IDA

INSTITUTE FOR DEFENSE ANALYSES

DARPA Digital Tutor: Assessment Data

J.D. Fletcher
John E. Morrison

Draft Final
September 2012
IDA Document D-4686
Log: H 12-001207/1
Copy

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 DASW01-04-C-0003, Task DA-2-2896, "Technical Review and Analyses for Education Dominance," for the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgements

The assessments reported here could not have been accomplished without support from the men and women of the following commands: Naval Education and Training Command Center for Information Dominance; Center for Information Dominance Corry Field Detachment; Center for Information Dominance Monterey Detachment; Naval Base San Diego; Center for Information Dominance San Diego Detachment; Naval Network Warfare Command; and the United States Army Garrison Presidio of Monterey. Particular thanks for their many essential contributions are due to the Navy Fleet Systems Engineering Teams, George Kevlin, Jason Paper, and Tim Scarborough—and of course to Greg Hayes of NETWARCOM. Many thanks are also due to Dr. Rebecca Grier for her thorough and responsive review of this report. Finally, thanks are due to many analysts from IDA's Information Technology Systems Division but especially to Brendan Farrar-Foley for his help and time spent in organizing IWAR 2 data.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-4686

DARPA Digital Tutor: Assessment Data

J.D. Fletcher
John E. Morrison

Summary

The DARPA Digital Tutor effort serves two broad purposes—meeting a Navy operational need and advancing the technology of computer applications in instruction. It applies principles from a number of cognitive and instructional theories, but its approach is pragmatic and eclectic rather than theoretic. It is an attempt to make the advantages of one-on-one tutorial instruction scalable and readily accessible. Its strategy has been to observe in systematic and specific detail the practice of individuals who are expert in both a subject matter and tutoring and then capture their instructional techniques and capabilities in computer technology.

Based on an analysis of need and criticality, DARPA selected "A" school and some "C" school training for the Navy Information Systems Technician (IT) rating for this effort. Five assessments of the evolving DARPA Digital Tutor have been performed. In addition to standard tests of statistical significance, effect sizes ("sigma") were also calculated. In common use, an effect size equal or greater than 0.75 but less than 1.10 is considered "large" and an effect size equal or greater than 1.10 is considered "very large." Both are rarely found in research on instruction.

This report summarizes results from the first four assessments, which have been reported elsewhere, and discusses Assessment Five (IWAR 2) in more detail.

Assessment One compared the IT knowledge of students who had learned primarily from human tutoring with those of students who had completed the existing "A" school IT training. Assessment Two (IWAR 1) compared both the knowledge and skills of the human tutored students with sailors who had an average 7.2 years of IT experience in the Fleet. Both assessments showed substantial differences in favor of the tutored students in IT troubleshooting and IT knowledge, with some effect sizes in excess of 2.00.

Assessment Three compared the IT knowledge of students who had completed the 4 weeks of the digitized tutor (DT) then available with that of graduates of the standard IT "A" school and their instructors. The DT students outscored the "A" school students with an effect size of 2.81 and the instructors with an effect size of 1.26.

Assessment Four compared both the knowledge and practical exercise skills of DT students who had completed the 7 weeks of the Digital Tutor then available with those of graduates who had completed 19 weeks of revised, primarily classroom "A" school IT training. The DT students outscored the "A" school graduates in both practical exercises (e.g., troubleshooting) and knowledge, with most effect sizes well in excess of 1.10.

Assessment Five (IWAR 2), which is highlighted in this report, examined the first completed 16-week version of the Digital Tutor. It compared the skills and knowledge of DT graduates with those of graduates from the (then) 35-week IT Training Continuum (ITTC) course and those of Fleet ITs with an average of 9.1 years of experience. The assessment was conducted, as was IWAR 1, in two weeklong sessions each involving 4 days of practical exercises, interviews with a Review Board, and Knowledge Testing. Three groups consisting of 12 DT, Fleet, and ITTC participants were examined—6 participants in each session. Troubleshooting exercises were drawn from a database of about 20,000 trouble tickets that had been referred by the Fleet to shore-based Fleet System Engineering Teams for solution.

Participants were divided into teams of three for the Troubleshooting exercises. The DT teams outscored the Fleet ITs in 2–1/2 days of troubleshooting with a "large" effect size of 0.85 and ITTC graduates with a "large" effect size of 1.13. These differences are statistically significant. The DT teams solved 74 percent of the problems they attempted compared with 51 percent for the Fleet teams and 38 percent for the ITTC teams. DT teams also solved three of the problems classified as "very hard" compared with none solved by either the Fleet or ITTC teams. In their solution attempts, the probability that DT teams would leave a harmful action in the system was 0.14 compared with 0.41 for the Fleet teams and 0.33 for the ITTC teams. These differences are statistically significant. The average number of unnecessary steps while attempting to solve a troubleshooting problem was 0.48 for the DT teams, compared with 1.24 for Fleet teams and 1.43 for ITTC teams. These differences are also statistically significant.

Participants remained in teams of three for the Security exercise. Fleet teams outperformed both the DT and ITTC teams on the Security exercise by finding and correcting 69 percent of embedded security violations compared to 44 percent and 45 percent found and corrected by DT and ITTC teams, respectively. These differences favor the Fleet teams, but they are not statistically significant.

In the Network Design and Development exercise, the DT teams received statistically significant higher ratings on critical objectives than Fleet and ITTC teams, with "very large" and "large" effect sizes, respectively. They also received statistically significant higher ratings than ITTC teams on secondary objectives and overall, with "very large" effect sizes in both cases. DT ratings were larger, but not statistically significant for secondary objectives nor overall in comparison with Fleet ratings.

On the Knowledge Test, the DT graduates outscored the Fleet ITs with an effect size of 4.30 and ITTC graduates with an effect size of 3.38. Both differences are statistically significant. IT knowledge as tested in these assessments is important, accounting for about 40 percent of the variance in performance of practical exercises, but it is an enabler of performance rather than a direct measure of performance itself.

Additional analyses found that:

- Digital Tutor students in IWAR 2 outscored those who received human tutoring in IWAR 1, but not at a statistically significant level.
- The advantage of DT over ITTC training on the Knowledge Test was about the same for both low-scoring and high-scoring students.
- Reading ability appeared to be unrelated to Knowledge Test scores for the DT students, but reading vocabulary was mildly related to Knowledge Test scores for ITTC students (accounting for about 35 percent of the variance).
- Armed Forces Qualification test scores were mildly related to Knowledge Test scores for the DT students (accounting for about 37 percent of the variance), but more strongly related to Knowledge Test scores for the ITTC students (accounting for about 59 percent of the variance).
- Although success in troubleshooting, as measured in Assessment Four, was related to DT scores on the Knowledge Test (accounting for about 41 percent of the variance), the relationship was effectively zero for classroom-trained "A" school students.

In sum, the DARPA Digital Tutor effort appears to have achieved its goals. The design of the Digital Tutor is likely to be a significant advance in the development of training overall and of instructional technology in particular. Moreover, the Digital Tutor has shown that in 16 weeks it can produce students who outperform students with more than double that time in classroom instruction and sailors with 7–9 years of Fleet experience. These comparisons have included lengthy tests of knowledge and job-sample, practical exercises, both of which found levels of performance by the Digital Tutor students at levels that are unprecedented in assessments of training effectiveness. The greater efficiency, absence of harmful errors, and ability to solve problems at the highest levels of difficulty demonstrated by Digital Tutor students suggest both monetary and operational returns of substantial value to the Navy.

Contents

1. Background

This report presents data and findings from five assessments of the DARPA Digital Tutor. The first four assessments have been reported earlier (see Fletcher 2010, 2011) and are only summarized here. IWAR 2, a fifth assessment and the first to test a complete, 16-week digitized version of the tutor, is the focus of this report. All assessments except the first were conducted by the Institute for Defense Analyses (IDA) acting as an independent third party.

The DARPA Digital Tutor effort serves two broad purposes—meeting operational needs of the Navy and advancing the technology of computer use in instruction. It applies principles from a number of cognitive and instructional theories, but its approach is pragmatic and eclectic rather than theoretic.

The Digital Tutor development was initiated by DARPA's Training Superiority Program and continued under its successor, Education Dominance. Both programs were initiated by Dr. Ralph Chatham. In preparing for this effort he reviewed a wide variety of technical training courses, or "schools," across the Department of Defense. He assessed (1) the criticality of the human performance they were intended to produce; (2) the need, as recognized across all echelons of operational and training commands, for their improvement and revision; and (3) the difficulty of meeting that need through conventional training techniques. He identified about 40 technical domains as targets for DARPA investment. Navy training for the Information Systems Technician (IT) rating was the most prominent among these and was chosen along with three other domains for DARPA research and development.

Design and development of the IT Digital Tutor began with a detailed and comprehensive analysis by the developer to identify the knowledge and skills required for expert (well beyond journeyman) IT performance in the Fleet. This analysis was pursued with particular care and vigor for the DARPA Digital Tutor. It included numerous observations and interviews with Fleet IT personnel aboard Navy vessels. It identified specific knowledge and skills required for, and especially characteristic of, expert IT performance. It focused on high-level conceptual IT knowledge and generalizable skills that would maximize retention and transfer of the training provided.

Findings from this analysis established many of the instructional objectives for the Digital Tutor and standards for performance. Additional objectives and standards were derived from instructional content covered by the 16-week IT "A" school, conducted by

the Center for Information Dominance at Corry Station, Florida, and from content in the following 5 "C" schools:

- Journeyman-Network Core (JNETCORE).
- Advanced Network Analyst (ANA).
- Information Systems Security Manager (ISSM).
- Network Security Vulnerability Technician (NSVT).
- Navy Tactical Command Support System (NTCSS) Manager.

The objectives and standards identified by this analysis guided the search for expert (human) tutors. The search began by locating candidate tutors with widely recognized and peer-acknowledged expertise in requisite sub-domains. These individuals were then examined for their ability to tutor learners in one-on-one (one instructor with one learner) settings—an ability that differs appreciably from instructional expertise in one-on-many classroom settings (e.g., Graesser, Person, & Magliano, 1995; Graesser, D'Mello, & Cade, 2011). About half of the candidate tutors were "auditioned" in half-hour sessions to assess their ability to tutor students representative of sailors to be trained as ITs. Twenty-four individuals were chosen by this process to provide tutorial instruction in their specific area of IT expertise so that it could be captured by computer.

Total replication of human tutorial capabilities by computer is currently not possible, and may never be, but significant aspects of it have been captured in software as the history of tutoring systems suggests (e.g., Carbonell, 1970; Sleeman & Brown, 1986; Psotka, Massey, & Mutter, 1988; Woolf & Regian, 2000, Graesser, D'Mello, & Cade, 2011; Van Lehn, 2011; Kulik & Fletcher, 2012). Computer technology has capabilities (e.g., memory speed, capacity, and accuracy) at levels that humans lack. These capabilities augment tutorial processes in ways not otherwise readily available. As these tutorial systems evolve, they may well incorporate unique qualities, characteristics, and capabilities of their own—not unlike the evolution of automobiles from horseless carriages. Nonetheless, an essential first step in developing the DARPA Digital Tutor was to begin with an effort to capture and clone expert human tutoring as a way to meet Navy operational needs and advance the technology.

2. Assessments One through Four

All training and assessments required for the Digital Tutor effort were made possible by cooperation with the Navy's Center for Information Dominance (CID). CID supplied spaces, facilities, and students at the IT school at Corry Station; spaces and facilities at the CID detachment in the Monterey Defense Language Institute; and spaces and facilities at the CID detachment, San Diego Naval Base, where ready access to the Fleet ITs needed for participation in IWAR 1 and 2 was also available.

There have now been five assessments of the Tutor. The first two concerned sailors who were primarily trained by human tutors. The second of these was designated as IWAR 1. It provided summative evaluation of the human tutoring. The third and fourth assessments provided formative evaluation of the Digital Tutor as the tutoring was being digitized. The final assessment was designated as IWAR 2. It provided summative evaluation of a fully digitized version of the Digital Tutor.

None of these assessments involved training that was purely human or purely computer-based. About 1 week of the Digital Tutor was available and used by sailors who participated in the first two assessments, which primarily used human tutoring. All instruction that was conducted by humans, and then increasingly by computer, typically involved daily tutoring lasting 5–6 hours followed by a 2-hour instructor-led study hall. The content, quality, and structure of these study halls were variable and at the discretion of the instructor in charge.

A. Calculating and Interpreting Effect Sizes

Along with other statistics in this report, effect sizes are included where possible. They measure effects with the following calculation,

$$\text{Effect Size} = \frac{\text{Mean of Group 1} - \text{Mean of Group 2}}{\text{Standard Deviation}}$$

Effect size is therefore a measure of standard deviations. It is strictly a descriptive statistic, like means and standard deviations. It does not address statistical probability.

Statistically significant results, which account for the probability of their occurrence, may be found with small effect sizes—and vice versa.[1]

Effect size may be called "sigma" in colloquial discussions because it is usually designated with the Greek letter sigma (σ) in mathematical notation. In this report, effect sizes are calculated as Cohen's *d,* which is based on pooled standard deviations (Cohen, 1988). It assumes that all subjects are drawn from a common population and that the standard deviation of any sample of subjects is an estimate of the full population standard deviation. When there are more than two groups involved, all groups are pooled to estimate this population standard deviation—based on the pooled variance (mean-squared 'error') across the groups.

Researchers continue to debate the best way to calculate effect size. Pooling standard deviations is expected to provide a more stable estimate of the population standard deviation. This practice is similar to using pooled variance to test pairwise differences following analysis of variance (ANOVA). Other researchers (e.g., Glass, 1976) have argued for using the standard deviation of the control or comparison group only as the denominator in effect size. Means reported here generally include standard deviations enclosed within parentheses, or they are provided in tables to allow readers to calculate alternative values of effect size.

Cohen (1988) offered some rough guidelines to help interpret effect size values. He characterized effect sizes of 0.20 as "small," 0.50 as "medium," and 0.80 as "large." As shown in Table 1, Thalheimer and Cook (2002) extended Cohen's guidelines by providing specific limits to these three categories and adding three more levels: "negligible," "very large," and "huge" to the nomenclature. We employ these terms to describe IWAR 2 effects in the current study. There are no generally accepted terms for effect sizes of 2.00 or greater.

Table 1. Terms to Describe Effect Size Values.

Effect Size Values	Description
$-0.15 \leq d < 0.15$	Negligible
$0.15 \leq d < 0.40$	Small
$0.40 \leq d < 0.75$	Medium
$0.75 \leq d < 1.10$	Large
$1.10 \leq d < 1.45$	Very large
$1.45 \geq d$	Huge

Note. Adapted from Thalheimer and Cook (2002).

[1] We set a cut-off for statistical significance of $p < 0.05$ for this report. That is, if a result could occur by chance more than 5 times in 100, we assumed it was not statistically significant.

B. Assessment One—April 2009

A human-tutored IT training course began in January 2009 with 15 sailors who had completed recruit training and would otherwise have begun 16 weeks of initial IT technical training (or "A" school training) at CID in Corry Station. IT Training for these 15 sailors was conducted for 16 weeks at the CID Detachment in the Defense Language Institute in Monterey, California, where the necessary spaces—cubicles for one-on-one tutoring sessions and rooms for computers and other equipment—were available. In all tutoring sessions, video, audio, and system instrumentation data were recorded for detailed observation and adaptation of the expert-level IT knowledge, skill, and tutorial techniques to be incorporated in the Digital Tutor.

