digital Tutor Knowledge Test scores. Although small, the negative correlation between GMRT comprehension and the Knowledge Test scores of Digital Tutor graduates remains an opportunity for speculation.

Table 16. Correlations of ASVAB and GMRT with Knowledge Test

	Knowledge Test	
ASVAB	**Digital Tutor**	**ITTC**
AFQT (Armed Forces Qualification Test)	0.61[a]	0.77[b]
PC (Paragraph Comprehension)	0.36	0.64[a]
VC (Verbal Comprehension)	0.43	0.75[b]
WK (Word Knowledge)	0.43	0.55
GMRT	**Digital Tutor**	**ITTC**
T-ESS (Total Reading)	0.00	0.50
C-ESS (Reading Comprehension)	–0.32	0.27
V-ESS (Reading Vocabulary)	0.33	0.59[a]

[a]$p < .05$; [b]$p < .01$
AFVAB = Armed Forces Vocational Ability Battery
GMRT = Gates-MacGinitie Reading Test
ITTC = Information Technology Training Continuum

Overall, ASVAB and GMRT results suggest that superior IT knowledge more than reading ability was responsible for the higher Knowledge Test scores ($g = 3.54$) of Digital Tutor participants.

D. Discussion

With respect to the four objectives of Assessment 5 (IWAR 2), we found that:

- *The Digital Tutor provided its graduates with substantial IT knowledge and skill*—The Digital Tutor graduates were technically prepared to perform a wide variety of advanced IT duties needed at Navy duty stations.

- *With the exception of the Security Test, the skills and knowledge demonstrated by Digital Tutor graduates were superior to those of experienced Fleet ITs*—Performance by Digital Tutor teams was generally superior to that of experienced Fleet teams and of statistical and practical significance; the Digital Tutor teams solved about twice as many IT troubleshooting problems as Fleet teams, with very large effect sizes for both troubleshooting and IT knowledge.
- *Digital Tutor graduates demonstrated IT skills and knowledge that were superior to those of students who received classroom and laboratory instruction of greater duration*—With a statistically significant and very large effect size in comparing overall troubleshooting performance of 16-week Digital Tutor graduates with 35-week ITTC graduates, we conclude that Digital Tutor graduates acquired substantially more troubleshooting skills in much less time than the classroom and laboratory instruction. This finding is reinforced by the statistically significant and very large effect size found in comparing the IT Knowledge Test scores of the two groups. Other analyses found that these results were independent of reading ability and learning was more equitably distributed among digitally tutored than classroom students.
- *Digital tutoring produced superior troubleshooting skills, but findings for IT knowledge were less clear*—Troubleshooting exercises suggested a statistically significant and large effect size favoring digital over human tutoring. Overall, knowledge testing favored the digitally tutored sailors with a large effect size indicating practical significance, but the difference was not statistically significant.

The following comments summarize other findings and reflections on this work.

1. Corroboration from the Veteran's Project

Corroboration of the Navy findings was provided by a follow-on project to prepare military veterans for the civilian workforce (Fletcher 2014). In this effort, the Digital Tutor was modified for civilian IT employment as an 18-week IT training course, which was then completed by 97 mostly unemployed military veterans, few of whom had prior IT experience. There were no academic dropouts from the training. Six months after graduation, all but one of the veterans who sought IT employment had been hired with a median annual salary of $73,000, which is roughly equivalent to that of a Network Administrator II with 3–5 years of experience (Salary.com 2014).

2. Comparison Groups

The assessments reported here were subject to the usual vicissitudes of field research. Both experimental and comparison groups were small. Whether the comparison groups

provided sufficiently robust control groups remains subject to discussion. Also subject to discussion is the extent and adequacy of the Fleet ITs' expertise. For the latter, we had to rely on Fleet professional certifications and fitness reports, as well as the judgment of Fleet commanders and colleagues.

