IDA

INSTITUTE FOR DEFENSE ANALYSES

# DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor Assessments

J.D. Fletcher

February 2011

IDA Document NS D-4260
Log: H 11-000117

**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.

**Copyright Notice**

4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-4260

# DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor Assessments

J.D. Fletcher

# Summary

This report presents findings from two assessments of the Digital Tutor (DT) being developed by the Education Dominance Program of the Defense Advanced Research Projects Agency (DARPA). This tutor is providing initial specialized skill training ("A" school and some additional "C" school training) for the Navy's Information Systems Technology (IT) rating. DT instruction is conducted as a spiral curriculum in which material initially presented is reiteratively elaborated and deepened in the light of additional course material, which itself is similarly elaborated and deepened.

The assessments determined how well the DT was meeting IT training requirements and preparing students for Fleet IT duties. These assessments were performed in April 2010 with the 4 weeks of DT training then available for student use and again in November 2010 with another group of students who had completed the 7 weeks of DT training then available. Both assessments were performed at the Navy's Center for Information Dominance (CID), Corry Station, Pensacola, Florida.

The April assessment used a Written Knowledge test to compare IT knowledge acquired by the 4-week DT students with that of students who had finished the "A" school Integrated Learning Environment (ILE) training, which takes on average 8 weeks to complete. The DT students scored significantly higher than ILE students with an effect size of 2.81 standard deviations ("sigmas") on the knowledge test. The DT students also scored significantly higher, with an effect size of 1.25 standard deviations, than CID IT instructors on the test.

The November assessment compared IT capabilities of four groups: the 7-week DT students, students who had completed ILE training, IT of the Future (IToF) students who had completed its 19 weeks of training, and CID IT instructors. The assessment again used a Written Knowledge test, which was taken by all four groups. DT and IToF students also performed Practical Troubleshooting exercises, networked Packet Tracer exercises, both of which used real world problems taken from the Fleet, and interviews by a three-member Oral Examination Board whose members did not know from which of the two training programs the interviewees were drawn.

The DT students outscored

- ILE students on the Written Knowledge test, with an effect size of 4.68.
- IToF students, with an effect size of 1.95.

- Instructors, with an effect size of 1.35.
- IToF students in the Practical Troubleshooting exercises, with an effect size of 1.90.
- IToF students in the Packet Tracer exercises, with an effect size of 0.74 for scores not weighted for difficulty and an effect size of 1.00 for scores that were weighted for difficulty.
- IToF students on the Oral Reviews, with an overall effect size of 1.34.

Scores of the DT students on the Written Knowledge test accounted for about 40% of the variance in their scores on the Practical Troubleshooting exercises, indicating that the spiral approach taken by the DT curriculum is successfully preparing students to apply their newly acquired knowledge in practice.

Summary findings are that the DT training will:

- Provide the knowledge and practical skills necessary to perform IT duties required by CID training and needed to perform Navy IT operational duties.
- Provide its students with considerably more IT knowledge and skills in substantially less time than IToF or ILE training.

# Contents

# 1. Background

The benefits of tutorial instruction—one tutor for one student—have long been noted (e.g., Bloom 1984; Fletcher 1990; Graesser and Person 1994; Graesser, D'Mello, and Cade 2009). These benefits include intense interactivity and time on task; high learner motivation; and the efficiencies of instruction specifically tailored to learner needs, interests, and capabilities. This approach has been deemed an instructional imperative but, with a few critical exceptions, an economic impossibility (Scriven 1975). The advent of computer technology and adaptive computer-based instruction promises to make it affordable.

Tutorial instruction delivered by computer also promises to reliably (and affordably) compress the many years needed by military technicians to develop high levels of technical expertise into a few months. Efforts to realize this promise has been supported since the late 1960s, much of it sponsored by the Department of Defense (e.g., Brown, Burton, & Bell 1975, Carbonell 1970; Fletcher 2009; Luckin, Koedinger, and Greer 2007; Psotka, Massey, and Mutter 1988; Sleeman and Brown 1982). It is a fundamental objective of DARPA's Education Dominance program.