The first assessment was undertaken at the request of the CID Commanding Officer. It used a written paper and pencil test to compare the IT knowledge of the 15 students in Monterey after the first 10 weeks of primarily human-tutored training with those of 17 "A" school Integrated Learning Environment (ILE) graduates at CID/Corry Station.

The ILE course used computer-based training to present self-paced instruction that was designed to run for 12 weeks. On average, its students finished the course in about 10 weeks of ILE instruction. At the time of the assessment, Monterey students had received the single, first week of computer-based Digital IT training that was then available followed by 9 weeks of the human tutored IT training. The daily schedule in both cases was 5–6 hours of instruction followed by a study hall of about 2 hours.

CID/Pensacola instructors developed the written Knowledge Test for this assessment. It included multiple-choice, network-diagram, and essay questions. The test was administered in April 2009 to the 15 Monterey students and the 17 "A" school graduates. They averaged scores of 77.7 (11.8) and 39.7 (18.7) points, respectively. This difference is statistically significant at $p < 0.01$ and suggests an effect size of 2.36, which would be characterized by Thalheimer and Cook (2002) as "huge." It is roughly equivalent to raising the scores of 50th percentile students to the 99th percentile of performance. Separate analyses showed no differences between the Armed Forces Qualification Test (AFQT) and IT Qualifying Scores of the two groups.

C. Assessment Two (IWAR 1)—July-August 2009

DARPA's IWAR 1 was conducted jointly with the assistance of the Navy Network Warfare Command. It provided both formative and summative assessment of the Monterey human tutoring. It involved 5-day assessments of the knowledge and skills acquired by the Monterey students who completed the course (Fletcher, 2010). Participants consisted of 12 Monterey graduates[2] and 12 Navy ITs with an average of 7.2

[2] Three students of the original 15 had been dropped late in the course for nonacademic reasons.

years of Fleet IT duty experience. Space and computer equipment limitations required the assessment to be conducted over 2 weeks, in two 5-day sessions, with 6 Monterey students and 6 Fleet ITs in each session.

1. Measurement

Testing consisted of:

- 4 days of practical exercises with:
  - 13.25 hours of hands-on troubleshooting typical Fleet IT casualties,[3] installed in both virtual and physical systems that mirrored shipboard IT systems;
  - 4 hours of security testing;
  - 7 hours of IT system design and development.
- 4 hours of paper-and-pencil Knowledge Testing.

2. Results

a. Practical Exercises

The practical exercises consisted of troubleshooting, security, and system design and development. They provided the most direct evidence of the extent to which the training achieved its technical and operational goals.

For troubleshooting, the Monterey graduates and Fleet ITs worked in teams of three. The four teams of Monterey graduates solved an average of 24.8 (3.10) problems compared to 19.8 (5.74) solved by the four Fleet IT teams. Although this difference favors the Monterey teams, it could occur by chance about 9 times out of 100 ($p < 0.09$) and was therefore not judged to be statistically significant, leaving uncertain if Monterey team performance was better than Fleet team performance. It was at least equal to it. These data also indicate an effect size of 1.06, which would be classified as "large."

In troubleshooting discipline and technique, the Monterey teams left fewer harmful changes uncorrected (a total of 8 compared with 18 by the Fleet teams). They also verified more problems (97 percent compared with 85 percent) and solutions (95 percent compared with 77 percent).

On the Security test, the 4 Monterey graduate teams and the 4 Fleet IT teams averaged 23.75 (9.29) and 37.25 (8.77) points, respectively, which were 23 percent and

[3] These problems were derived from over 20,000 West Coast and East Coast Fleet trouble reports requesting shore-based technical assistance.

35 percent of the total possible. These differences favoring the Fleet IT teams are statistically significant ($p < 0.05$) with a "very large" effect size of 1.49.

In one sense, the Security test results validated the approach used in developing the instruction because the expert human tutor scheduled for the security section was pulled away after about 2 days, requiring rapid selection of last-minute substitutes who were not of expert quality in either tutoring or the subject matter. This problem and its consequences appear to have been carried over to the digitized version, because tutorial and subject-matter expertise on which to model the digitized instruction was absent.

In System Design and Development, the Monterey graduates and the Fleet ITs participated as six-member, self-organized teams in each of the two 5-day IWAR sessions. The two Monterey teams successfully accomplished 32 percent of the objectives, and the two Fleet teams successfully accomplished 34 percent. The Monterey teams scored a total of 84.5 points out of 220 and the Fleet teams scored 113.5 each. The small sample sizes precluded statistical analysis of significance.

b. Knowledge Test

On the Knowledge Test, the Monterey graduates and Fleet ITs averaged 146.7 (68.0) and 86.7 (43.9) points, respectively, out of a total of 278. This difference is statistically significant ($p < 0.01$) and suggests a "large" effect size of 1.02, indicating that 50th percentile Monterey students scored at about the same level as 85th percentile Fleet ITs.

Sailors in the remaining three assessments received Digital Tutor training for about 6 hours a day, supplemented by a study hall of about 2 hours led by an instructor.

D. Assessment Three—April 2010

Assessment Three (Fletcher, 2011) examined the knowledge and skills of 20 students who had completed 4 weeks of the computerized DT training then available. Testing involved comparisons with 31 students who had graduated from the self-paced ILE "A" school (averaging about 10 weeks in duration) and with 10 CID IT instructors.

1. Measures

A written Knowledge Test was administered to all three groups in two 2-hour sessions. This test was based on the original Knowledge Test developed by CID.

2. Results

The mean Knowledge Test scores for the 20 DT students, 31 ILE graduates, and 10 IT instructors were 128.4 (14.5), 63.8 (27.0), and 99.8 (34.0), respectively. All pairwise comparisons were statistically significant. Their effect sizes are shown in Table 2. The

effect size difference favoring DT students over ILE graduates was 2.81, which would be characterized as "huge." This result suggests that 50th percentile DT students scored at about the 99th percentile of ILE graduates. The effect size difference favoring DT students compared with instructors was 1.26, or "very large," suggesting that 50th percentile DT students scored at about the 90th percentile of CID IT instructors. The effect size difference favoring instructors over ILE graduates was 1.25, or "very large," suggesting that 50th percentile instructors scored at about the 89th percentile of the ILE graduates.

Table 2. Assessment Three Effect Sizes for All Pairwise Comparisons.

	ILE	Instructors
DT	2.81	1.26
ILE	—	1.25

E. Assessment Four—November 2010

This assessment (Fletcher, 2011) examined the knowledge and skills of 20 students who had completed the 7 weeks of computerized DT training then available.

1. Participants

The assessment compared the IT knowledge and skills of four groups:

- 20 DT students who had completed 7 weeks of the DT training.
- 20 IT of the Future (IToF) students who had graduated from a new, revised 19-week IT "A" School primarily consisting of classroom instruction.
- 17 graduates of the original ILE self-paced "A" School.
- 10 CID instructors, who had been trained to present IToF material.

2. Measures

The DT students and the IToF graduates were examined in Troubleshooting Exercises, Packet Tracer Exercises, and individual interviews conducted by a three-member Review Board consisting of experienced Navy ITs led by a senior FSET (Fleet Systems Engineering Team) member. All four groups took the Written Knowledge Test.

3. Results

a. Practical Exercises

The DT and IToF students participated as individuals in Troubleshooting Exercises and were scored by pairs of experienced ITs who had to agree on a single score, ranging

from 0 to 5. In practice, the scores assigned by individual members of each pair rarely deviated by more than one point. Troubleshooting consisted of 15 trouble tickets presented on virtual systems. These trouble tickets were chosen from the database of 20,000 trouble tickets mentioned earlier. The virtual systems used Fleet software but only one server, thereby reducing hardware requirements while maintaining software fidelity.

The DT students averaged 26.55 (14.09) points in these exercises, and IToF graduates averaged 5.65 (6.56) points. The variance for the IToF students exceeds the average because many IToF students scored zeros in this exercise. The difference is statistically significant ($p < 0.01$), with an effect size of 1.90, or "huge," suggesting that the 7-week DT students were performing troubleshooting at about the 98th percentile of IToF graduate performance.

The Packet Tracer Exercise consisted of 18 trouble tickets presented on virtual systems. Results were based on two types of scores: weighted for problem difficulty or not weighted. The DT students averaged 36.91 (16.2) unweighted points on these exercises, and IToF graduates, who had received more training with the Packet Tracer program, averaged 25.29 (15.3) unweighted points. This difference is statistically significant ($p < .05$), with a "medium" effect size of 0.74, suggesting that the DT students were performing at about the 77th percentile of IToF graduates. The DT students averaged 30.39 (15.9) points weighted for problem difficulty and IToF graduates averaged 15.85 (13.0) weighted points. This difference is statistically significant ($p < .01$), with a "very large" effect size of 1.00, suggesting that the 7-week DT students were performing at about the 84th percentile of the 19-week IToF graduates.

b. Review Board Interviews

The Review Board interviews were conducted with the sailors individually. Board members were not told from which group each sailor was drawn. Time only permitted interviews with 7 DT students and 6 IToF graduates drawn at random from participants in this assessment. The Board rated each sailor on a nonlinear scale. It awarded a 1 to a sailor who demonstrated less than 3 months of experience; a 2 for evidence of 3 months of experience; a 3 for evidence of 1–3 years of experience; a 4 for evidence of 4–5 years of experience; and a 5 for evidence of more than 5 years of experience. Each of the 3 Board members could award up to 30 points covering 6 topics, making a total of 90 points possible per sailor. The DT students averaged 41.64 (12.93) points in these interviews, and IToF graduates averaged 18.80 (20.85) points. This difference is statistically significant ($p < .01$), with a "very large" effect size of 1.34 suggesting that the DT students scored at about the 91st percentile of IToF graduates.

c. Knowledge Test

Finally, Table 3 shows mean and standard deviation scores on the Written Knowledge Test. The difference between the IToF graduates and the CID instructors is not statistically significant, but all other pairwise means are statistically significant at ($p <$ 0.01) or greater.

Table 3. Assessment Four: Knowledge Test Means, Standard Deviations, and Number of Observations for 7-Week DT Students, 19-Week IToF Graduates, ILE Graduates, and CID IT Instructors.

Group	Mean	Std Dev	N
DT	207.90	37.30	20
IToF	145.75	25.18	20
ILE	64.52	19.96	17
IT instructors	149.30	53.96	10

Effect sizes are shown in the matrix provided as Table 4. The effect sizes from pairwise comparisons of DT with Instructors are "very large," and the effect size from pairwise comparisons of IToF with Instructors are "negligible." The remaining three effect sizes would be characterized as "huge."

Table 4. Assessment Four: Knowledge Test Effect Sizes.

	IToF	ILE	Instructors
DT	1.95	4.68	1.35
IToF	—	3.54	–0.10
ILE	—	—	–2.65

3. Assessment Five: IWAR 2

As in IWAR 1, IWAR 2 was conducted jointly with the assistance of the Navy Network Warfare Command. It provided formative and summative assessment of a completed 16-week version of the Digital Tutor. It was conducted in two successive 5-day sessions in late March and early April 2012. Eighteen participants (six participants from each of three groups) were examined in each of the two sessions. The number of participants and scheduling were determined by the availability of computer systems and available spaces for the assessment.

A. Background

1. Objectives

IWAR 2 was conducted to answer four basic questions:

- Has the DT program achieved its training objectives in providing students with Fleet-required IT knowledge and skill?
- How do the knowledge and skills acquired by DT graduates compare with those of experienced Fleet ITs?
- Has the DT program captured in digital form the human tutoring effectiveness found in IWAR 1?
- How does a tutoring system such as the DARPA DT compare to classroom instruction in producing targeted knowledge and skills?

The first two questions address the goal of meeting a Navy operational need. The third and fourth questions address the goal of advancing the technology of computers used in education and training.

2. Participants

IWAR 2 participants consisted of three groups:

- 12 graduates from the 16-week DT training—Six Digital Tutor graduates were drawn from each of two 20-student classes that completed DT courses on 2 March 2012 and 9 March 2012. No certification testing had been given to these graduates before they participated in IWAR 2.

- 12 graduates from 35 weeks of ITTC training—These students were the first 12 who had passed their certification exam out of the full class of 30 students who graduated on 9 March 2012. All had enlisted in the Navy as 6 Year Obligators. They were awarded Navy Enlisted Code (NEC) 2790 ("Information Systems Technician") after completing the 19-week "A" IT School, which requires Computing Technology Industry Association (CompTIA A+), and Microsoft Certified Professional (MCP) XP certification for completion. From follow-on 16-week ITTC training, they received NEC 2791 ("Information Systems Administrator"), which requires Security + certification, and some Cisco Certified Network Associate Routing & Switching (CCNA) training.
- 12 Fleet ITs with 4–15 years (or an average of 9.1 years) of experience as Fleet ITs, all with NEC 2791 and most with additional certifications—none were ITTC alumni. They were chosen to be representative of Fleet ITs at this level of experience. Nine of the Fleet ITs were assigned to small-deck ships, which generally offer a wider range of IT experience than that available to the remaining three Fleet ITs, who were assigned to aircraft carriers where specialized IT experience is more common. However, most of the Fleet ITs had both large-deck and small-deck experience.

3. Support Teams

A "White Team" made up of senior Navy ITs and three members of the Navy's FSET was essential in conducting IWAR 2. Members of this team interviewed participants in the Review Boards, organized participants for IWAR practical exercises, ensured that exercise parameters and procedures were observed, scored all performance in the exercises, and coordinated IWAR activities and proceedings with the Technical Support team. In scoring exercise performance, White Team members divided into three 3-member teams—generally headed by an FSET—and rotated among the three participating groups (DT, Fleet, and ITTC), rating teams from each group an equal number of times.

Also essential to IWAR 2 was the Technical Support Team, which was provided by the research contractor and was responsible for proper initialization, management, and operation of the IWAR hardware and software. IWAR 2 did not experience any significant technical disruptions during either of its two 5-day sessions, which is commendable, especially for such a complex integration of disparate computer systems during an exercise of IWAR 2 intensity.

4. Facilities

IWAR 2 participants were tested in three separate classrooms provided by the Navy's 3rd Fleet at the San Diego Naval Base. Each classroom contained three IT

systems—one physical system, with a full complement of servers and software, and two identical virtual systems running on virtualized hardware hosted on a single server but supporting the same software. The systems were designed to mirror those typically found on Navy vessels and installations.

Three systems in each room allowed one to be prepared for the next exercise while two teams worked to solve problems on the others. Each system consisted of a working network made up of Common PC Operating System Environment (COMPOSE) 3.0 running on three servers with a backbone, although the virtual systems did not have the throughput capacity of a full hardware implementation. Three workstations were provided for each of the virtual systems.

The physical system had a server rack with four servers, one UNIX system, two backbone switches, and four edge switches. Connected to the rack were three workstations on a network, again made to resemble operational Navy systems as closely as possible. It required more time to prepare the physical systems than the virtual systems for a troubleshooting problem. 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, both IWAR 1 and IWAR 2 presented more virtual than physical system problems. Some limited access to the Internet was provided to participants as needed for specific problems, but more general World Wide Web access was not allowed.

The exercise classrooms were also instrumented with video cameras and microphones. Participant activity was available live and time stamped for later review and analyses using tools developed for DT development and training.

A fourth, larger classroom housed the hardware and software that the Technical Support Team used to control the practical exercises. This team was responsible for injecting troubleshooting problems, correcting any system technical issues that arose, and ensuring that the systems could be restarted quickly when that was required. Spare hardware for every system component was available, along with spare disks with cloned images for the servers so that any technical problems could be quickly resolved.