Differences in content of training, training objectives, and training approaches may also be questioned in the use of ITTC and IToF graduates as comparison groups. These differences seem particularly relevant for education, where the eventual application of what is learned is much less certain than it is in training, which prepares individuals for specified, known tasks and jobs. The challenge for training, then, is to produce graduates who are as well qualified as possible for the work to be done. When criterion measures of job and task competencies are valid and accurately reflect the tasks and jobs to be performed in the fleet or the field, choices of instructional content, approaches, and style are free to vary in any way that raises measures of readiness for task, job, and career performance.

3. Focus—Learning or Theory

An early motivation for the development of computer-assisted instruction was to validate theories of learning, memory, and cognition (Atkinson 1968; Suppes 1964). Researchers versed in mathematical psychology hoped to derive new hypotheses from empirically derived mathematical models of cognition, learning, and memory. These hypotheses would then be incorporated into computer programs for delivering instruction and tested for their validity with detailed data collected from computer interactions with learners. The primary intent was to advance cognitive theory.

This motivation contrasts with the pragmatic and eclectic approach taken by the developers of the DARPA Digital Tutor. Rather than focus on theory to be used in the development of instruction, their focus was to produce learning and derive theory from that—an approach much in the spirit of Simon's *Sciences of the Artificial* (1969).

These two objectives are obviously interdependent and both are essential, but which objective has priority guides priorities for the development of specific instructional interventions.

4. Blending with Human Monitors and Mentors

Although total cloning of human tutors by computer is not yet possible, significant aspects of cloning have been captured in software as the history and development of tutoring cloning suggests (e.g., Carbonell 1970; Feurzeig 1969; Graesser, D'Mello, and Cade 2011; Kulik and Fletcher, in press; Sleeman and Brown 1982; Psotka, Massey, and Mutter 1988; Van Lehn 2011; Woolf 2009). Still, human intervention seems likely to retain a critical and unique role in human-centered activities—including education and training.

Navy IT instructors provided blending needed to support instruction provided by the Digital Tutor. After 4 weeks of training, however, the Digital Tutor students outscored their Navy IT instructors on a knowledge test, with a very large effect size. Nonetheless, the Navy instructors were essential in conducting the course. They provided fallback and guidance for communication between the Digital Tutor and its students; proctoring to manage the flow and discipline of Digital Tutor classrooms and study halls; and, most important, orientation into the culture, traditions, and practices of the Navy.

5. Discovery and Guidance

The Digital Tutor's indirect approach to providing guidance, hints, and correct answers is notable. It did not provide learners with correct answers or direct hints, and it did not articulate learners' misconceptions. It uses information structures to guide learners in solving problems and discovering underlying concepts. It differs from some discovery learning, which throws learners into the subject and challenges them to fend for themselves. Through the use of reflection and review of what learners have already learned, the Digital Tutor works to guide them in finding their own way and devising for themselves the conceptual issues that underlie practical problems.

6. Abductive Reasoning

Expertise requires a combination of pattern recognition and subject-matter knowledge to enable what Charles Sanders Peirce (Douven 2011) described as abductive reasoning, or inference to the best (most economic) explanation—an ability to deal with unfamiliar, amorphous situations and discern what is fundamentally relevant to their functioning or malfunctioning. Peirce viewed abductive reasoning as a form of higher order induction that, unlike deduction, does not guarantee its conclusions, but yields confirmation-theoretic explanations that make some hypotheses more credible than others.

Abductive reasoning may be fundamental to expertise. It could be the basis for Clark and Wittrock's (2000) X-ray vision. The Digital Tutor's practice of asking learners to explain why something did or did not work is in accord with an inferential search for the best explanation and may account for some of its success. More direct research on abductive reasoning as a cognitive process and its contribution to successful problem-solving, decision-making, and learning seem in order.

7. Mixed-Initiative Dialogue

The DARPA Digital Tutor relied considerably on the use of information structures (i.e., intelligence) to provide pedagogical support for the learner and to parse free-form learner responses, but it did not attempt full, free-form tutorial conversations. How essential full natural language interaction with computers is for instruction remains to be

determined. Current capabilities hint at the value of such interaction, and the promise of fully conversational tutoring remains attractive and viable, but not yet within our grasp.