# 2. DARPA's Education Dominance Program and the Digital Tutor

DARPA's Education Dominance Program is developing techniques of instruction that reach beyond memorization and straightforward application of facts and procedures to develop deeper conceptual and analytical understanding of technical subjects. This understanding has been shown to enable long-term retention and transfer of knowledge and skills, as well as their creative application in dealing with the novel and unexpected situations that inevitably arise in military operations (Wisher, Sabol, and Ellis 1999, Kiszley 2007). For this reason, the program employs a spiral curriculum approach in which material initially presented is reiteratively elaborated and deepened in the light of additional course material, which itself is similarly elaborated and deepened as the instruction proceeds.

Noting that large differences can be observed and measured in student learning produced by different teachers, DARPA, through its research contractor, is analyzing and then capturing in computer technology the best practices of expert human tutors. Its goals are to develop a Digital Tutor (DT) that, in a matter of months, produces individuals with the knowledge and skills of experts possessing many years of experience and then to implement this capability as a core component of Navy training.

After a review of many critical military training programs, DARPA focused its effort on the initial specialized skill training ("A" school training) used to prepare and qualify sailors for the Navy Information Systems Technician (IT) rating. The tutor that is now under development will provide 16 weeks of IT training. DARPA tasked the Institute for Defense Analyses (IDA), as a Federally Funded Research and Development Center for the Office of the Secretary of Defense, to provide independent assessments of progress and development of this DT. Thus far, IDA has conducted three such assessments.

The first assessment (Phase 1 IWAR) was conducted over a 5-week period in July-August 2009. It compared the knowledge and skills of Navy ITs who had many years of Fleet experience with those of students who had less than a year of Navy experience but who had completed 16 weeks of IT training using the Education Dominance curriculum. One week of the DT itself, which was then available, was included in this training. Expert human tutors who were also expert in specific topics of the curriculum delivered the remaining 15 weeks of training. The Education Dominance students were found to be superior in knowledge, measured by scores on a written test, and in practical skills, measured by performance of troubleshooting tasks using Navy IT equipment, to the knowledge and skills of Fleet ITs who averaged 7 years of experience in performing Navy IT duties (Fletcher 2010).

Two subsequent assessments were administered by IDA as the full digital version of the tutor was being developed. IDA administered the first of these in April 2010, assessing students who had completed the 4 weeks of DT training that was then available. Seven weeks of DT training were available in November 2010, and IDA administered the second assessment to another group of DT students at that time. Both assessments were conducted at the Navy Center for Information Dominance (CID), Navy Technical Training Center, Corry Station, Pensacola. This report summarizes findings from these two assessments.

# 3. April 2010 Digital Tutor Assessment

The purpose of this assessment was to determine how well the 4 weeks of DT training then available were preparing students with the knowledge required by CID IT training and Fleet IT duties. This was primarily done by comparing the progress of DT students with that of other IT students using standard CID curriculum materials and approaches.

## A. Participants

IDA tested three groups:

- **20 DT students** who had completed 4 weeks of the planned 16-week DT course in the week prior to this assessment.
- **31 Integrated Learning Environment (ILE) students** who had completed that course in the week prior to this assessment. The ILE course is the current "A" school training intended to provide the knowledge and skills in information systems technology initially needed by Navy ITs. It is designed for 11 weeks of training but students complete it, on average, in about 8 weeks.
- **13 CID instructors**. Four of the instructors were drawn from the DT course and six were drawn from the ILE course.

Both groups of students were chosen at random from the IT "A" school students who were available when training began. The average age of both groups was about 20. The average Armed Forces Qualification Test (AFQT) percentile score for the DT students was 73.7, and the average for the ILE students was larger, at 78.3.

## B. Test Development, Administration, and Scoring

The April assessment consisted of a single "closed-book, closed-notes" written test of IT knowledge. The test was intended to be sufficiently difficult to avoid ceiling effects (too many scores near the maximum) and floor effects (too many scores near zero). It was arbitrarily divided into two parts. Part 1 consisted of 63 items; Part 2 consisted of 89 items—152 items in all. Participants were given 90 minutes to finish each part, with a 30-minute break between. Nearly all participants finished each part in less than an hour.

IDA assembled the test from items collected from

- 51 items comprised in a test produced earlier by CID.