5. IWAR 2 Assessment

IWAR 2 assessment consisted of five major activities: three types of practical exercises, interviews with a Review Board, and completion of a Written Knowledge Test. All participants completed the Written Knowledge Test before beginning IWAR 2. Review Board interviews and the Practical Exercises were scheduled as shown in Table 5.

Table 5. Schedule for Each of the Two 5-day IWAR Sessions.

Monday	Tuesday	Wednesday	Thursday	Friday
Review Board Interviews with IWAR participants.	Practical Trouble-shooting exercises (Six teams of three individuals—two teams at the same time in each room)	Practical Trouble-shooting exercises continued	Practical Trouble-shooting 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)

6. Practical Exercises

As in IWAR 1 and as shown in Table 5, there were three types of practical exercises: Troubleshooting by three-member teams over a period of 2.5 days, a Security exercise performed by the same teams for a period of about 6 hours, and a System Design and Development exercise conducted for a period of about 6 hours by all six members of each IWAR 2 group in a self-organized team.

a. Troubleshooting Exercises

As job sample exercises, troubleshooting problems were the core assessments in IWAR 2. As in IWAR 1, IWAR 2 participants in the troubleshooting exercise were organized into three-person teams—two teams from each group of participants. In troubleshooting, one team worked to solve a problem on the physical hardware system while the other team worked to solve a problem on one of the virtual systems. Differences in team performance were assessed on group means calculated from four data points (four teams for each group). Six teams were tested in each week's session using two forms of the troubleshooting exercise problems that differed in length and content.

Troubleshooting problems were presented as they are at Navy duty stations, as Trouble Tickets. These were again drawn from the database of about 20,000 trouble tickets that had been referred to shore-based FSET organizations for solution. Figure 1 shows a sample trouble ticket that would be presented to IWAR participants, who were then required to solve the problem, describe the solution, and document the steps they had taken to correct it. Figure 2 shows the setup instructions for this sample troubleshooting problem.

A total of 210 troubleshooting problems were developed for IWAR 2. Of these, 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—initially 92 in the first week's session and 90 in the second week's session. Each team began the exercises for a session with troubleshooting problems presented in the same order as

those for the other teams. Fifteen minutes after a problem was presented, teams were free to move to the next problem when they chose to do so. The teams were free to use their own notes and the reference materials that were provided by the technical team on compact discs.

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 1. 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 2. Example Troubleshooting Problem Description and Setup Instructions.

b. Security

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

c. System Design and Development

This exercise was performed by all six participants from a participant group (DT, ITTC, or Fleet) operating as a self-organized team. They were given hardware, including servers, routers, cables, and switches; software, such as Windows XP and Microsoft Office; and a block of 128 IP addresses. These materials were sufficient to design and implement a system that met a set of specifications consisting of both critical and secondary objectives. Figure 3 shows Sample Objectives 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, depending on their performance. Different objectives were presented in the two sessions. 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.

7. Review Board Interviews

This review consisted of an individual 20–30 minute interview with a three-member Review Board. Three boards operated in parallel and interviewed an equal number of IWAR participants randomly assigned to the Board on the first day of each IWAR session. As described earlier, each Board was led by an FSET, assisted by two senior 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 initially know, aside from the Fleet ITs, which participants came from which of the two remaining groups. As in Assessment Four, differences between sailors from the two groups became evident as the interviews progressed.

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[a]/IP settings must be configured automatically.	0—Design not functional or complete 1—Clients using APIPA[a] addressing 3—Some (not all) clients have DHCP[a] IP[a] 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[a] not functional 3—Forward lookup configured properly, reverse hookup not functional 5—Forward and reverse lookups configured correctly

[a] APIPA (Automatic Private IP Addressing); DHCP (Dynamic Host Configuration Protocol); DNS (Domain Name System); (IP) Internet Protocol; TCP (Transmission Control Protocol).

Figure 3. Example Objectives and Scoring for a System Design and Development Exercise.

Each participant was examined on a 0–5 scale with regard to the following six core topics: Networking, Workstations, Domain Controllers, Domain Name System, Disk Management, and Exchange. The interview began with a common question, but was free to proceed after that as the board members chose. Participants who demonstrated effectively no knowledge of a topic were assign 0 points; any participant who demonstrated more knowledge than that possessed by members of the Board was awarded 5 points. (A DT student with no prior IT experience received three 5s. He was

the only participant to receive any.) Each of the 3 Board members scored each participant so that a total of 90 points could be awarded for the 6 topics.

Each Board member also assigned scores on a 0–5 scale for the level of Satisfaction a manager might have with the participant's likely performance on an IT team. Finally, each Board member assigned scores on a 0–5 scale based on the participant's demonstrated Confidence in his or her own IT capabilities. These scores were assigned by the three Board members separately and then added so the maximum score any student could receive for either Satisfaction or Confidence was 15.

8. IT Knowledge Test

The IT Knowledge Test consisted of three parts. All participants completed the Knowledge Test before engaging in IWAR 2. The IWAR 2 Knowledge Test primarily consisted of short-answer questions assembled from items prepared from a variety of sources, including IDA, CID instructors, and the Digital Tutor developer.

With minor copy edits to 5 items, Part 1 (74 items) and Part 2 (65 items) were effectively the same as Parts 1 and 2 of the Knowledge Test administered in IWAR 1. Part 3 (133 items) included items added to cover in more detail topics addressed in Parts 1 and 2, thereby creating redundancy in some topics while adding other topics not tested earlier. Generally, items were worth one point, but all three parts included a number of two-point questions.

As in previous assessments, all test items were vetted for their central relevance to Navy IT assignments and appropriateness for Navy ITs. Although vetting was performed at different times for Parts 1 and 2 and later for Part 3, it was the same for all three parts. First, all items were examined in detail by members of the IDA professional research staff specializing in IT. Not all items were accepted from all sources, and others were substantially edited. Second, the items were reviewed by IDA technical support IT staff members who took the test and provided feedback. Third, the items were similarly vetted by the Navy Network Warfare Command before being assembled by IDA into the final version of the test.

All three parts were administered under "closed-book, closed-notes" conditions. The test was intended to be sufficiently difficult to avoid ceiling effects (too many scores near the maximum) or floor effects (too many scores near zero). 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—in the parlance of instructional testing, it was a "power test" rather than a "speed test."

Table 6 shows the topics and number of items assessed in the IWAR 2 Knowledge Test along with their maximum score. DT and ITTC students took the Knowledge Test

on 12 March 2012, before IWAR 2, in separate rooms at CID, Pensacola. The test was taken by 38 of the 39 DT students (including the IWAR 2 participants) who graduated from classes ending 2 and 9 March 2012 and by the ITTC students who graduated on 2 March and were to participate in IWAR 2. Eight of the 12 Fleet ITs took the Knowledge Test on 20 March, the week before IWAR 2 began. The remaining four Fleet ITs took the test on 26 March when the first IWAR 2 session began. These four were all scheduled to participate in the second session.

After IWAR 2, the remaining 16 ITTC students who had not participated in IWAR 2 and a group of 16 CID instructors took the Knowledge Test. Their scores are included in some of the Knowledge Test results reported here.

Table 6. Knowledge Test Topics, Number of Items, and Points Possible for Each Part and Their Totals.

Topic	Items	Points
Part 1		
PC Hardware	10	11
Client Support Fundamentals	15	19
Windows Server Fundamentals	17	17
Windows DNS (Domain Name System) Server	10	10
Active Directory	12	13
Exchange	10	16
Part 1 Subtotals	74	86
Part 2		
Group Policy	12	17
CISCO IOS (Internet Operating System)	11	15
OSPF (Open Shortest Path First) Protocol	11	21
Switching	12	17
UNIX Operating System	9	9
Security	10	17
Part 2 Subtotals	65	96
Part 3		
Hardware	22	24
Number Systems	4	4
Internet Protocol	19	27
Routers and Routing	10	19
OSPF	7	8
Windows Operating Systems	19	20
Windows Permissions	4	4
Exchange Server	8	10
Group Policy	6	8

Topic	Items	Points
The OSI (Open Systems Interconnection) Model	2	2
DNS (Domain Name Server)	11	14
Domains	5	7
Windows Server—Printing	1	1
DHCP (Dynamic Host Configuration Protocol)	5	6
Active Directory	3	5
Server Management	5	8
Others	2	2
Part 3 Subtotals	133	169
Overall Grand Totals	272	351

B. IWAR 2 Results

1. Troubleshooting Exercise

Data in this section are based on efforts to solve 140 troubleshooting problems, the maximum attempted in the exercise. Data collected addressed three issues: quality of problem solution, harmful changes that were left in the system after solving the problem, and unnecessary steps taken while solving the problem. Problems varied in difficulty and were assigned to five levels of difficulty ranging from "very easy" to "very hard." Table 7 gives brief descriptions of these difficulty levels and their frequency among the 140 problems attempted.

Table 7. Description and Frequency Distribution of Troubleshooting Problem Difficulty Levels.

Level of Difficulty	Description	Number of Problems
1-Very Easy	Solved by the average "Power User."	**10**
2-Easy	Solved by the average IT technician.	**26**
3-Average	Solved by an average network administrator.	**59**
4-Hard	Solved only by experienced network administrators.	**40**
5-Very Hard	Solved only by seasoned IT professionals.	**5**

The trouble tickets used in all Digital Tutor Troubleshooting Exercises were drawn from problems that could not be solved locally (e.g., aboard ship) and were reported to the FSET for assistance. An initial determination of problem difficulty could then be made from the ensuing e-mail traffic. Final determination was made from a review of a participant team's solution path modulated by the experience of the reviewing team.

The problems were presented to three-member teams, four drawn from each of the three groups of participants: DT, Fleet, and ITTC. Two teams from each group

participated in each of the two IWAR sessions. As shown in Table 8, the quality of solution was rated on a 0–5 point scale. A problem was considered correctly solved if its solution was rated four or higher.

Table 8. Scoring for Troubleshooting Problems.

Score	Description
5	Solved as described in the White Team's instructions or deemed equal in quality.
4	Solved in accord with the 5-point standard, but omits items such as documentation or full implementation.
3	A weak solution with specific reasons why it is substandard (e.g., a work-around that would require later upgrading).
2	A solution that relieves the symptom but does not solve the underlying problem.
1	An attempt that does not solve the problem.
0	No attempt.

The final score for the team attempting to solve the problem was determined by consensus among the three White Team members scoring the attempt.

Errors were similarly coded, but in two categories. White Team examiners distinguished between unnecessary steps made during an attempted solution and harmful changes left in the system after (correctly or incorrectly) finding a solution. A significant number of Fleet trouble tickets arise from harmful changes made by IT technicians attempting to solve other problems.

When a team made multiple harmful changes, only the most severe change was recorded. The severity level of each change was scored by the White Team based on the skill level needed to find and correct it (see Table 7).

Taking unnecessary steps to solve a problem is a much less serious matter than taking harmful actions. Unnecessary-step scores measure the efficiency and precision of teams solving a problem. These scores were therefore simply tallies of unnecessary steps taken in a solution attempt. The maximum score for unnecessary steps for a single problem was held to 5, even if more than five unnecessary steps were taken.

2. Successful Solutions to Troubleshooting Problems

Figure 4 shows the number of attempts and successful solutions for troubleshooting problems. A success was defined as receiving a score of 4 or 5 on the problem. The data indicate that DT teams attempted a total of 140 problems and successfully solved 104 of them, or 74 percent of problems attempted. Fleet teams attempted 101 problems and successfully solved 52 (51 percent) of them. ITTC teams attempted 87 problems and successfully solved 33 (38 percent) of them. In brief, the DT teams attempted and solved more problems with a substantially higher probability of success than the other teams.

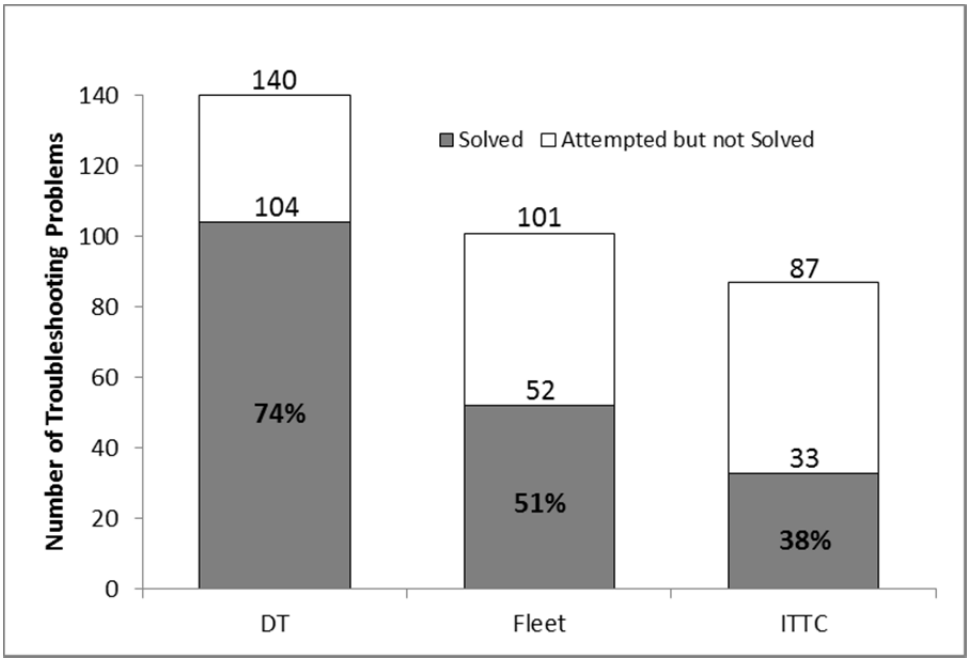


Figure 4. Troubleshooting Problems Attempted and Solved by DT, Fleet, and ITTC Teams.

DT teams received a total of 529 points over the 140 solved problems. As shown in Figure 5, they averaged a Troubleshooting score of 3.78 (1.91). The Fleet teams received a total of 280 points over the 140 problems, averaging a score of 2.00 (2.26), and the ITTC teams received a total of 197 points, averaging 1.41 (2.09) points. In brief, the White Team consistently awarded substantially higher solution quality scores to the DT teams than to the other teams.

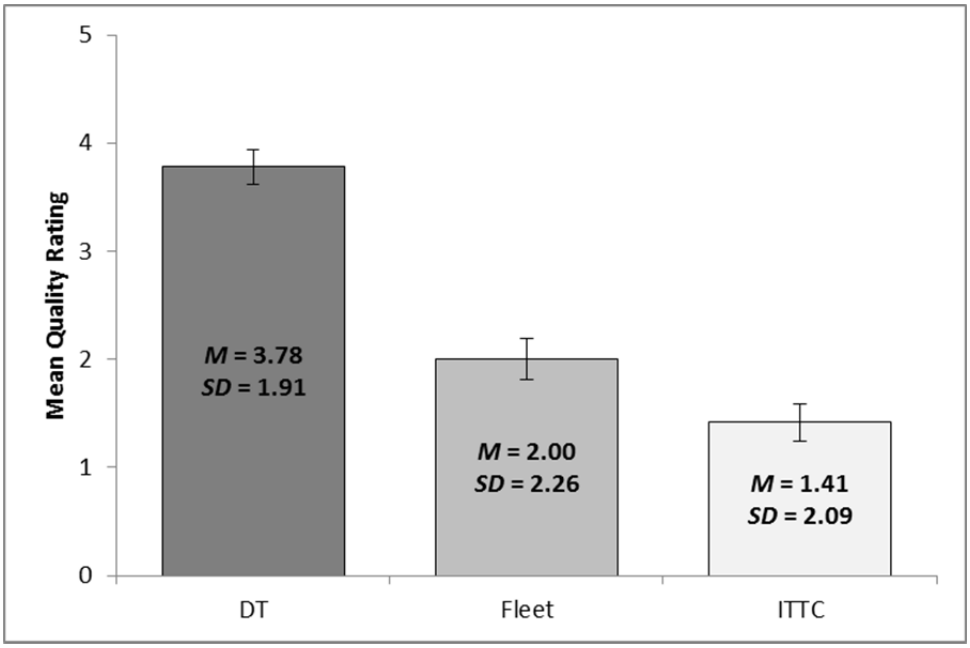


Figure 5. Mean Quality Ratings for Troubleshooting Problem Solving Received by DT, Fleet, and ITTC Teams. (Error bars for this figure and later ones denote standard errors of the mean.)