Nonetheless, conversational tutoring as envisioned for early systems like Mentor (Feurzeig 1969) and SOPHIE (Brown, Burton, and DeKleer 1982) seems as inevitable now as it did in the mid-1980s. This capability combined with the vast resources of human knowledge and information (and misinformation) available in the global information infrastructure suggests the eventual availability, if not inevitability, of tutorial, decision-aiding, and problem-solving conversations that are generated in real time and made available, through technology, anytime and anywhere (Fletcher, Tobias, and Wisher 2007). The interim capabilities applied by the DARPA Digital Tutor indicate that much can be done while we are waiting.

8. Minimize Cost or Maximize Effect

Self-pacing has long been a featured aspect of technology-based learning. Keyed to prior knowledge, learning rates appear to differ by a ratio of at least 4:1 (Gettinger and White 1980). Much of this difference can be accounted for by prior knowledge (Dochy, Segers, and Buehl 1999; Tobias 1989). Allowing learners to pass quickly through material they have mastered and concentrating on material they have yet to learn have been shown to decrease time to reach targeted learning objectives and reduce costs for instruction by as much as one-third (Fletcher 2004). However, both education and training institutions have difficulty in adjusting to students who complete their learning at arbitrary times.

A solution used by the Digital Tutor was to exploit the instructional agility of computer technology and allow each learner additional opportunities to learn in the time set aside for each subtopic. The Digital Tutor holds time spent within subtopic areas constant, but it allows learners who are proceeding rapidly through the material to learn more by presenting more difficult problems, related topics that could not be adequately taught to all learners in the time available, and challenge problems that provide minimal (or no) tutoring support. In this way, the potential of learners who could achieve higher and much needed levels of proficiency in the time available is not squandered.

9. Cost

Estimates of current per-student cost for the Digital Tutor include continued research for its development as it is being refined and improved. Cohn and Fletcher (2010) used these costs to calculate net present-day costs (including research and development) compared with classroom costs plus the 7 years of on-job training needed to reach levels of performance demonstrated by Digital Tutor students immediately upon graduation. Assuming a discount rate of 4%, as recommended by Levin and McEwan (2001), the cost would be about $180,000 or about 62% more per learner for ITTC than Digital Tutor training, primarily due to the need for continued on-job training in Fleet assignments.

Aside from these fiscal considerations, the operational return on investment may be even more substantial by avoiding losses that result from failure of IT systems in the Fleet or ships involved in combat engagements. The unpredictable and uncertain quantification of operational losses limits their use in pre-engagement or even pre-deployment analysis, but they may be the most critical factor in decisions concerning the use of training that accelerates acquisition of expertise.

10. Novice, Journeyman, and Expert

These three levels roughly characterize the knowledge and skills a course of instruction intends to produce. Much initial (ab initio) training aims for either novice or, at most, journeyman levels of ability, trusting expertise to be developed by job experience. An assumption is that the expense to develop expertise beyond entry or journeyman level in initial training is prohibitive. The DARPA Digital Tutor suggests otherwise. Accelerating expertise, produced in the time now allocated to ab initio training, may be at hand and could be more routinely targeted. Cohn and Fletcher (2010) and Fletcher (2014) suggest that the costs of lengthy on-job experience and training used instead to develop expertise may make it prohibitively expensive not to aim for expertise in initial training for technical occupations.

11. Return on Investment

The Digital Tutor was relatively expensive to develop and is, at present, expensive to use for instruction. On the other hand, return on this investment appears to justify its cost. A problem is that the funding to develop or apply a tutor must be spent up front. The return comes later, gradually, in degrees, and is often realized by a different organizational entity than the one that funded the development. The DARPA Digital Tutor was supported by a research and development budget. Continued funding will need to come from other sources.