- 25 items produced by CID instructors for this occasion.
- 5 items added by IDA.
- 93 items prepared by the DT developer.

This process yielded a pool of 174 items. The items were vetted in detail by four members of the IDA research staff specializing in IT issues. IDA edited errors and reformatted a number of items. Duplicates were discarded along with other items that were judged too ambiguous to adapt or use, yielding a test of 152 items. Topics covered by the test were:

- Hardware.
- Number Systems.
- General Networking Concepts.
- Windows Operating System.
- Windows Permissions.
- User Accounts.
- File and Folder Sharing.
- Internet Protocol.
- Domains.
- Group Policy.
- Active Directory.

Of the 152 items, 25 were assigned partial credit and were scored 0, 1, or 2 points, so the test had a maximum test score of 177 points. Examples of these 2-point questions are “Name 2 examples of…” and “Explain the purpose of....”

IDA also administered the test to four members of its own IT support staff. Their comments on the test were useful and incorporated in the test. Overall, they judged it to be a reasonable and balanced assessment of IT knowledge.

Test administration at the CID was proctored by CID instructors and IDA personnel. The test was graded by three of the IDA researchers who had earlier vetted the test items. Open-answer questions were discussed by all graders to ensure consistency in the scoring.

## C. Results

Means, standard deviations, and numbers of observations (*N*) are shown in Tables 1–3. Probabilities that the observed differences might have occurred by chance are shown if they were judged to be statistically significant (for this purpose, a probability of occurring by chance less than 0.05).

Effect sizes are also shown in the tables. Effect sizes directly estimate the magnitude of impact different treatments have on groups being compared. Effect sizes are sometimes called “sigmas” because they estimate effects in standard deviation units, which are usually signified by the Greek letter sigma in statistical notation. Influential discussions of effect size have been provided by Glass and McGaw (1980), Hedges and Olkin (1985), Cohen (1988), and Rosenthal (1991), among others.

There are several ways to calculate effect size, but Cohen’s *d* appears to be the most common metric and is used here. It is roughly calculated as:

$$d = (\text{Mean of Group 1} - \text{Mean of Group 2}) / \text{“Pooled” Standard Deviation.}$$

This metric keys on the assumption that both groups under consideration provide estimates of the single population from which they are drawn, and therefore a pooled standard deviation is likely to be closer to that population’s standard deviation than an estimate obtained from either of the groups alone. There are alternatives to this assumption, but common practice may suffice for this report. Cohen’s *d* is generally more conservative than other measures of effect size.

Effect sizes measured by Cohen’s *d* that are less than 0.20 are commonly assumed to indicate only minor effects, those between 0.20 and 0.40 are regarded as small but real, and those between 0.40 and 0.60 are considered moderate. Those between 0.60 and 0.80 are considered large. Those above 0.80 are deemed very large and occur rarely in assessing the effects of different instructional techniques. All the effect sizes reported in Tables 1–3 would be regarded as very large.

Tables 1–3 each report the pair-wise comparisons available given the three groups available for this assessment. All three tables show statistical and practical significance. The effect size of 2.81 reported in Table 1, which compared DT and ILE student test scores, is roughly equivalent to increasing the performance of 50th percentile students to about the 99th percentile. This result seems notable given that the DT students had received only 4 of the 16 weeks of instruction planned for the DARPA spiral curriculum compared with the average 8 weeks of instruction received by the ILE students as they completed their course.

The results and effect size of 1.25 reported in Table 2, which compared instructor and ILE student test scores, were about what one would expect—the instructors scored higher than ILE students on the knowledge test. The effect size of 1.26 reported in Table 3, however, was surprising. It shows that DT students scored significantly higher on the Knowledge Test than did the CID instructors.

**Table 1. Comparison of DT student and ILE student test scores.**

| **Digital Tutor** | | | **Integrated Learning Environment** | | | **Probability** | **Effect Size (Sigma)** |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 128.4 (72.9%) | 14.5 | 20 | 63.8 (36.3%) | 27.0 | 31 | *p* < .01 | 2.81 |

**Table 2. Comparison of instructor and ILE student test scores.**

| **Instructors** | | | **Integrated Learning Environment** | | | **Probability** | **Effect Size (Sigma)** |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 99.8 (56.7%) | 34.0 | 10 | 63.8 (36.3%) | 27.0 | 31 | *p* < .01 | 1.25 |

**Table 3. Comparison of DT student and instructor test scores.**

| **Digital Tutor** | | | **Instructors** | | | **Probability** | **Effect Size (Sigma)** |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 128.4 (72.9%) | 14.5 | 20 | 99.8 (56.7%) | 34.0 | 10 | *p* < .01 | 1.26 |

Table 4 shows high, median, and low scores for the two groups.