Table 9 shows results from independent-group *t*-tests and effect sizes for the three pairwise contrasts suggested by Figure 5. All these contrasts are statistically significant. Pairwise effect sizes for the DT teams would be characterized as "large." The contrast between ITTC and Fleet teams would be characterized as "small."

Table 9. Pairwise Contrasts for Mean Total Scores of IWAR 2 Troubleshooting Teams[4]

	M_{diff}	*t*	*d*
DT vs. Fleet	1.78	7.12[b]	0.85
DT vs. ITTC	2.37	9.47[b]	1.13
ITTC vs. Fleet	–0.59	–2.34[a]	–0.28

[a] $p < .05$, [b] $p < .0001$

Figure 6 displays problem attempts and solutions by the three groups of IWAR participants arranged by difficulty. It shows that the DT teams attempted and correctly solved more difficult troubleshooting problems than did either the Fleet or the ITTC teams and that they solved larger proportions of these problems. They were the only

[4] In this and similar tables. M_{diff} refers to a difference in means, *t* is a t-score statistic, *d* is Cohen's *d* effect size, $p < .05$ signifies a statistically significant difference that could occur less than 5 times out of 100, and $p < .0001$ signifies a statistically significant difference that could occur less than 1 time out of 10,000.

group that attempted to solve "very hard" problems, solving three of the five problems in this category. Further, DT teams correctly solved 65 percent of the "hard" problems they attempted, compared with 33 percent for Fleet teams and 17 percent for ITTC teams. Similar results were obtained for problems of "average," "easy," and "very easy" difficulty, with the DT teams solving larger proportions of problems at each level of difficulty.

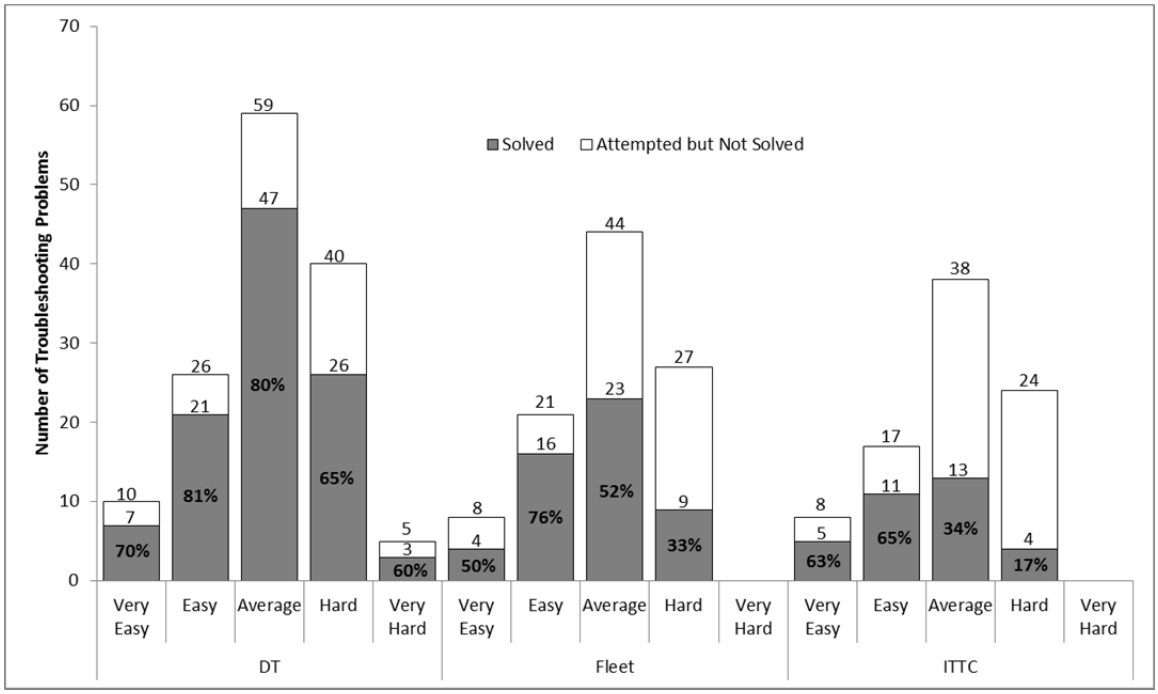


Figure 6. Troubleshooting Problems Attempted and Solved by Difficulty Level.

The correlation between item difficulty and solution quality score was effectively zero for the DT teams. Similar correlations for the Fleet and ITTC teams were small and negative (more difficult, lower score), but large enough to be statistically significant. The quality of solutions by the Fleet and ITTC teams degraded slightly as problem difficulty increased.

Assuming that problem difficulty is rated on an interval scale, a quality ratings (0–5) for a troubleshooting problem can be multiplied by its difficulty rating (1–5) to yield a weighted score. In this analysis, DT teams were awarded an average of 11.18 (6.86) points on each problem compared with an average of 5.35 (6.62) points for Fleet teams and an average of 3.64 (5.72) points for ITTC teams. As shown in Table 10, these results are similar to those obtained for unweighted scores (see Table 9).

Table 10. Pairwise Contrasts for Mean Scores Weighted for Difficulty on Troubleshooting Problems.

Contrast	M_{diff}	*t*	*d*
DT vs. Fleet	5.83	7.13[a]	0.86
DT vs. ITTC	7.54	9.73[a]	1.19
ITTC vs. Fleet	–1.71	–2.31[b]	–0.28

[a]$p < .001$; [b]$p < .05$

In summary, DT 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 more difficult problems.

3. Harmful Errors and Unnecessary Solution Steps

Two types of problems in IT troubleshooting were examined in IWAR 2—harmful changes made and left uncorrected in solving a problem and unnecessary steps taken. A significant number of Fleet trouble tickets arise from harmful changes made during troubleshooting while correcting other problems. A goal of the Digital Tutor is to reduce their frequency. Occurrence of these errors was therefore recorded. Another goal of the Digital Tutor is to produce efficiency in troubleshooting and problem solving. Unnecessary steps, which are benign relative to harm done to IT systems during troubleshooting, were also recorded. Both types of errors are examined in this section.

a. Harmful Changes

Figure 7 depicts the number (left-side panel) and proportion (right-side panel) of harmful changes made by the three IWAR 2 groups. In attempting to solve a problem, DT teams made fewer harmful changes than either Fleet or ITTC teams (20 versus 41 and 29, respectively). Further, the probability that a DT team would make a harmful change during problem solving (right-hand panel) was less than half that of Fleet and ITTC teams—0.14 versus 0.41 and 0.33, respectively, based on the number of problem solutions attempted by the DT, Fleet, and ITTC teams.

Table 11 shows results from analyses of the harmful changes proportions. These results indicate that the proportions for DT teams were statistically significantly lower than those for either Fleet or ITTC teams. The effect sizes of –0.61 (DT compared with Fleet teams) and –0.44 (DT compared with ITTC teams) would be characterized as "medium." The Fleet and ITTC team error rates did not differ significantly from each other, and the effect size for the difference in their rates would be characterized as "small."

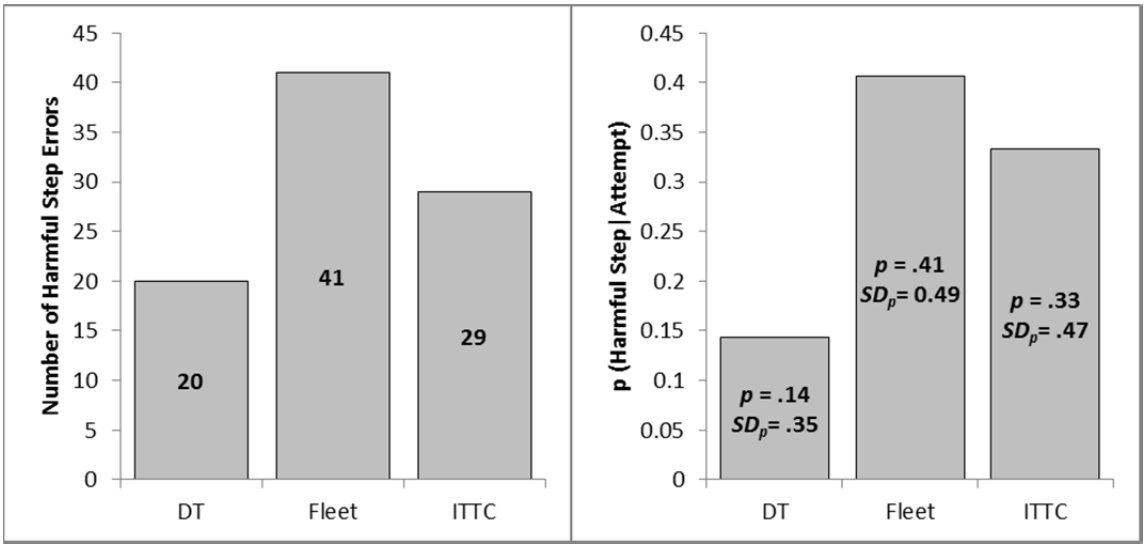


Figure 7. Number (Left-Hand Panel) and Proportions with Standard Deviations (Right-Hand Panel) of Harmful Changes Made by the IWAR Teams.

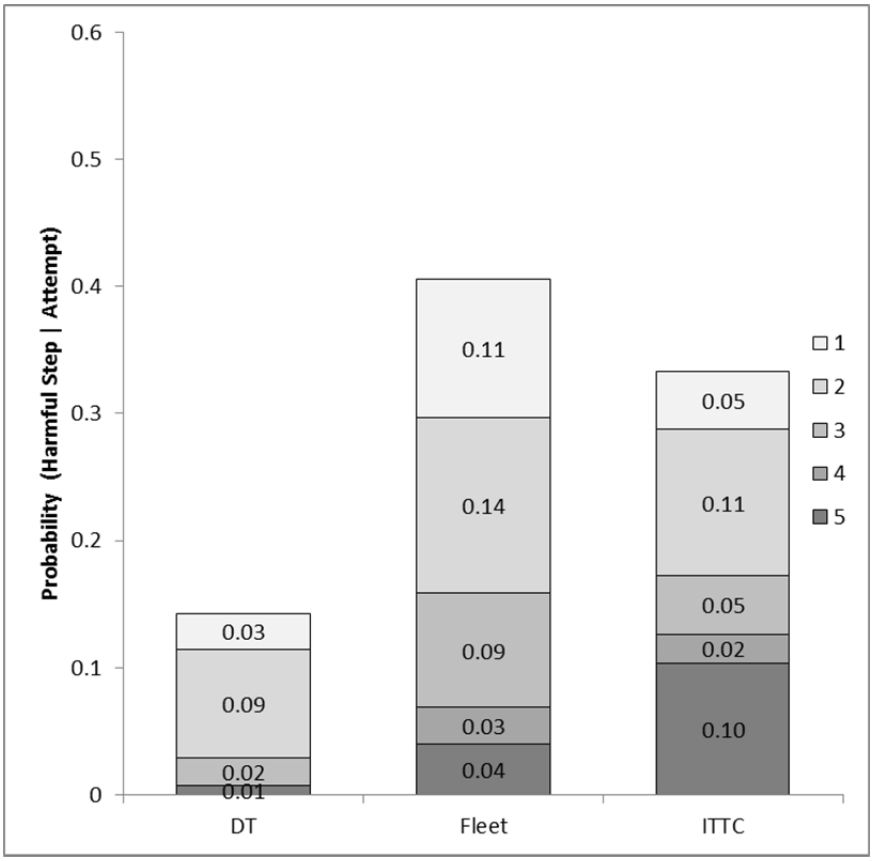
Table 11. Results from Unweighted Pairwise Comparisons of Harmful Action Rates.

Contrast	M_{diff}	*t*	*d*
Harmful Actions			
DT vs. Fleet	–0.26	–4.68[b]	–0.61
DT vs. ITTC	–0.19	–3.24[a]	–0.44
ITTC vs. Fleet	–0.07	–1.15	–0.17

[a]$p < .01$, [b]$p < .0001$

Analogous to the weighting for solution score difficulty, the rates of harmful changes were weighted by degree of severity. The weightings were based on the level of IT ability needed to find and correct a harmful change (see Table 7). They ranged from 1 (least severe) to 5 (most severe). Figure 8 shows the proportions of harmful changes made by the three IWAR 2 groups at each level of severity. It indicates that DT teams made fewer harmful changes at each level of severity than those of Fleet and ITTC teams. One out of 100 problem attempts by the DT teams included a severely harmful action (rated 4 or 5) compared with 1 out of 14 for Fleet teams and 1 out of 8 for ITTC teams.

There appears to be no established procedure for the IWAR assessment to take account of solution quality, problem difficulty, and severity of harm captured in a single composite score. A descriptive statistic was devised for IWAR by assigning problem difficulty level as a score for problems solved correctly (quality score of 4 or 5) and then subtracting from that the severity score of any harmful error that might have been made and left uncorrected. Under this procedure, teams could receive negative scores if they made severely harmful errors while solving relatively easy problems.



NB: The DT teams made no harmful changes of level 4 severity.

Figure 8. Probabilities at Each Severity Level of Harmful Changes Made by the Three IWAR Groups During Problem Attempts.

Figure 9 displays the total composite score points calculated in this manner for the three IWAR groups. There were 267 points awarded to the DT teams, 43 points awarded to the Fleet teams, and –7 points awarded to the ITTC teams. To assess the variability between problem solutions, the mean total score per problem was calculated by dividing these total points by the number of problems. A one-way ANOVA indicated that the differences between these group means were statistically significant: F (2, 417) = 41.229, MSE = 3.689, $p < .0001$,[5] but inferential statistics based on this metric should be viewed with caution.

[5] This notation signifies an F-ratio of 61.59 with 2 and 23 degrees of freedom, a mean squared error of 957.56, and a probability of occurring by chance less than once in 10,000.