A similar situation may exist for many educational reforms. It is not unusual for leaders in the practical world to ask, in effect, what a pound of education or training is worth. Many education and training investments are supported out of faith, but faith only goes so far. Education and training researchers may need to take the next step and routinely defend the monetary value of the instructional innovations they produce, as Levin and McEwan (2001), Ross, Barkaoui, and Scott (2007), Harris (2009), and others have suggested. It may be time, if not past time, for researchers in education and training to begin routinely including cost and return-on-investment analyses in their assessments.

12. Authoring Systems

The effort to enable individuals who are not computer-specialists (e.g., subject-matter experts, instructional specialists, classroom teachers) to develop (“author”) computer-

assisted instruction is long-standing (Buck and Hunka 1995; Tenczar et al. 1974). This effort has been successful in developing computer instruction that achieves lower level instructional objectives, often through the use of drill and practice using what Carbonell (1970) described as ad-hoc, frame-oriented instruction. Authoring systems enable instructional specialists who may not be proficient computer programmers to produce effective computer-assisted instruction by trading off pedagogical flexibility for programming simplicity.

Application of computer intelligence in computer-assisted instruction, such as that applied by the Digital Tutor, is almost exclusively coded by expert programmers. As instructional objectives aim at increasingly abstract and complex conceptual levels of learning, the instructional flexibility and human-computer synergy required to present them may make authoring by non-programmers infeasible. The need to budget for individuals who are proficient in computation, the subject matter, and tutorial techniques may be inescapable. Return on investment may well compensate for the cost.

7. Final Word

Researchers may argue (endlessly) about whether this effort and its findings represent a breakthrough for the use of technology in education and training. At a minimum the findings indicate the value of what can be done with sufficient resources applied to best effect.

That this effort was conducted in training rather than an education setting should not matter to educators. The knowledge scores speak to the promise of digital tutoring for education. They remind us that training and education exist on the same continuum. Most training (an effort to prepare people to do something) contains elements of education, and most education (an effort to prepare learners to live well) contains elements of training. There may be purity at both ends of the continuum, but it is rare in practice.

The memory-retrieval accuracy and computation speed of computer technology far surpass those of humans. These capabilities augment tutorial processes, economically and operationally, in ways not otherwise available and with increasing affordability. As these digital tutorial systems evolve, they may well develop and incorporate unique qualities, characteristics, and capabilities of their own—not unlike the evolution of automobiles from horseless carriages and radio from wireless telegraph. Like Columbus we may set out with one objective in mind and end up with something entirely unexpected. The voyage and its promise of reward should justify the effort.

References

Anderson, L. W., and D. R. Krathwohl, eds. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Taxonomy of Educational Objectives.* Columbus, OH: Allyn and Bacon.

Anzai, Y., and H. A. Simon, H.A. 1979. "The Theory of Learning by Doing." *Psychological Review* 86:124–40.

Atkinson, A. B. 1970. "On the Measurement of Inequality." *Journal of Economic Inequality* 2:244–63.

Atkinson, R. C. 1968. "Computerized Instruction and the Learning Process." *American Psychologist 23*(4): 225–39.

———. 1972. "Ingredients for a Theory of Instruction." *American Psychologist* 27:921–31.

Atkinson, R. C., and J. A. Paulson. 1972. "An Approach to the Psychology of Instruction." *Psychological Bulletin* 78:49–61.

Bloom, B. S. 1984. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as one-to-one Tutoring." *Educational Researcher* 13:4–16.

Bloom, B., M. Engelhart, E. Furst, W. Hill, and D. Krathwohl. 1956. *Taxonomy of Educational Objectives, the Classification of Educational Goals, Handbook I: Cognitive domain*. New York: David McKay Company.

Bourne, L. E, Jr., W. D. Raymond, and A. F. Healy. 2010. "Strategy Selection and Use During Classification Skill Acquisition." *Journal of Experimental Psychology: Leaning, Memory, & Cognition* 36:500–514.

Brown, J. S., R. R. Burton, and J. DeKleer. 1982. "Pedagogical, Natural Language and Knowledge Engineering in SOPHIE I, II, and III." In *Intelligent Tutoring Systems*, edited by D. Sleeman and J. S. Brown, 227–82. New York, NY: Academic Press.