**Table 4. High, median, and low test scores for DT and ILE students (from a maximum score of 177).**

| | **High** | **Median** | **Low** |
|---|---|---|---|
| DT | 151 (85%) | 131 (74%) | 91 (51%) |
| ILE | 128 (72%) | 52 (29%) | 23 (13%) |

Additional results are shown in Table 5, which compares the knowledge test scores of the four DT instructors with those of DT students. The table shows that the DT instructors outscored their students, even though the opposite result was obtained (Table 3) when *all* instructors were included in this comparison. The difference is not statistically different and far from conclusive, since data from so few instructors were available, but the possibility that DT instruction benefits instructors as well as their students seems of interest.

**Table 5. Comparison of DT instructor and DT student test scores.**

| Digital Tutor Instructors | | | Digital Tutor Students | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 133.5<br>(75.9%) | 21.2 | 4 | 128.4<br>(72.9%) | 14.5 | 20 | Not significant | 0.33 |

In sum, it appears that students using the DT are acquiring the knowledge needed to meet CID objectives and to perform Fleet IT duties. It also appears that the DT students are acquiring substantially more of this knowledge in considerably less time than ILE students.

# 4. November 2010 Digital Tutor Assessment

The purpose of this assessment was to determine how well the 7 weeks of DT training then available was preparing students with the knowledge and, in addition, the practical troubleshooting skills required by CID IT training and Navy Fleet IT duties. Again, assessment was primarily accomplished by comparing the progress of DT students with that of IT students using other CID training materials and approaches.

## A. Participants and Measures

IDA tested four groups:

- **20 DT students** who had completed 7 weeks of the planned, spiral 16-week DT curriculum in the week prior to this assessment.
- **20 IT of the Future (IToF) students** who had completed the current 19-week version of that course. Like the ILE course, the IToF course is intended to provide the basic, “A” school training in information systems technology initially needed by Navy ITs.
- **18 Integrated Learning Environment (ILE) students** who had completed that course. It is designed for 11 weeks but students complete it, on average, in about 8 weeks. The ILE course is the current, “A” school training intended to provide the knowledge and skills in information systems technology initially needed by Navy ITs.
- **10 CID instructors from the ILE course**, all of whom had been trained to present subjects in the IToF course.

The average AFQT percentile score for the DT students was 76.75. It was 84.55 for the IToF students and 78.29 for the ILE students. Average and median age of each group was about 20. The DT and ILE students were selected at random from the IT “A” School pool. The median ILE AFQT score was 8 points higher than that of DT students. Sixteen of the 20 IToF students had enlisted under the Navy’s Advanced Technical Field program, which requires higher AFQT scores. Their median AFQT score was 12.5 points higher than that of DT students.

Both the DT and IToF students completed their IT training the week before this assessment. All ILE students completed IT training within the 2 weeks before this assessment. Most of these students completed it the week before.

IDA used four types of assessments:

- **Written Knowledge Test**—4 hours, delivered as two 90-minute parts to all four groups. The two parts covered different IT topics, but allocation of topics to parts was arbitrary. The topics covered were the following:
  - Hardware.
  - Number Systems.
  - General Networking Concepts.
  - Windows Operating System.
  - Windows Permissions.
  - User Accounts.
  - Windows Server (Printer).
  - File and Folder Sharing.
  - Internet Protocol.
  - Exchange Server.
  - Domains.
  - Domain Name System.
  - Group Policy.
  - Open Systems Interconnection Model.
  - Dynamic Host Configuration Protocol.
  - Active Directory.
- **Practical Troubleshooting Exercise**—4 hours, 15 trouble tickets presented to DT and IToF students. Students, working individually, responded to trouble tickets that were adapted from a collection of about 20,000 that had not been solved aboard ships in the Fleet over the last 3 years. Problems were represented on two virtual networks. Each student started with the full set of problems within each network and was free to address them in any order. The students began with one network and one set of problems, solved those that they could, and then switched over to the other network. Students chose when to switch, but they could not go back to the first network after doing so. These procedures were adapted to accommodate the limited number of computer resources available. Students were permitted to use their class notes for this exercise on the assumption that they would have access to similar materials in their duty stations.