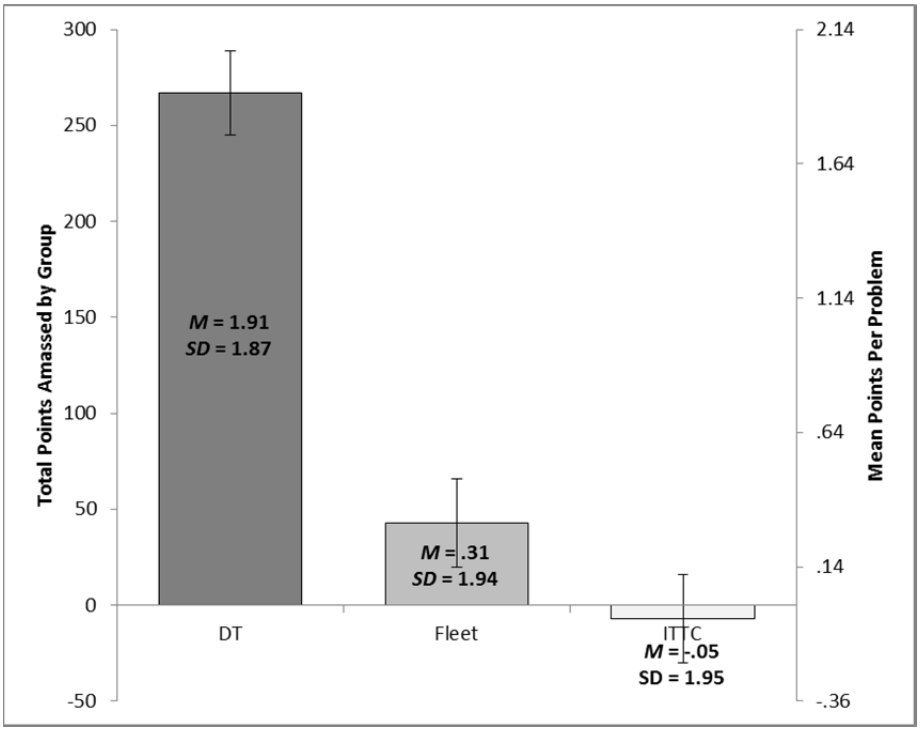


Figure 9. Means and Standard Deviations of Composite Score Totals for Three IWAR Groups. Error Bars are Standard Errors of the Mean.

Data concerning harmful changes help to explain the positive comments of White team members in favor of DT students and considerable concern about the likely actions Fleet and ITTC students might make in IT problem solving. The monetary and operational consequences of these errors have not been quantified, but they are likely to be considerable.

b. Unnecessary Steps

Figure 10 provides an indication of problem-solving efficiency by the three groups. It depicts the number (left-hand panel) of solution attempts with one or more unnecessary steps and proportion (right-hand panel) of attempts with unnecessary steps taken to solve troubleshooting problems. As shown, there were fewer attempts in which DT teams took unnecessary steps than those found for either Fleet or ITTC teams (36 versus 52 and 48 steps, respectively). Further, the proportion (number of solution attempts with at least one unnecessary step divided by total attempts) found for the DT teams was about half that for Fleet and ITTC teams (0.26 versus 0.51 and 0.55, respectively).

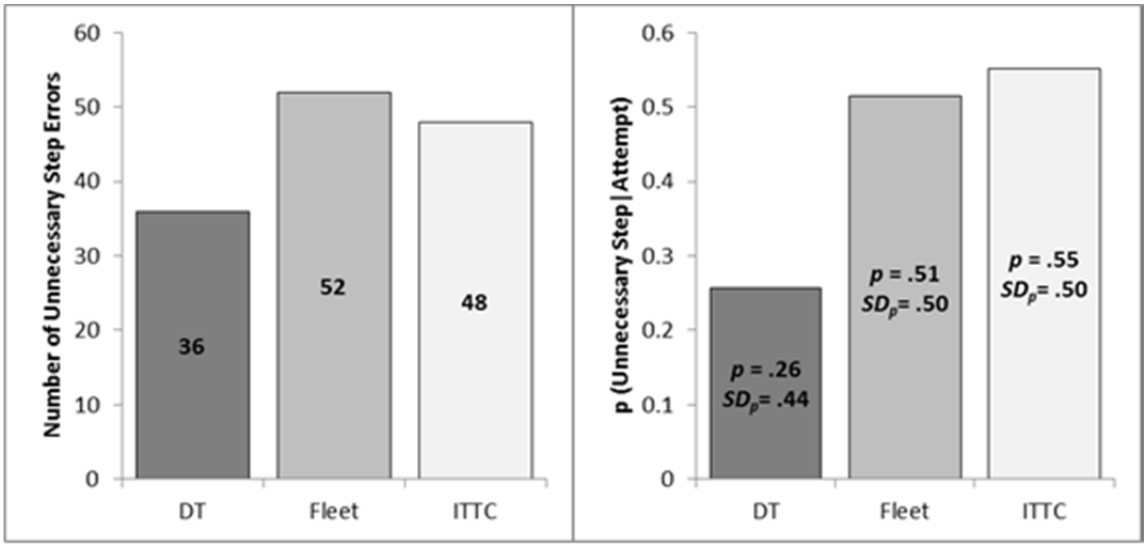


Figure 10. Number (Left-Hand Panel) and Proportions (Right-Hand Panel) of Unnecessary Steps Made by the Three IWAR Groups.

Table 12 shows results from analyses of the unnecessary-step proportions as shown in Figure 10. These results found that the proportions for DT teams were statistically significantly lower than those for either Fleet or ITTC teams. The effect sizes for the contrast between DT and the comparison groups were both "medium" at –0.54 and –0.62, respectively. Proportions for unnecessary steps taken by the Fleet and ITTC teams did not differ significantly from each other and show an effect size that would have been characterized as "negligible."

Table 12. Results from Pairwise Contrasts of Unnecessary Step Rates.

Contrast	M_{diff}	*t*	*d*
Unnecessary Steps			
DT vs. Fleet	–0.26	–4.17[a]	–0.54
DT vs. ITTC	–0.29	–4.56[a]	–0.62
ITTC vs. Fleet	0.04	0.53	0.08

[a] $p < .0001$

Figure 11 shows the average number (and standard deviation) of unnecessary steps made by the three groups per problem-solving attempt. With an average of 0.48 unnecessary steps, DT teams averaged about a third of the unnecessary steps taken by Fleet (1.24) and ITTC (1.43) teams per problem solution. Efficiency in solving IT problems may not be, as suggested above, as serious a matter as harmful errors likely to spawn additional trouble tickets, but it remains of concern, especially given the exigencies that are likely to arise in military operations—along with the increased

possibility of harmful changes created while "Easter-egging"—randomly searching for a problem solution.

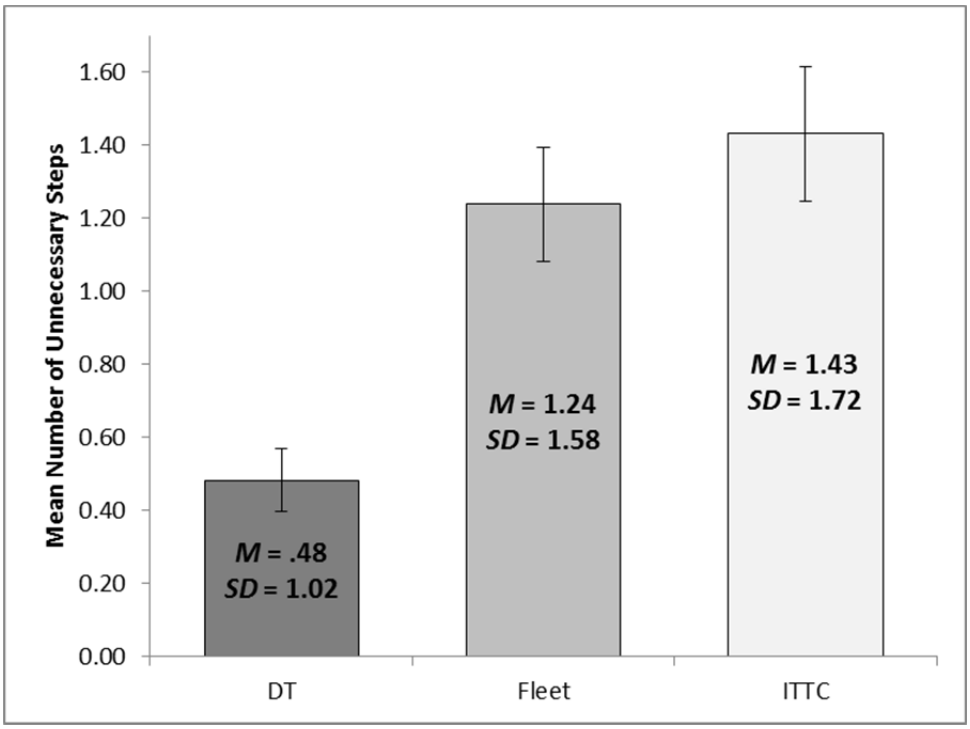


Figure 11. Average Number of Unnecessary Steps Taken Per Problem-Solving Attempt by the Three IWAR Groups.

Table 13 shows results from pairwise analyses of the unnecessary-step means shown in Figure 11. These results suggest that the mean number of unnecessary steps per troubleshooting problem solution attempt were statistically significantly lower for DT teams than for either Fleet or ITTC teams. The effect sizes for the contrast between DT and the comparison groups were both "medium" at –0.54 and –0.67, respectively. The unnecessary-step averages for the Fleet and ITTC teams did not differ significantly from each other and show an effect size (0.14) that would have been characterized as "negligible."

Table 13. Results from Pairwise Contrasts of Unnecessary Step Means.

Contrast	M_{diff}	*t*	*d*
DT vs. Fleet	-0.76	-4.10[a]	-0.54
DT vs. ITTC	-0.95	-4.92[a]	-0.67
ITTC vs. Fleet	0.19	0.94	0.14

[a] $p < .0001$

In sum, the DT teams were less likely to make either harmful changes or take unnecessary steps in troubleshooting than either the Fleet or ITTC teams. Moreover, the severity of harmful changes and frequency of unnecessary steps made by DT teams was found to be statistically significantly lower than those of either the Fleet or ITTC teams.

4. Review Board Interviews

Performance in the Review Board interview was rated on a 6-point scale (0 for no knowledge of the topic and 5 for more knowledge of the topic than the Board members). Each IWAR participant was assessed by three judges consisting of an FSET assisted by two senior Navy ITs. Their median rating was recorded as the Review Board score. Figure 12 displays means and standard deviations of the Review Board scores.

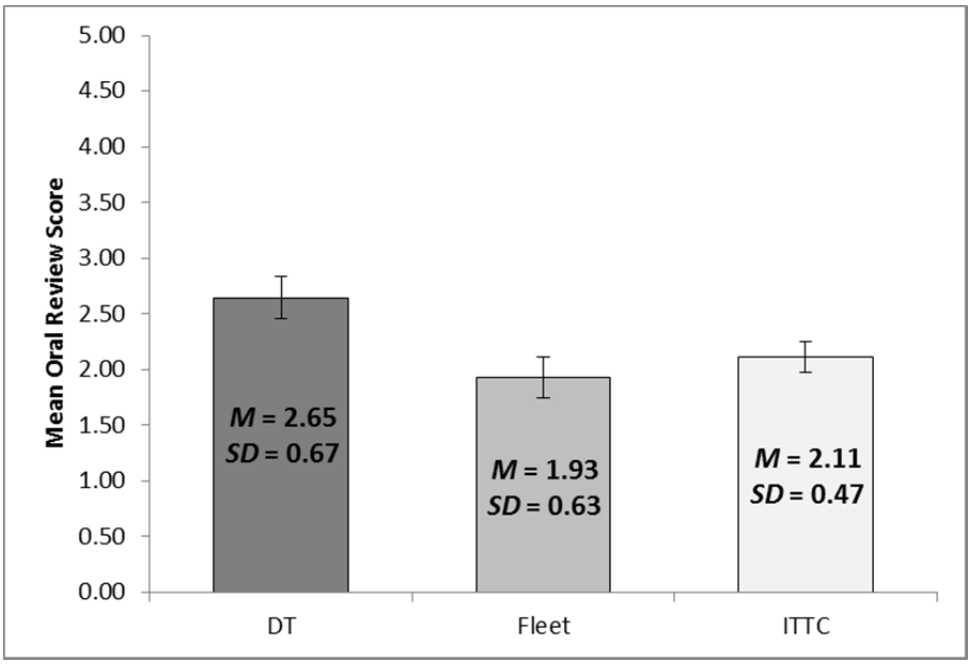


Figure 12. Means and Standard Deviations of Review Board Scores for Three IWAR Groups.

Although the mean differences (Figure 12) between groups are small, the variability in Review Board scores is also small. As a result (Table 14), the differences between DT and both Fleet and ITTC participants are statistically significant. Their effect sizes are "very large" and "large," respectively. The differences between Fleet and ITTC participants are not statistically significant. The effect size for their means would be characterized as "negligible."

Table 14. Results from Pairwise Contrasts on Review Board Scores.

Contrast	M_{diff}	*t*	*d*
DT vs. Fleet	0.72	2.83[b]	1.20
DT vs. ITTC	0.54	2.11[a]	0.90
ITTC vs. Fleet	0.18	0.72	0.31

[a]$p < .05$, [b]$p < .01$

5. Security Exercise

IWAR participants in the security exercise were organized into the same three-person teams that were used in the troubleshooting exercise. As in the Troubleshooting Exercise, differences were assessed on group means calculated from four data points (four teams for each group). Six teams were tested in each week's session using two forms of the security exercise that differed in length and content. 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 the total points possible in the session.

Figure 13 shows the mean percentage of total points on the security exercise for the three IWAR groups. The Fleet teams outperformed the other two groups. Because of the small number of data points, however, a one-way ANOVA of the means indicated no statistically significant differences among means: $F(2, 9) = 2.181$, $MSE = .037$, $p > .10$.

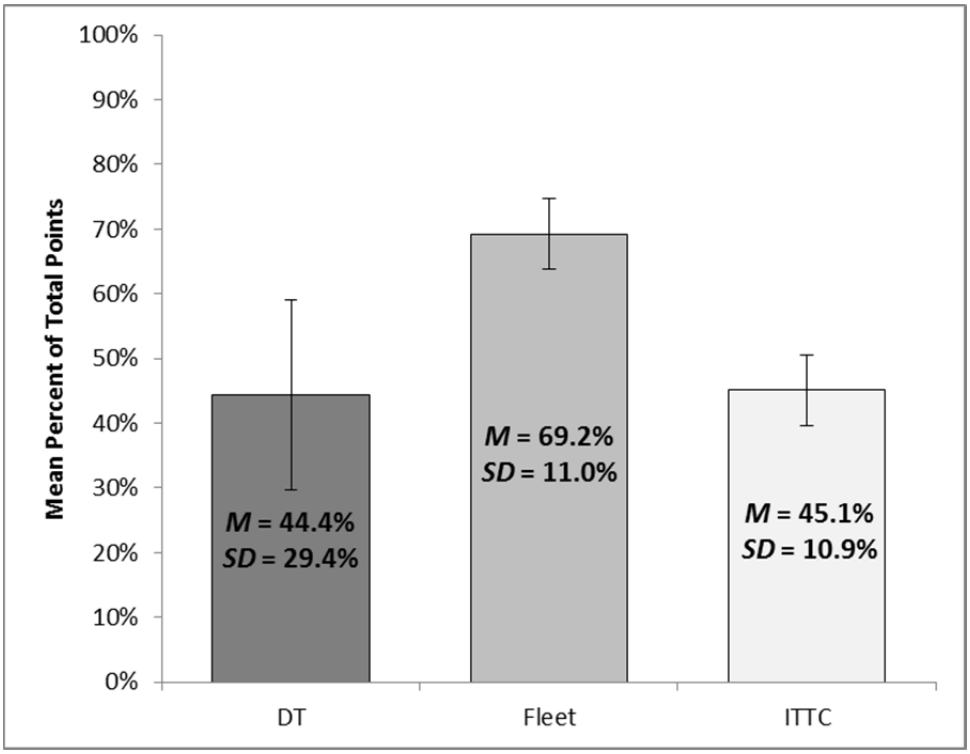


Figure 13. Means and Standard Deviations of Percentage of Total Points on Security Exercise for Three IWAR Groups.