Brown, J. S., R. R. Burton, and F. Zdybel. 1973. "A Model-Driven Question-Answering System for Mixed Initiative Computer-Assisted Construction." *IEEE Transactions on Systems, Man, and Cybernetics, SMC-3*, 248–57.

Buck, G., and S. Hunka. 1995. "Development of the IBM 1500 Computer-Assisted Instructional Language." *IEEE Annals of the History of Computing* 17 (1): 19–31.

Carbonell, J. R. 1970. "AI in CAI: An Artificial Intelligence Approach to Computer-Assisted Instruction." *IEEE Transactions on Man-Machine Systems* 11:190–202.

Chant, V. G., and R. C. Atkinson. 1978. "Application of Learning Models and Optimization Theory to Problems of Instruction." In *Handbook of Learning and*

Cognitive Processes, vol. 5, edited by W.K. Estes. Hillsdale, NJ: Erlbaum Associates.

Chatham, R. E., and J. V. Braddock. 2001. *Training Superiority and Training Surprise.* Washington, DC: Defense Science Board, Department of Defense.

Clark, R., and M. Wittrock. 2000. "Psychological Principles in Training." In *Training and Retraining: A Handbook for Business, Industry, Government, and the Military,* edited by S. Tobias and J. D. Fletcher, 51–84. New York: Macmillan Reference.

Cohen, J. 1988. *Statistical Power Analysis for the Social Sciences,* 2nd ed. Hillsdale, NJ: Erlbaum.

Cohn, J., and J. D. Fletcher. 2010. "What is a Pound of Training Worth? Frameworks and Practical Examples for Assessing Return on Investment in Training." *Proceedings of the InterService/Industry Training, Simulation and Education Annual Conference*. Arlington, VA: National Training and Simulation Association.

Craik, F. I., and R. S. Lockhart. 1972. "Levels of Processing: A Framework for Memory Research." *Journal of Verbal Learning and Verbal Behavior* 11:671–84.

Crowder, N. A. 1959. "Automatic Teaching by Means of Intrinsic Programming." In *Automatic Teaching: The State of the Art,* edited by E. Galanter, 109–16. New York, NY: John Wiley and Sons.

Csikszentmihalyi, M. 1990. *The Psychology of Optimal Experience*. London: Harper Perennial.

Dawes, R. M. 1971. "A Case Study of Graduate Admissions: Application of Three Principles of Decision-Making." *American Psychologist* 26:180–8.

Dochy, F., M. Segers, and M. Buehl. 1999. "The Relation between Assessment Practices and Outcomes of Studies: The Case of Research on Prior Knowledge." *Review of Educational Research* 69 (2): 145–86.

Douven, Igor. 2011. "Abduction." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. http://plato.stanford.edu/archives/spr2011/entries/abduction/.

Ericsson, K. A. 2006. "The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance." In *The Cambridge Handbook of Experts and Expert Performance,* edited by K. A. Ericsson, N. Charness, P. J. Feltovich, and R. Hoffman, 683–722. New York: Cambridge University Press.

Ericsson, K. A., N. Charness, P. J. Feltovich, and R. R. Hoffman, eds. 2006. *The Cambridge Handbook of Expertise and Expert Performance*. New York: Cambridge University Press.

Feltovich, P. J., M. J. Prietula, and K. A. Ericsson. 2006. "Studies of Expertise from Psychological Perspectives." In *The Cambridge Handbook of Expertise and Expert Performance,* edited by K. A. Ericsson, N. Charness, P. J. Feltovich, and R. R. Hoffman, 41–67. New York: Cambridge University Press.

Feurzeig, W. 1969. *Computer Systems for Teaching Complex Concepts.* BBN Report 1742. Cambridge, MA: Bolt Beranek and Newman, Inc. DTIC AD 684 831.

Fletcher, J. D. 1992. “Individualized Systems of Instruction.” In *Encyclopedia of Educational Research,* 6th ed., edited by M. C. Alkin, 613–20. New York: Macmillan.