- **Packet Tracer Exercise—**2 hours, 18 trouble tickets presented to DT and IToF students. Students, working individually, used Cisco's Packet Tracer software to resolve trouble tickets, again adapted from the Fleet database of 20,000. Packet Tracer is primarily a training aid, but it can be used for testing. It displays the network topology of a multiple-router environment, thereby enabling instruction and assessment to take place without a need for physical devices while providing more visibility and control over time than would otherwise be available. It simulates network traffic with continuous and real-time updating and control over underlying network logic and activity. Students began with the full set of 18 trouble tickets and were free to address them in any order they chose. Both groups of students had sufficient experience with the Packet Tracer program in training to use it in an assessment. The students were not permitted to use their class notes in this exercise.
- **Oral Examination Board—**20–30 minutes per student, 7 DT and 6 IToF students selected at random. The Board consisted of the Pacific Area of Responsibility (AOR) Lead for the Fleet Systems Engineering Team (FSET), an IT Chief Petty Officer, and a Fire Control Technician First Class Petty Officer—all had been identified and selected for their IT knowledge and expertise. The examinations were "blind" in that members of the Board did not initially know which students came from which group, although differences between students from the two groups became evident as the Oral Examinations progressed.

As noted, the Written Knowledge test was administered to all four groups. The other three assessments were only administered to DT and IToF students.

## B. Test Development, Administration, and Scoring

As in the April assessment, the tests were intended to be sufficiently difficult to avoid ceiling effects (too many scores near the maximum) and floor effects (too many scores near zero).

All items on all tests were reviewed and screened by four members of the IDA research staff as well as the Pacific AOR Lead FSET.

At least one CID instructor and one IDA researcher proctored all testing

### 1. Written IT Knowledge Test

IDA assembled the Written Knowledge test from:

- 51 items produced by CID and administered in April 2009 to IT and DT students.
- 25 items produced by CID instructors for the DT testing in April 2010.

- 20 items developed by IDA specifically for this test.
- Additional items developed or collected from various sources by DT technicians.

IDA edited errors and reformatted a number of items. Duplicate questions were discarded along with other items that were judged too ambiguous to adapt or use.

Part 1 and Part 2 of the Written Knowledge test had 143 and 150 questions, respectively. Twenty-five items on Part 1 and 6 items on Part 2 permitted partial credit and were scored 0, 1, or 2. In all, there were 293 questions on the Written Knowledge test with 324 points possible.

The Written Knowledge test was scored by IDA group members working together to ensure consistency and accuracy in grading—especially for the written, open-answer questions. The grading was blind in that IDA did not know which tests had come from which group.

### 2. Practical Troubleshooting

The 15 items used in the Practical Troubleshooting exercise were screened and selected by IDA from 164 trouble ticket candidates drawn from the Fleet database and judged appropriate for the 7-week DT content. Four pairs of IT experts scored exercise performance. Each pair had to agree on the scores to be assigned. In practice, these scores rarely deviated by more than 1 point. Responses to each problem were rated on a scale ranging from 0 (no attempt) to 5 (correct procedures and correct solution), for a maximum possible score of 75 points.

### 3. Packet Tracer

IDA selected and screened the 18 trouble tickets used in the Packet Tracer exercise from a larger set provided by DT technicians from the Fleet database of 20,000. Neither IToF nor DT students had any experience with the specific trouble ticket items used in these two exercises.

Both unweighted and weighted scoring was used for the Packet Tracer exercise. Unweighted scores depended on the number of tasks needed to resolve the problem. One point was assigned for each task completed. Unweighted points for each problem ranged from 1 to 12, with a maximum possible score of 75 points for the exercise.