The analyses summarized in Table 15 indicate that although none of the three contrasts is statistically significant, the effect sizes for contrasts between the DT and Fleet teams and the ITTC and Fleet teams are substantial. Both effect sizes favor the performance of the Fleet teams and both may be characterized as "very large."

Table 15. Results from Pairwise Contrasts on Security Exercise Scores.

Contrast	M_{diff}	*t*	*d*
DT vs. Fleet	–24.8%	–1.83	–1.30
DT vs. ITTC	–0.6%	–0.05	–0.03
ITTC vs. Fleet	–24.1%	–1.78	–1.26

As in Assessment Four performed in November 2010, the Security test results somewhat validate the approach used in developing the Digital Tutor because of the unavailability of an expert human tutor model for the Security section.

6. Network Design and Development

IWAR participants executed this task as six-person teams. Three teams were tested in each week's session. Team performance was rated with respect to multiple objectives using a 5-point scale. There were 24 objectives (5 critical and 19 secondary) assessed in

the first week's session and 21 objectives (5 critical and 16 secondary) in the second week's session. Figure 14 shows the number of design and development objectives successfully met (scores of 4 or 5) by the three IWAR groups.

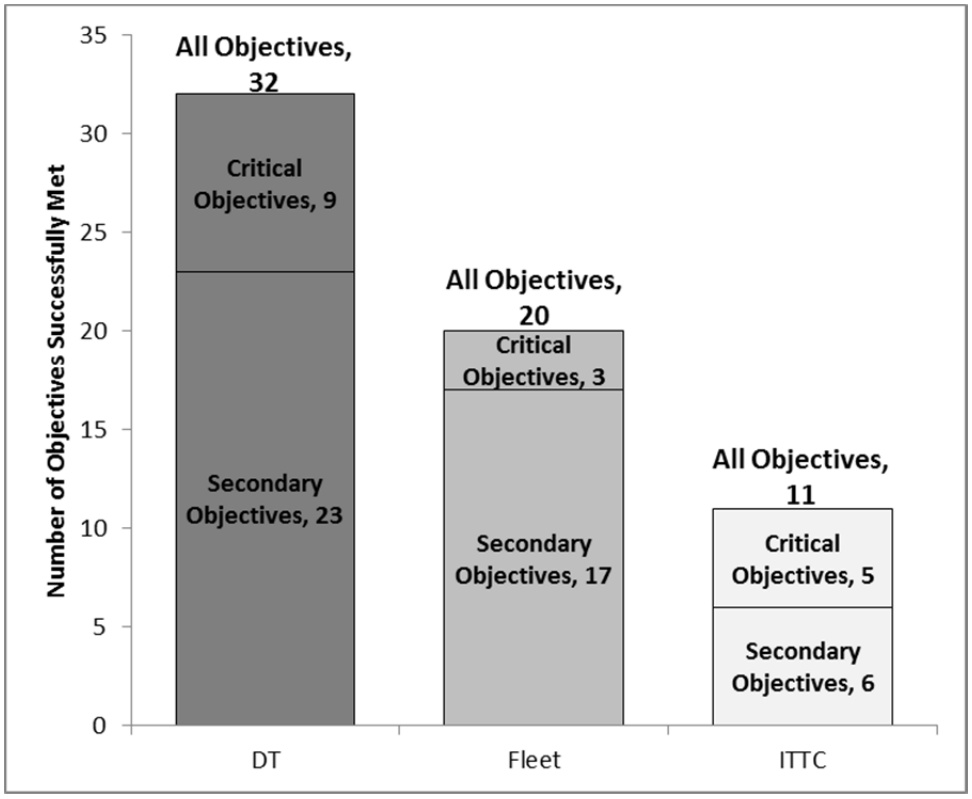


Figure 14. Number of Objectives Successfully Achieved by the Three IWAR Group Teams.

Figure 15 shows the means and standard deviations of total scores obtained by the three groups (one team per each weekly session).

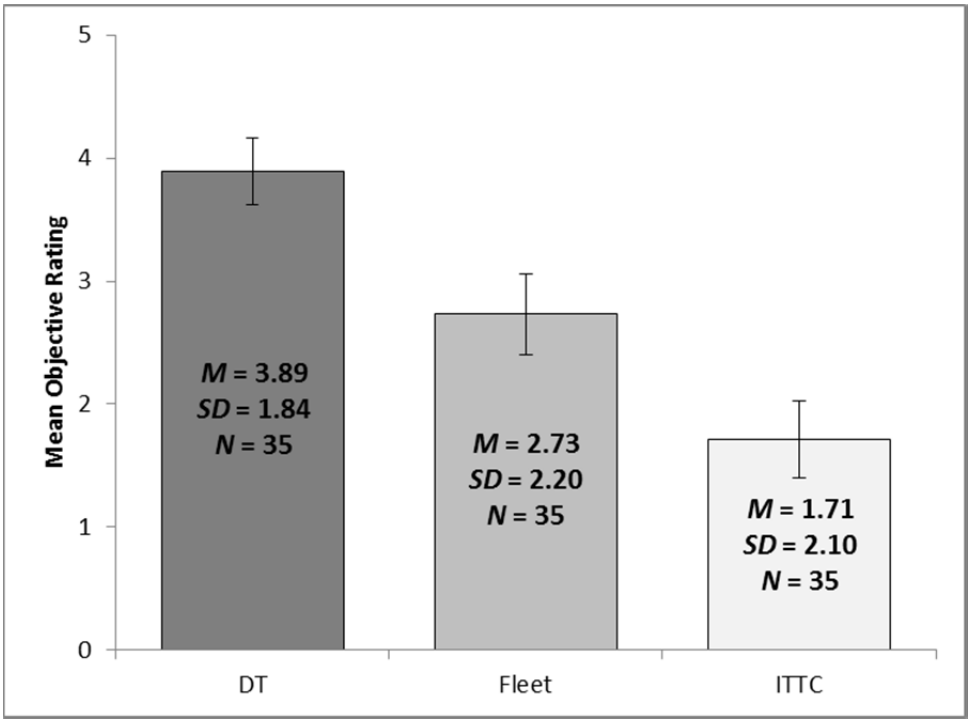


Figure 15. Means and Standard Deviations of Ratings on the Design and Development Objectives.

Analysis of these ratings (Table 16) indicates that the DT teams scored statistically significantly more points than the Fleet and ITTC teams in meeting the Critical Objectives. They also scored statistically significantly more points compared with the ITTC teams for Secondary Objectives but not compared with Fleet teams. These results are reflected in the total scores. Compared to ITTC, the DT total score was statistically significantly higher with a "very large" effect size. Although the DT teams had a "medium" effect size and a larger total score compared with Fleet teams, the score difference between the two was not statistically significant.

In general, but not in every case, the DT teams successfully achieved more of the design objectives and received higher scores than the Fleet and ITTC teams.

Table 16. Results from Pairwise Comparisons of Scores on the Design and Development Exercise.

Contrast	M_{diff}	*t*	*d*
Critical Objectives			
DT vs. Fleet	2.90	4.44[c]	1.51
DT vs. ITTC	1.70	2.60[a]	0.89
ITTC vs. Fleet	1.20	1.84	0.63
Secondary Objectives			
DT vs. Fleet	0.66	1.01	0.34
DT vs. ITTC	2.31	3.54[c]	1.21
ITTC vs. Fleet	–1.66	–2.54[b]	–0.86
All Objectives			
DT vs. Fleet	1.16	1.77	0.60
DT vs. ITTC	2.18	3.34[b]	1.14
ITTC vs. Fleet	–1.02	–1.57	–0.53

[a] $p < .05$, [b] $p < .01$, [c] $p < .001$

7. Knowledge Test

Data in this section are based on individual participant scores on the Knowledge Test. The Knowledge Test for IWAR 2 was divided into three parts. The total test comprised 272 items worth 349 points. The items were organized into 29 separate topics that referred to specific content areas.

a. Total Score

Figure 16 displays the means and standard deviations of scores on the Knowledge Test for the three groups of participants. As the figure shows, the DT group scored considerably higher than either the Fleet or ITTC group. The ITTC group scored somewhat higher than the Fleet group, although the difference is smaller. The overall between-group difference was significant: $F(2,33) = 61.59$, $MSE = 957.56$, $p < .0001$.

The Fleet ITs were at some disadvantage on the Knowledge Test. The test assesses current knowledge of considerable breadth and depth. Although the Fleet ITs attend "C" schools and receive sustainment training and technical updates, their duties may prevent them from receiving current information in a timely manner. Also, the knowledge they receive from these sources may be limited to specific duty station requirements without covering the wide spectrum of IT issues targeted by the Knowledge Test.

As a result, Fleet ITs may be expert in some topics, but rusty in others that they have not visited since their "A" school training. Working in teams, as required by the practical exercises, should mitigate limitations arising from such specialized experience,

and that possibility is supported by findings from the practical exercises. But the Fleet ITs were on their own when taking the Knowledge Test.

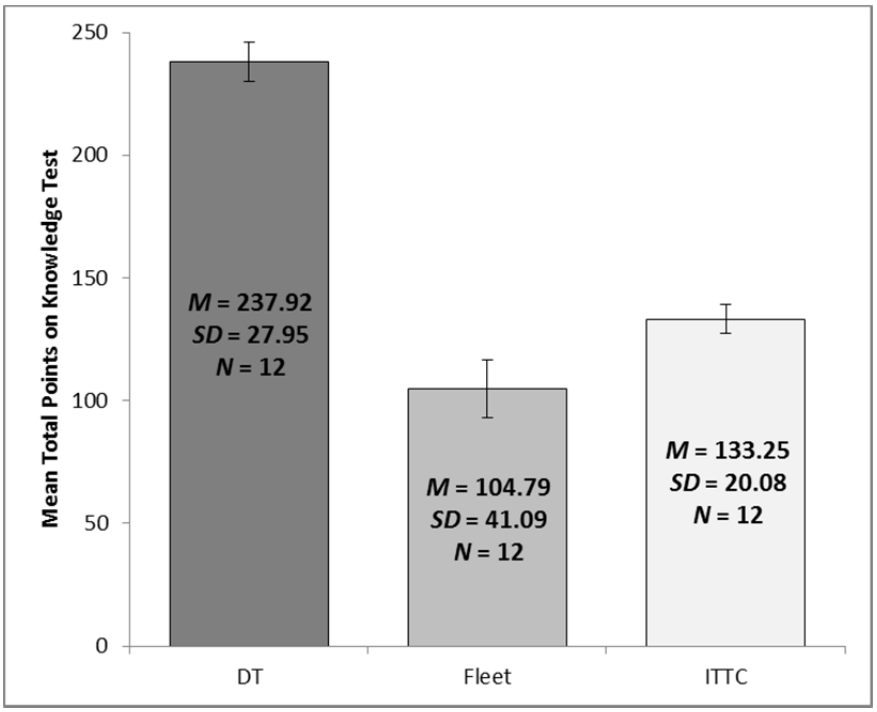


Figure 16. Means and Standard Deviations on Knowledge Test.

Table 17 indicates that all three contrasts in Figure 16 were statistically significant. The effect sizes of the two pairwise contrasts of DT scores with Fleet scores (4.30) and ITTC scores (3.38) would be characterized by the Thalheimer and Cook (2002) rubric as "huge." The effect size (0.92) for the ITTC contrast with Fleet was "large" in accord with that rubric.

Table 17. Results from Pairwise Contrasts on Total Knowledge Scores.

Contrast	M_{diff}	*t*	*d*
DT vs. Fleet	133.13	10.54[b]	4.30
DT vs. ITTC	104.67	8.29[b]	3.38
ITTC vs. Fleet	28.46	2.25[a]	0.92

[a] $p < .05$, [b] $p < .0001$

b. Non-IWAR Groups

In addition to the three groups (DT, Fleet, and ITTC) that participated in the IWAR program, three other comparison groups took the Knowledge Test:

- 26 Students who received instruction on the digital tutor, but did not participate in IWAR 2 ("DT Non-IWAR").
- 17 Students who participated in the ITTC instructional program, but did not participate in IWAR 2 ("ITTC Non-IWAR").
- 16 CID School instructors.

Figure 17 shows means and standard deviations for Knowledge Test scores awarded to all six groups.

The results indicate that both of the DT groups scored higher than all other comparison groups, with the DT IWAR group scoring highest on the Knowledge Test. In contrast, the ITTC IWAR, ITTC Non-IWAR, and Instructor scores differed only slightly. Comparisons of Knowledge Test mean scores using the Tukey b post hoc procedure (α = .05) revealed three homogeneous subsets:

- DT IWAR students, who scored higher than the other six groups.
- DT Non-IWAR students, who scored lower than DT IWAR yet higher than participants in the four remaining conditions.
- Fleet, ITTC IWAR, ITTC Non-IWAR, and A School Instructor groups, all of which scored about the same on the Knowledge Test.

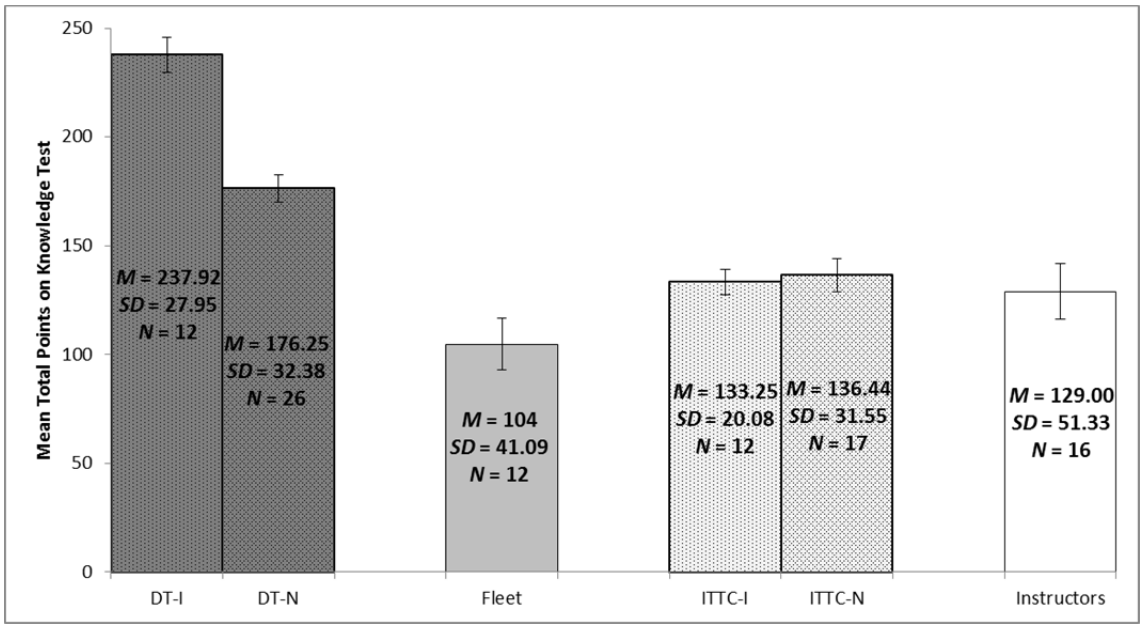


Figure 17. Means and Standard Deviations of Knowledge Test scores for IWAR (I), Non-IWAR (N), Fleet, and CID Instructor Comparison Groups.

Table 18 displays the 15 pairwise comparisons that can be made between these 6 groups.[6] The effect sizes for the DT-IWAR contrasts would all be characterized as "huge"; the DT Non-IWAR contrasts are smaller and mostly "very large." All contrasts involving DT groups are statistically significant.