———. 2004. “Technology, the Columbus Effect, and the Third Revolution in Learning.” In *The Design of Instruction and Evaluation: Affordances of Using Media and Technology,* edited by M. Rabinowitz, F. C. Blumberg, and H. Everson, 139–57. Mahwah, NJ: Lawrence Erlbaum Associates.

———. 2009. “Education and Training Technology in the Military.” *Science* 323:72–75.

———. 2010. *Phase 1 IWAR Test Results.* IDA Document D-4047. Alexandria, VA: Institute for Defense Analyses.

———. 2011. *DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor Assessments.* IDA Document NS D-4260. Alexandria, VA: Institute for Defense Analyses.

———. 2014. “Digital Tutoring in Information Systems Technology for Veterans: Data Report.” IDA Document D-5336. Alexandria, VA: Institute for Defense Analyses.

Fletcher, J. D., and M. R. Rockway. 1986. “Computer-Based Training in the Military.” In *Military Contributions to Instructional Technology,* edited by J. A. Ellis, 171–222. New York, NY: Praeger Publishers.

Fletcher, J .D., S. Tobias., and R. K, Wisher. 2007. “Learning Anytime, Anywhere: Advanced Distributed Learning and the Changing Face of Education.” *Educational Researcher* 36(2): 96–102.

Gettinger, M., and M. A. White. 1980. “Evaluating Curriculum Fit with Class Ability.” *Journal of Educational Psychology* 72:338–44.

Gick, M. L., and K. J. Holyoak. 1980. “Analogical Problem Solving.” *Cognitive Psychology* 12:306–55.

Goldberg, L. R. 1970. “Man Versus Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences.” *Psychological Bulletin* 73:422–32.

Gott, S. P., and A. M. Lesgold. 2000. “Competence in the Workplace: How Cognitive Performance Models and Situated Instruction Can Accelerate Skill Acquisition.” In *Advances in Instructional Psychology,* edited by R. Glaser, 239–327. Hillsdale, NJ : Erlbaum.

Graesser, A. C., S. K. D’Mello, and W. Cade. 2011. “Instruction based on tutoring. In *Handbook of Research on Learning and Instruction,* edited by R. E. Mayer and P. A. Alexander, 408–26. New York: Routledge Press.

Graesser, A. C., N. K. Person, and J. P. Magliano. 1995. “Collaborative Dialogue Patterns in Naturalistic One-on-One Tutoring.” *Applied Cognitive Psychology* 9:495–522.

Groen, G. J., and R. C. Atkinson. 1966. “Models for Optimizing the Learning Process.” *Psychological Bulletin* 66:309–20.

Harris, D. N. 2009. “Toward Policy-Relevant Benchmarks for Interpreting Effect Sizes: Combining Effects with Costs.” *Education Evaluation and Policy Analysis* 31:3–29.

Healy, A. F., J. A. Kole, and L. E. Bourne. 2014. “Training principles to advance expertise.” *Frontiers in Psychology* 5:1-4.

Hedges, L. V. 1981. “Distribution Theory for Glass’s Estimator of Effect Size and Related Estimators.” *Journal of Educational Statistics* 6 (2): 107–28.

Hoffman, R. R., P. Ward, P. J. Feltovich, L. DiBello, S. Fiore, and D. H. Andrews. 2014. *Accelerated Expertise: Training for High Proficiency in a Complex World.* New York: NY: Psychology Press.

Jamison, D., J. D. Fletcher, P. Suppes, and R. C. Atkinson. 1976. “Cost and Performance of Computer-Assisted Instruction for Education of Disadvantaged Children.” In *Education as an Industry*, edited by J. Froomkin, D. T. Jamison, and R. Radner. National Bureau of Economic Research. Cambridge, Massachusetts: Ballinger Publishing.

Jamison, D. T., P. Suppes, and S. Wells. 1974. “The Effectiveness of Alternative Instructional Media: A Survey.” *Review of Educational Research* 44:1–67.