Weighted scoring was accomplished first by assigning 2, 4, or 6 points to each problem based on an assessment of the problem as easy, medium, or difficult. After this assignment, 0–3 points were added, depending on the number of tasks required to resolve the problem and the number of (virtual) machines that had to be accessed in performing these tasks. Total weighted points for the problems ranged from 2 to 9, with a maximum total score of 74 points for the weighted scores.

### 4. Oral Examination

Each student in the Oral Examination 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. Students who demonstrated effectively no knowledge of a topic were assigned 0 points; those who were judged to possess typical knowledge and skills of "A" School students were assigned 2 points; those who demonstrated knowledge and skills typical of ITs with 1–3 years' experience received 3 points; 4 points were assigned to those who demonstrated knowledge and skills typical of ITs with more than 3 years' experience; students who demonstrated more knowledge than members of the Board possessed were awarded 5 points. Each of the 3 Board members scored each student so that a total of 90 points could be awarded to a student for the 6 topics.

Each Board member also assigned scores on a 0–5 scale to each student for the level of Satisfaction a manager might have with the student's likely performance on an IT team. Each Board member also assigned scores on a 0–5 scale based on the student's demonstrated confidence in his or her IT capabilities. These scores were assigned by the 3 Board members separately and then added so the maximum score any student could receive for either Satisfaction or Confidence was 15.

Note that the Oral Examinations were scored on a ranking scale—4 is greater than 2, for example—not on a ratio scale. That is, it would be incorrect to say that 4 is twice as large as 2, nor could we say that the difference between 1 and 3 is the same as the difference between 2 and 4. The scores are rankings and their averages only report an average ranking for a group.

## C. Results

### 1. Written Knowledge Test

Tables 6–11 compare the written test results. One test had to be discarded from the ILE group, leaving 17 rather than 18 students in that group.

DT students' scores were significantly higher from a statistical standpoint than those of IToF students, ILE students, and (again) the instructors (Tables 6–8), with effect sizes of 1.95, 4.68, and 1.35, respectively, all of which would be considered very large. An effect size of 1.95 is roughly equivalent to increasing performance at the 50th percentile to that of the 97th percentile, and an effect size of 1.35 is roughly equivalent to increasing performance at the 50th percentile to that of the 91st percentile. An effect size of 4.68 is beyond those included in standard charts. Here, we may be exceeding the practical use of effect size as a metric.

ITоF students significantly outscored the ILE students (Table 9) with another effect size (3.54) that is beyond standard charts, and which, again, may be beyond the practical use of effect size as a metric. There was, however, no practical or statistical difference between ITоF student scores and instructors' scores on the Written Knowledge Test (Table 10).

As in the April assessment, ILE students were again outscored by their instructors (Table 11).

**Table 6. Written knowledge test scores of DT and ITоF students.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 207.90<br>(64%) | 37.30 | 20 | 145.75<br>(45%) | 25.18 | 20 | *p* < .001 | 1.95 |

**Table 7. Written knowledge test scores of DT and ILE students.**

| Digital Tutor | | | Integrated Learning Environment | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 207.90<br>(64%) | 37.30 | 20 | 64.52<br>(20%) | 19.96 | 17 | *p* < .001 | 4.68 |

**Table 8. Written knowledge test scores of DT students and instructors.**

| Digital Tutor | | | Instructors | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 207.90<br>(64%) | 37.30 | 20 | 149.30<br>(46%) | 53.96 | 10 | *p* < .01 | 1.35 |

**Table 9. Written knowledge test scores of IToF and ILE students.**

| IT of the Future | | | Integrated Learning Environment | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 145.75 (45%) | 25.18 | 20 | 64.52 (20%) | 19.96 | 17 | *p* < .001 | 3.54 |

**Table 10. Written knowledge test scores of IToF students and Instructors.**

| IT of the Future | | | Instructors | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 145.75 (45%) | 25.18 | 20 | 149.30 (46%) | 53.96 | 10 | Not significant | 0.10 |

**Table 11. Written knowledge test scores of ILE students and Instructors.**

| Integrated Learning Environment | | | Instructors | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 64.52 (20%) | 19.96 | 17 | 149.30 (46%) | 53.96 | 10 | *p* < .001 | 2.35 |

Table 12 shows high, median, and low scores on the Written Knowledge test for the student groups.