Table 18. Effect Sizes for All IW ARS and Non-IWARS Groups on Knowledge Test Scores.

Contrast	M_{diff}	*t*	*d*
DT-IWAR vs. DT-Non	61.67	4.95[c]	1.73
DT-IWAR vs. Fleet	133.13	9.14[c]	3.73
DT-IWAR vs. ITTC-IWAR	104.67	7.19[c]	2.93
DT-IWAR vs. ITTC-Non	101.48	7.54[c]	2.84
DT-IWAR vs. Instructors	108.92	8.00[c]	3.05
DT-Non vs. Fleet	71.46	5.74[c]	2.00
DT-Non vs. ITTC-IWAR	43.00	3.45[b]	1.21
DT-Non vs. ITTC-Non	39.81	3.58[b]	1.12
DT-Non vs. Instructors	47.25	4.17[b]	1.32
ITTC-IWAR vs. ITTC-Non	–3.19	–0.24	–0.09
ITTC-IWAR vs. Fleet	28.46	1.95	0.80
ITTC-IWAR vs. Instructors	4.25	0.31	0.12
ITTC-Non vs. Fleet	31.65	2.35[a]	0.89
ITTC-Non vs. Instructors	7.44	0.60	0.21
Instructors vs. Fleet	24.21	1.78	0.68

[a]$p < .05$, [b]$p < .001$, [c]$p < .0001$

C. IWAR 2 Summary

IWAR 2 provides various findings regarding the effectiveness of DT relative to Fleet and ITTC conditions. Table 19 summarizes the findings and suggests two patterns that were repeated across the different performance measures:

- With the exception of the Security Exercise, DT participants outperformed both the Fleet and ITTC teams. The effect sizes are mostly "large," "very large," and even "huge" in accord with the Thalmeier & Cook (2002) rubric.
- Differences between Fleet and ITTC participants were generally much smaller and neither consistently positive or negative.

[6] Three of the contrasts (DT-IWAR vs. Fleet, DT-IWAR vs. ITTC-IWAR, and ITTC-IWAR vs. Fleet) repeat contrasts made earlier. The effect sizes here are slightly smaller due to increased variability from inclusion of Non-IWAR groups. The overall findings remain the same.

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

Table 19. Summary of Results from IWAR 2.

Performance Measure	Direction[a]	Significance[b]	Effect Size[c]
DT versus Fleet			
Problem Solving (PS) Total Score	+	< .0001	Large
PS (Fewer) Harmful Actions	+	< .0001	Medium
PS (Fewer) Unnecessary Actions	+	< .0001	Medium
Oral Review	+	< 0.01	Very Large
Security Exercise	-	N.S.	Very Large
Network Design & Development	+	N.S.	Medium
Knowledge Test Total Score	+	< 0.0001	Huge
DT versus ITTC			
Problem Solving (PS) Total Score	+	< .0001	Large
PS (Fewer) Harmful Actions	+	< 0.01	Medium
PS (Fewer) Unnecessary Actions	+	< .0001	Medium
Oral Review	+	< 0.05	Large
Security Exercise	-	N.S.	Negligible
Network Design & Development	+	< 0.01	Very Large
Knowledge Test Total Score	+	< 0.0001	Huge
ITTC versus Fleet			
Problem Solving (PS) Total Score	-	N.S.	Small
PS (Fewer) Harmful Actions	+	N.S.	Small
PS (Fewer) Unnecessary Actions	+	N.S.	Negligible
Oral Review	+	N.S.	Small
Security Exercise	-	N.S.	Very Large
Network Design & Development	-	N.S.	Small
Knowledge Test Total Score	+	< .05	Large

[a] Signs indicate either consistent (+) or inconsistent (-) with the following hypotheses: DT > Fleet, DT > ITTC, and ITTC > Fleet.

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

[c] Effect size using Thalheimer and Cook (2002) nomenclature.

The exception to this pattern is the Security Exercise where Fleet teams outperformed both DT and ITTC teams by a large margin. Because of the small number of data points in this comparison, the differences could not be judged as statistically

significant. However, this exception emphasizes the importance of selecting experts in both one-on-one tutoring and the subject matter when developing digital tutors modeled on tutoring.

Overall, if the DT students had simply matched Fleet IT performance in the practical exercises, the goals of this program would have been met. It is notable that they generally outscored the Fleet participants by substantial margins. From a training and monetary standpoint it is also notable that in every instance, they outscored ITTC graduates who had been almost twice as long in training as DT students.

4. Additional Analyses

Additional analyses of data were undertaken to address matters of more specific interest.

A. Topic Scores

Are there differences among topics in DT effectiveness?

As shown in Table 20, the Knowledge Test covered 29 separate topics (including 1 identified as "other"). Some topics appear twice because they were included in Part 1 or Part 2 and in Part 3 of the IWAR 1 test. Parts 1 and 2 needed to be identical to those in IWAR 2 to enable the comparison between human (IWAR 1) and digital (IWAR 2) tutoring. Some of these topics were elaborated in Part 3, which was unique to IWAR 2. Table 20 shows statistical significance and effect sizes for the three group contrasts over all topics. On 85 percent of the topics, DT graduates scored statistically significantly ($p <$ .05) higher than Fleet ITs. Similarly, on 76 percent of the topics the DT graduates scored statistically significantly ($p < .05$) higher than the ITTC graduates. The ITTC graduates and the Fleet ITs had statistically significant differences on 31 percent of the topics.

The effect sizes for DT vs. Fleet comparisons and for DT vs. ITTC comparisons varied widely. The two lowest effect sizes are negative. They indicated that DT participants scored lower than both Fleet ($d = –0.81$) and ITTC ($d = –0.80$) participants on the topic of security. Neither of these contrasts was statistically significant. The largest effect sizes concerned Routers and Routing. They were 6.00 for DT vs. Fleet and 4.10 for the DT vs. ITTC.

Effect sizes involving the DT group averaged 2.46 for DT vs. Fleet and 1.98 for DT vs. ITTC. Well over 60 percent of the topics showed "huge" effect sizes for the DT vs. Fleet contrasts (20 of 29), as well as for the DT vs. ITTC contrasts (18 of 29). Effect sizes greater than 3.00 were obtained for the DT vs. Fleet contrast for the following topics:

- Client Support Fundamentals ($d = 3.54$).
- Windows DNS (Domain Name System) Server ($d = 4.03$).
- Active Directory ($d = 3.62$).
- CISCO IOS (Internet Operating System) ($d = 4.86$).
- Routers and Routing ($d = 6.00$).

- OSPF (Open Shortest Path First) Protocol - Part 3 (d = 5.40).
- Windows Operating Systems (d = 3.12).
- DNS (Domain Name Server) (d = 5.50).
- Domains (d = 3.88).
- DHCP (Dynamic Host Configuration Protocol) (d = 3.02)
- Server Management (d = 4.09).

The six topics in which the differences between DT and ITTC graduates were not statistically significant were the following:

- PC Hardware (d =0.74).
- Group Policy (d = 0.31).
- Unix Operating Systems (d = –0.03).
- Security (d = -1.96).
- Windows Permissions (d = 1.16)
- Group Policy (d = 1.21)
- Windows Server (d = 0.00)[7]

Pairwise contrasts between the Fleet and ITTC groups were much smaller, with an average effect size of 1.10. Nearly half the effect sizes (14 of 29) comparing Fleet and ITTC scores on the separate topics were rated as "negligible" (i.e., $d < 0.15$). The lowest effect size between Fleet and ITTC groups was –1.10 for the Number Systems topic, with Fleet scoring statistically higher than ITTC participants. The largest effect size between Fleet and ITTC groups was 2.24 ("huge") for the Open Shortest Path First (OSPF) topic in Knowledge Test Part 3, with Fleet participants scoring statistically higher on that topic than ITTC participants.

[7] There was only one test item for this topic.

Table 20. Statistical Significance and Effect Size (*d*) for Contrasts on Knowledge Test Topics.

	DT vs. Fleet			DT vs. ITTC			ITTC vs. Fleet		
Topic	M_{diff}	*t*	*d*	M_{diff}	*t*	*d*	M_{diff}	*t*	*d*
PC Hardware	3.33	3.61[c]	1.48	1.67	1.81	0.74	1.67	1.81	0.74
Client Support Fundamentals	8.75	8.68[d]	3.54	5.58	5.54[d]	2.26	3.17	3.14[b]	1.28
Windows Server Fundamentals	4.21	4.09[c]	1.67	5.67	5.51[d]	2.25	–1.46	–1.42	–0.58
*Windows DNS (Domain Name System) Server	5.13	9.87[d]	4.03	4.88	9.39[d]	3.83	0.25	0.48	0.20
*Active Directory	7.21	8.87[d]	3.62	6.92	8.51[d]	3.48	0.29	0.36	0.15
*Exchange	4.00	3.05[b]	1.24	7.13	5.42[d]	2.21	–3.13	–2.38[a]	–0.97
*Group Policy	2.21	2.09[a]	0.85	0.33	0.31	0.13	1.88	1.77	0.72
CISCO IOS (Internet Operating System)	9.67	11.90[d]	4.86	6.46	7.95[d]	3.25	3.21	3.95[c]	1.61
*OSPF (Open Shortest Path First) Protocol	10.67	7.27[d]	2.97	7.88	5.37[d]	2.19	2.79	1.90	0.78
Switching	7.46	6.68[d]	2.73	7.33	6.57[d]	2.68	0.13	0.11	0.05
UNIX Operating System	–1.67	–1.39	–0.57	–0.04	–0.03	–0.01	–1.63	–1.35	–0.55
Security	–2.50	–1.99	–0.81	–2.46	–1.96	–0.80	–0.04	–0.03	–0.01
Hardware	6.17	3.81[c]	1.55	5.75	3.55[b]	1.45	0.42	0.26	0.10
Number Systems	2.25	6.04[d]	2.46	3.25	8.72[d]	3.56	–1.00	–2.68[a]	–1.10
Internet Protocol	11.83	7.32[d]	2.99	3.58	2.22[a]	0.91	8.25	5.10[d]	2.08
Routers and Routing	13.42	14.70[d]	6.00	9.17	10.05[d]	4.10	4.25	4.66[d]	1.90
*OSPF (Open Shortest Path First) Protocol	5.42	13.23[d]	5.40	3.17	7.74[d]	3.16	2.25	5.50[d]	2.24
Windows Operating Systems	7.08	7.63[d]	3.12	6.92	7.45[d]	3.04	0.17	0.18	0.07
Windows Permissions	0.67	1.86	0.76	0.42	1.16	0.47	0.25	0.70	0.28
*Exchange Server	1.92	2.86[c]	1.17	1.83	2.74[b]	1.12	0.08	0.12	0.05
*Group Policy	1.42	2.57[b]	1.05	0.67	1.21	0.49	0.75	1.36	0.56
The OSI (Open Systems Interconnection) Model	0.75	2.94[c]	1.20	0.67	2.61[a]	1.07	0.08	0.33	0.13
*DNS (Domain Name Server)	8.17	13.47[d]	5.50	5.67	9.35[d]	3.82	2.50	4.12[c]	1.68
Domains	4.08	9.51[d]	3.88	3.08	7.18[d]	2.93	1.00	2.33[a]	0.95
Windows Server—Printing	–0.17	–0.96	–0.39	0.00	0.00	0.00	–0.17	–0.96	–0.39

	DT vs. Fleet			DT vs. ITTC			ITTC vs. Fleet		
Topic	M_{diff}	t	d	M_{diff}	t	d	M_{diff}	t	d
DHCP (Dynamic Host Configuration Protocol)	3.33	7.40[d]	3.02	1.75	3.89[c]	1.59	1.58	3.52[b]	1.44
*Active Directory	2.58	5.64[d]	2.30	2.50	5.46[d]	2.23	0.08	0.18	0.07
Server Management	4.75	10.02[d]	4.09	3.50	7.38[d]	3.01	1.25	2.64[a]	1.08
Others	1.00	4.15[c]	1.69	1.42	5.88[d]	2.40	–0.42	–1.73	–0.71

[a] $p < .05$, [b] $p < .01$, [c] $p < .001$, [d] $p < .0001$.

* Topics covered in Part 1 or 2 and elaborated in Part 3.

In sum, although DT students scored higher than Fleet ITs on 85 percent of IT topics and higher than ITTS students on 76 percent of IT topics, there are differences in the success with which the DT program presents its topics. These differences suggest areas for improvement in the program.

B. IWAR 2 Versus IWAR 1

Is digitized tutoring as effective as human tutoring?

Human tutors provided nearly all IT instruction for IWAR 1 participants, but it was digitized for IWAR 2. Other than the delivery medium, the tutoring systems employed in IWAR 1 and IWAR 2 were basically the same in instructional content and approach. Comparing Knowledge Test results on Parts 1 and 2 from IWAR 1 and IWAR 2 is not experimentally ideal nor a final answer to the question of human versus digitized tutoring, but it presents a unique opportunity to obtain some relevant data.

Combining scores from right-most bars of the pairs in Figure 18 suggests that the overall mean was higher for digital than for human tutoring, but this difference is not statistically significant: ($t(22) = 1.558$, $p < 0.10$).[8] A breakdown of the overall score revealed that the advantage of the digital tutor condition was statistically significant for items in Part 1 (t (22) 3.226, $p < 0.01$), but not in Part 2 ($t(22) = 0.468$, $p < 0.60$) of the Knowledge Test.

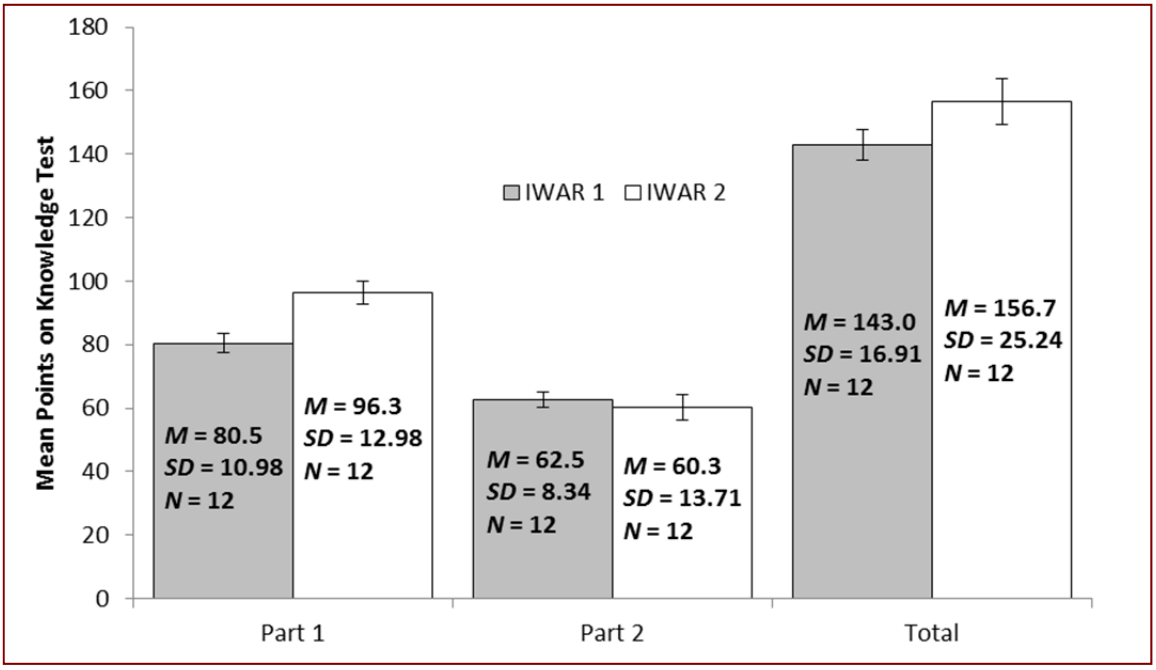


Figure 18. Means and Standard Deviations of Knowledge Test Scores for Human (IWAR 1) and Digital (IWAR 2) Tutoring.