Kulik, J. A. 1994. “Meta-Analytic Studies of Findings on Computer-Based Instruction.” In *Technology Assessment in Education and Training,* edited by E. L. Baker and H. F. O’Neil, Jr., 9–33. Hillsdale, NJ: Erlbaum.

Kulik, J. A., and J. D. Fletcher. In press. “Effectiveness of Intelligent Tutoring Systems.” *Review of Educational Research*.

Lesgold, A., S. Lajoie, M. Bunzo, and G. Eggan. 1988. “SHERLOCK: A Coached Practice Environment for an Electronics Troubleshooting Job.” In *Computer Assisted Instruction and Intelligent Tutoring Systems: Establishing Communication and Collaboration,* edited by J. Larkin, R. Chabay, and C. Scheftic, 201–38. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lesgold, A., and M. Nahemow. 2001. “Tools to Assist Learning by Doing: Achieving and Assessing Efficient Technology for Learning.” In *Cognition and Instruction: Twenty-five Years of Progress,* edited by D. Klahr and S. Carver, 307–46. Mahwah, NJ: Erlbaum.

Levin, H. M., and P. J. McEwan. 2001. *Cost-Effectiveness Analysis*. Thousand Oaks, CA: Sage.

Mayer. R. E. 2002. “Rote Versus Meaningful Learning.” *Theory into Practice* 41 (4): 226–32.

MacGintie, W. H., R. K. MacGintie, K. Maria, L. G. Dreyer, and K. E. Hughes. 2007. *Gates-MacGintie Reading Tests Level AR Forms S&T: Manual for Scoring and Interpretation*. Rolling Meadows, IL: Riverside Publishing.

McDaniel, M. A., M. J. Cahill, M. Robbins, and C. Wiener. 2014. “Individual Differences in Learning and Transfer: Stable Tendencies for Learning Exemplars Versus Abstracting Rules.” *Journal of Experimental Psychology, General* 143 (2): 668–93.

Meehl, P. E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Literature*. Minneapolis, MN: University of Minnesota Press.

Niemiec, R., and H. J. Walberg. 1987. "Comparative Effects of Computer-Assisted Instruction: A Synthesis of Reviews." *Journal of Educational Computing Research* 3:19–37.

Psotka, J., L. D. Massey, and S. A. Mutter, eds. 1988. *Intelligent Tutoring Systems: Lessons Learned.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Quillian, M. R. 1969. "The Teachable Language Comprehender: A Simulation Program and Theory of Language." *Communications of the ACM* 12 (8): 459–75.

Ross, J. A., K. Barkaoui, and G. Scott. 2007. "Evaluations that Consider the Cost of Educational Programs: The Contribution of High Quality Studies." *American Journal of Evaluation* 28:477–92.

Salary.com. 2014. "Network Administrator II." Retrieved 8 August 2014, http://swz.salary.com/SalaryWizard/Network-Administrator-II-Salary-Details.aspx?andyearsofexperience=3.5

Shalit, H. 1985. "Calculating the Gini Index of Inequality for Individual Data." *Oxford Bulletin of Economics and Statistics* 47:185–89.

Simon, H. A. 1969. *The Sciences of the Artificial.* Cambridge, MA: MIT Press.

Sleeman, D., and J. S. Brown, eds. 1982. *Intelligent Tutoring Systems*. New York, NY: Academic Press.

Sternberg, R., and J. Hedlund. 2002. "Practical Intelligence, g, and Work Psychology." *Human Performance* 15:143–60.

Suppes, P. 1964. "Modern Learning Theory and the Elementary-School Curriculum." *American Educational Research Journal* 1:79–93.

Suppes, P., J. D. Fletcher, and M. Zanotti. 1975. "Performance Models of American Indian Students on Computer-Assisted Instruction in Elementary Mathematics." *Instructional Science* 4:303–13.

———. 1976. "Models of Individual Trajectories in Computer-Assisted Instruction for Deaf Students." *Journal of Educational Psychology* 68:117–27.

Suppes, P., and M. Morningstar. 1972. *Computer-Assisted Instruction at Stanford 1966–68: Data, Models, and Evaluation of the Arithmetic Programs.* New York: Academic Press.

Tenczar, P., et al. 1974. "PLATO IV Authoring." *International Journal of Man-Machine Studies* 6:445–63.

Tobias, S. 1989. "Another Look at Research on the Adaptation of Instruction to Student Characteristics." *Educational Psychologist* 24:213–27.

Tobias, S., and T. M. Duffy, eds. 2009. *Constructivist Instruction: Success or Failure?* New York: Routledge.

U.S. Department of Education. 2009. Institute of Education Sciences. What Works Clearinghouse. July. *Middle School Math Intervention Report: Cognitive Tutor Algebra I*. Retrieved from http://whatworks.ed.gov.

U.S. Naval Education and Training Command. 2010. *Naval Education and Training Command Integrated Learning Environment Course Development and Life-Cycle Maintenance Manual* (NAVEDTRA 136). Pensacola, FL: Naval Education and Training Command.

VanLehn, K. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46 (4): 197–221.

Vinsonhaler, J. F., and R. K. Bass. 1972. "A Summary of Ten Major Studies on CAI Drill and Practice." *Educational Technology* 12:29–32.

Wetzel-Smith, S. K., and W. H. Wulfeck. 2010. "Training Incredibly Complex Tasks." In *Performance Enhancement in High Risk Environments,* edited by J. V. Cohn and P. E. O'Connor, 74–89. Westport, CN: Praeger.

Woolf, B. P., and J. W. Regian. 2000. "Knowledge-Based Training Systems and the Engineering of Instruction." In *Training and Retraining: A Handbook for Business, Industry, Government, and the Military,* edited by S. Tobias and J. D. Fletcher, 339–56. New York: Macmillan Library Reference.

Woolf, B. P. 2009. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing e-Learning*. Burlington, MA: Elsevier Academic Press.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
November 2014	Final	August 2014 – September 2014

4. TITLE AND SUBTITLE

Accelerating Development of Expertise:
A Digital Tutor for Navy Technical Training

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

J. D. Fletcher
John E. Morrison

5d. PROJECT NUMBER

5e. TASK NUMBER
BE-2-3831

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-5358
Log: H14-001221

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of the Under Secretary of Defense (P&R)
Training Readiness and Strategy Directorate
Defense Pentagon, Rm. 1E532
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

OUSD(P&R)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (28 July 2015).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Development of the Digital Tutor (DT) was undertaken by the Defense Advanced Research Projects Agency (DARPA) to accelerate the acquisition of expertise in initial technical training courses. DARPA chose Information Systems Technology (IT) as the Tutor's subject matter after extensive analysis revealed its ubiquity and criticality in defense and civilian operations. Five assessments of the Tutor have been completed and are documented by this report. The last four of these assessments were performed by the Institute for Defense Analyses (IDA) acting as an impartial, third-party reviewer. These assessments found that after 16 weeks of training with the Tutor, sailors who had no prior IT experience, scored higher in tests of IT knowledge and job-sample troubleshooting, often with effect sizes well in excess of two standard deviations, than others who received 35 weeks of classroom instruction and technicians with an average of 9 years of IT experience in the Fleet. Additional analyses found (a) the Digital Tutor to be significantly more effective than human tutoring in one case and better but not statistically significant in another, (b) effectiveness of the Tutor was not dependent on reading ability, (c) its effectiveness was more equitably distributed than that of classroom instruction, and (d) net present value to the Fleet from investment in the fully employed IT Tutor would exceed $300M annually.

15. SUBJECT TERMS

Digital Tutor; Intelligent Tutoring System; Computer Based Instruction; Computer Assisted Instruction; Training; Education; Navy Training; Information Systems Technology; Educational Technology; Problem Solving; Troubleshooting

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			DiGiovanni, Frank
Uncl.	Uncl.	Uncl.	SAR	68	19b. TELEPHONE NUMBER *(include area code)* 703-695-2618

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18