[8] This notation signifies a t-score value of 1.558 with 22 degrees of freedom, which could occur up to 10 times out of 100 by chance.

In sum, the students taught primarily by the Digital Tutor matched the performance of students tutored in one-on-one settings by highly qualified and carefully screened tutors. The Digital Tutor is a first, research version of a 16-week course of instruction. It is not unreasonable to expect that in future trials, Digital Tutor students will achieve higher levels of expertise, perhaps surpassing levels that are achievable with human tutoring.

C. Practical Exercise Correlations

Does IT troubleshooting performance depend on ability and/or IT knowledge?

The Practical Exercises in IWAR 1 and 2 were performed in teams (as they would be at Navy duty stations). Therefore, meaningful correlations between individual ability, IT Knowledge, and Practical Exercise performance could not be calculated from the IWAR 2 data.

But Assessment Four, which was performed in November 2010 and described earlier, assessed the performance of individual 8-week DT students compared with individual, 16-week IT of the Future (IToF) graduates on Practical Exercises involving Troubleshooting and Packet Tracer[9] problems. The DT students and the IToF graduates also took a written IT Knowledge Test, not unlike those used in IWAR 1 and 2. Also, their Armed Forced Qualification Test (AFQT) scores, which averaged about 8 points lower than those of the IToF graduates, were available. These data allowed us to assess, using linear regression, the relationships between ability as measured by the AFQT, IT knowledge, and performance in practical exercises. Table 21 shows the correlation coefficients from this assessment.

Table 21. Correlations From Assessment Four Data between Practical Exercises, AFQT Scores, and Knowledge Testing.

Correlations from Assessment Four	DT	IToF
1- Troubleshooting w/ AFQT	0.32	0.07
2- Troubleshooting w/ Knowledge Test	0.64	0.07
3- Packet Tracer (Weighted) w/ Knowledge Test	0.68	0.76
4- Packet Tracer (Unweighted) w/ Knowledge Test	0.73	0.63
5- Knowledge Test with AFQT	0.56	0.45
6- Packet Tracer (Weighted) w/ AFQT	0.31	0.51
7- Packet Tracer (Unweighted) w/ AFQT	0.33	0.41

[9] The Packet Tracer program can be used to enhance or test students' understanding of routing and message traffic through a network.

These correlations suggest that:

- The IT knowledge of the DT students was related to their ability to solve IT troubleshooting problems (Row 2).
- The IT knowledge of the DT students was more strongly related to their ability to solve troubleshooting problems than was overall ability, as measured by the AFQT test (Rows 2 and 5).
- Neither overall ability as measured by the AFQT test nor IT knowledge were related to the ability of IToF graduates to solve IT troubleshooting problems (Rows 1 and 2).
- IT knowledge was related to both the DT students' and IToF graduates' Packet Tracer performance (Rows 3 and 4).
- Ability as measured by the AFQT was related to both the DT students' and IToF graduates' Packet Tracer performance, but more closely for the IToF graduates than for the DT students (Rows 6 and 7).
- Ability as measured by the AFQT was related to both the DT students' and IToF graduates' Knowledge Test scores, but slightly more closely for the DT students than for the IToF graduates (Row 5).
- Ability as measured by the AFQT was related to Packet Tracer performance, but more for those with less IT knowledge (IToF graduates) than for those with considerably more knowledge (DT students) (Rows 6 and 7).
- Ability as measured by the AFQT was related to Troubleshooting performance, for those with more IT knowledge (DT students) and not at all for those with considerably less IT knowledge (IToF graduates) (Row 1).

In sum, IT troubleshooting performance of the DT students appeared to be mildly dependent on ability as measured by AFQT scores, but it was more strongly related to IT knowledge provided by the DT training. Troubleshooting performance by IToF students was found to be related to neither AFQT scores nor IT knowledge they had acquired through their classroom training.

D. Dependence on Verbal Ability

Does the effectiveness of digital tutoring depend on reading and other verbal abilities?

Because so much instruction in the Digital Tutor is conducted through reading, the extent to which reading ability affects students' progress is of interest. DT and ITTC groups were assessed on individual differences related to verbal ability. The Armed

Forces Vocational Ability Battery (ASVAB) and the Gates-MacGinitie reading test (GMRT) were used to assess this possibility.

Scores included from the ASVAB were AFQT, Paragraph Comprehension (PC), Verbal Comprehension (VC), and Word Knowledge (WK). Scores included from the GMRT were Total Reading (T-ESS), Reading Comprehension (C-ESS), and Reading Vocabulary (V-ESS). Extended Scale Scores (ESS) are normalized metrics, with favorable psychometric properties derived from GMRT raw scores. They provide equal interval units normalized to a mean of 500.

Table 22 compares DT and ITTC participants on ASVAB and GMRT measures. The data indicate that ITTC participants scored slightly higher than DT participants on all individual measures, but none of the differences were significant.

Table 22. Mean Comparisons of DT and ITTC Groups on Individual Difference Measures of Overall Verbal Ability (ASVAB) and Reading (GMRT).

	DT		ITTC			
Measure	***M***	***SD***	***M***	***SD***	***t***	***p***
ASVAB						
AFQT	80.83	12.31	84.17	8.75	–0.765	0.453
PC	55.92	7.61	58.17	4.71	–0.871	0.393
VC	56.25	6.18	58.00	3.98	–0.825	0.418
WK	56.08	5.48	57.00	4.97	–0.429	0.672
GMRT						
T-ESS	650.75	31.12	664.67	29.82	–1.118	0.275
C-ESS	651.17	29.65	659.42	19.13	–0.810	0.427
V-ESS	629.42	28.58	645.67	32.68	–1.297	0.208

Table 23 displays within-group correlations between ASVAB and GMRT measures of verbal ability and two IWAR criterion measures—the Knowledge Test and Review Board scores.

These correlations suggest at least four findings with regard to the ASVAB measures:

- The AFQT scores of both DT and ITTC graduates correlate positively with their Knowledge Test scores.
- Although participants with higher AFQT scores tended to perform better on the Knowledge Test, the advantage of DT over ITTC was about the same (80–90 points) for low-scoring as for high-scoring participants.
- The lower correlations between AFQT and the Knowledge Test for the DT graduates than for the ITTC graduates suggests that success in the DT program

may be less dependent on overall ability (AFQT) than it is in the ITTC program—perhaps because tutoring can adapt the pace, content, and sequence of instruction to individuals more precisely than classroom instruction.

- The relations of ASVAB measures of overall ability (AFQT) and verbal ability (PC, VC, and WK) to Review Board ratings were effectively zero—none accounted for more than 4 percent of the variance in these ratings for either group.

Table 23. Within-Group Correlations of ASVAB and GMRT Individual Difference Measures and Knowledge and Review Board Scores.

	Knowledge Test		Review Board	
Measure	**DT**	**ITTC**	**DT**	**ITTC**
ASVAB				
AFQT	0.61[a]	0.77[b]	0.16	0.11
PC	0.36	0.64[a]	–0.07	–0.08
VC	0.43	0.75[b]	–0.07	0.12
WK	0.43	0.55	–0.08	0.18
GMRT				
T-ESS	0.00	0.50	–0.59[a]	0.19
C-ESS	–0.32	0.27	–0.66[a]	0.11
V-ESS	0.33	0.59a	–0.38	0.26

[a]$p < .05$; [b]$p < .01$

The GMRT measures in Table 23 suggest the following:

- The GMRT measures of total reading skill (T-ESS) and reading comprehension (C-ESS) were significantly and negatively correlated with the Review Board ratings in the DT group.
- In contrast, the GMRT measures were positive, but non-significant, accounting for no more than 6 percent of the variance in Review Board ratings for the ITTC group.
- Best-fit linear functions derived from the correlations suggest that the advantage of DT over ITTC participants in the Review Board interviews was more pronounced for low-reading-ability subjects.
- Examination of data points also suggests that the negative correlation in the DT group is largely due to one data point: the participant who scored lowest on the GMRT test, scored highest on the Review Board interviews.
- The positive relationship between reading ability and scores on the Knowledge Test is smaller in every case for the DT than for the ITTC participants,

suggesting that their higher scores were less due to reading ability than actual IT knowledge.

In sum, reading ability appeared to be unrelated to Knowledge Test scores for the DT students, but reading vocabulary was related to Knowledge Test scores for ITTC students.

5. Final Comments

First, it should be emphasized that the Digital Tutor is a research project and a work in progress. IWAR 2 was the first full test of a 16-week version of the Digital Tutor. Despite promising summative results and helpful formative data provided by these assessments, the Digital Tutor is not yet a finished product. Significant improvements have been made in the Digital Tutor since IWAR 2. Obvious next steps, such as inclusion of a proper tutoring segment for security and improved training in areas such as Unix operating system, Windows permissions, and group policy, are in order. Others may well emerge as use of the Digital Tutor continues.

Second, as technically capable as DT graduates are, they are still novices when it comes to Navy culture and leadership skills. The Digital Tutor appears to provide a strong technical background in relevant IT knowledge and skills, but it is silent when it comes to issues of "sailorization." Navy residential training commands are aware of this issue and attend to it well in their "A" schools. Duty stations must be prepared to capitalize on the technical abilities of young DT trained sailors with very high levels of expertise, while also dealing with them as individuals who have much yet to learn about the Navy, Navy culture, and Navy leadership.

Third, the Digital Tutor is intended to produce higher level, conceptual understanding of IT systems well beyond memorization of facts and application of straightforward procedures. This intent is in accord with considerable research showing that such understanding promotes both retention of what is learned and the ability to apply ("transfer") it to situations and systems not specifically covered by initial instruction (e.g., Wisher, Sabol, Ellis, 1999; Kimball & Holyoak, 2000). To an appreciable extent, transfer is measured by IWAR practical exercises, but retention can only be measured after time has passed, and it remains a matter of concern and interest. Given the body of data now available, a comparison of the knowledge and skills retained by ITTC and DT graduates after 12–18 months would be invaluable in assessing return on investment in both approaches.

Fourth, validating the impact of any training program on operational effectiveness is critical. Such assessments are difficult to perform and generally limited to surveys of subjective impressions, if done at all. It is possible to approximate the sizable monetary return on investment from the Digital Tutor (e.g., Cohn & Fletcher, 2010), but its effect on operational performance remains to be determined and may be critical in guiding further investment in this technology. Beyond surveys, there may be opportunities to

gather more objective data on the operational impact of the Digital Tutor and its graduates, perhaps through systematic analysis of trouble tickets arriving from the Fleet. These opportunities seem worth pursuing.

Finally, as sizable as the Knowledge Test results are, knowledge is an indirect indicator—an enabler of the expert level of performance targeted by the Digital Tutor. This assumption is not unreasonable. It is supported by a correlation of 0.64 between IT knowledge and troubleshooting performance found for the DT graduates, suggesting that knowledge accounts for about 40 percent of their troubleshooting scores. But as training, rather than education, the Digital Tutor is intended to prepare individuals for specific jobs and tasks. The most direct measure of the Digital Tutor's ability to achieve training goals of interest to the Navy is performance on the practical exercises.

In sum, the DARPA Digital Tutor effort appears to have achieved its goals. The design of the Digital Tutor is likely to be of significance for the development of training in general and the advancement of instructional technology in particular. Moreover, the Digital Tutor has shown that in 16 weeks it can produce students who outperform students with more than double that time in classroom instruction and sailors with 7–9 years of Fleet experience. These comparisons have included lengthy tests of knowledge and job-sample, practical exercises, both of which found levels of performance by the Digital Tutor students at levels that are unprecedented in assessments of training effectiveness. The greater efficiency, absence of harmful errors, and ability to solve problems at the highest levels of difficulty demonstrated by Digital Tutor students suggest both monetary and operational returns of substantial value to the Navy.

Appendix A. References

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

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

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

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

Graesser, A. C., D'Mello, S. K., and Cade, W. 2011. Instruction based on tutoring. In *Handbook of Research on Learning and Instruction,* edited by R.E. Mayer and P.A. Alexander, 408–426. New York: Routledge Press.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology* 9, 359–387.

Kimball, D. R., & Holyoak, K. J. 2000. "Transfer and Expertise." In *The Oxford Handbook of Memory,* edited by E. Tulving & F. I. M. Craik, 109–122. New York: Oxford University Press.

Kulik, J.A., & Fletcher, J.D. 2012. "Effectiveness of Intelligent Tutoring Systems." IDA Document D–4464. Alexandria, VA: Institute for Defense Analyses.

Psotka, J., Massey, L. D., & Mutter, S. A. (Eds.). 1988. *Intelligent Tutoring Systems: Lessons Learned.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sleeman, D., & Brown, J. S. (Eds.). 1986. *Intelligent Tutoring Systems*. New York: Academic Press.

Thalmeier, W., & Cook, S. 2002. *How to Calculate Effect Sizes from Published Research: A Simplified Methodology.* Somerville, MA: Work-learning Research.

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

US Department of Education. 2011. *Procedures and Standards Handbook: Version 2.1*. Washington, DC: Institute of Education Sciences. What Works Clearinghouse. September.

Wisher, R. A., Sabol,, M. A., & Ellis, J. A. 1999. *Staying Sharp: Retention of Military Knowledge and Skills.* ARI Special Report 39. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Woolf, B. P., & Regian, J. W. 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 & J. D. Fletcher, 339–356. New York: Macmillan Library Reference.

Appendix B. Figures

Appendix C. Tables

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
September 2012	Final	June 2012 – September 2012

4. TITLE AND SUBTITLE

DARPA Digital Tutor: Assessment Data

5a. CONTRACT NUMBER
DASW01-04-C-0003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

J.D. Fletcher
John E. Morrison

5d. PROJECT NUMBER

5e. TASK NUMBER
DA-2-2896

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-4686
Log: H12-001207/1

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

Defense Advanced Research Projects Agency
Defense Sciences Office
3701 N. Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S)

DARPA

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The DARPA Digital Tutor effort is intended to meet a Navy operational need and advance the technology of computer applications in instruction. Its strategy has been to observe in systematic and specific detail the practice of individuals who are expert in both a subject matter and tutoring and then capture their instructional techniques and capabilities in computer technology. Based on an analysis of need and criticality, DARPA selected "A" school and some "C" school training for the Navy Information Systems Technician (IT) rating for this effort.

This report summarizes results from the first four assessments and discusses Assessment Five (IWAR 2) in more detail. These assessments showed that after 16 weeks of training with the Navy "A" school students scored substantially higher than students with 35 weeks of training, Fleet ITs with 7–9 years of experience, and "A" and "C" school instructors on practical troubleshooting tests, system design tests, and knowledge tests with effect sizes ranging from 0.80 to 4.50. Overall, the Digital Tutor appears to have achieved its goals. It produced human performance levels of substantial operational value to the Navy, and it demonstrated a promising, perhaps breakthrough, direction for the development of training in general and instructional technology in particular.

15. SUBJECT TERMS

Intelligent Tutoring System (ITS), Information Technology (IT), Training, Digital Tutor, Computer Based Instruction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					LTC William Casebeer
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	67	19b. TELEPHONE NUMBER *(include area code)* 703-526-4163

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