**Table 12. High, median, and low scores for the Written Knowledge test (from a maximum score of 324).**

| Student Group | High | Median | Low |
|---|---|---|---|
| **DT** | 271 (84%) | 204 (63%) | 133 (41%) |
| **IToF** | 210 (65%) | 139 (43%) | 113 (35%) |
| **ILE** | 70 (22%) | 58 (18%) | 27 (8%) |

## 2. Practical Troubleshooting Exercises

This exercise presented 15 Trouble Tickets on virtual systems. It was undertaken by DT and IToF students. Among the tests, this exercise seems closest to the IT duties required in Navy Fleet operations. Its items were directly drawn from the database of

Fleet trouble tickets described earlier. Students performed the exercise without a visual aid like Packet Tracer but with their class notes because reference material is available and commonly used in Fleet IT duty. Students in this exercise received a score of 0–5 for each of the 15 items: 0 for no attempt at solution and 5 for a complete and satisfactory resolution of the problem.

Aggregate scores for the two groups are shown in Table 13. The difference between the two group means is about 4.6. It is statistically significant, with an effect size of 1.90, which would be considered very large. Roughly, this effect size suggests an improvement from 50th percentile performance to that of 97th percentile performance. Notably, 9 of the IToF students scored 0, indicating that the dynamic range of this exercise was too limited and the test itself was too difficult for that group.

**Table 13. Practical Troubleshooting exercise scores of DT and IToF students.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| **Mean** | **Std Dev** | ***N*** | **Mean** | **Std Dev** | ***N*** | | |
| 26.55 | 14.09 | 20 | 5.65 | 6.56 | 20 | $p < .001$ | 1.90 |

Table 14 shows high, median, and low scores for the two groups that took the Practical Troubleshooting Exercises.

**Table 14. High, median, and low scores for the Practical Troubleshooting exercises (from a maximum score of 75).**

| Student Group | High | Median | Low |
|---|---|---|---|
| **DT** | 50 (67%) | 22 (29%) | 6 (8%) |
| **IToF** | 19 (25%) | 4 (5%) | 0.0 (0%) |

### 3. Packet Tracer Exercise

This exercise consisted of 18 Trouble Tickets presented on virtual systems using the Packet Tracer program. It was scored in two ways. In the unweighted case, scores simply consisted of the number of tasks needed to resolve the problem being presented. In the weighted case, scores were first based on the difficulty of the problem that was presented and then augmented in accord with the number of tasks needed to resolve it.

Table 15 shows results for DT and IToF scores that were not weighted. Table 16 shows high, medium, and low unweighted scores for both groups. Comparisons of weighted scores for the two groups are shown in Table 17, with their high, medium, and low scores shown in Table 18. In both comparisons, the results statistically favor the DT students, but more for weighted than unweighted scores. The effect size of 0.74 for unweighted scores would be classified as large, roughly amounting to a performance improvement from the 50th percentile to the 77th percentile. The effect size for weighted scores would be considered very large and suggests a performance improvement from the 50th percentile to the 84th percentile.

**Table 15. Unweighted Packet Tracer exercise scores of DT and IToF students.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 36.91 | 16.2 | 20 | 25.29 | 15.3 | 20 | *p* < .05 | 0.74 |

**Table 16. High, median, and low unweighted scores for the Packet Tracer exercises (from a maximum score of 75).**

| Student Group | High | Median | Low |
|---|---|---|---|
| **DT** | 63.5 (85%) | 34.5 (46%) | 9.1 (12) |
| **IToF** | 57 (76%) | 20.5 (27%) | 0 (0%) |

**Table 17. Weighted Packet Tracer Exercise Scores of DT and IToF students.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 30.39 | 15.90 | 20 | 15.85 | 13.0 | 20 | *p* < .01 | 1.00 |

**Table 18. High, median, and low weighted scores for the Packet Tracer exercises (from a maximum score of 74).**

| Student Group | High | Median | Low |
|---|---|---|---|
| **DT** | 60.7 (81%) | 28.4 (38%) | 7.2 (10%) |
| **IToF** | 50.7 (69%) | 10.9 (15%) | 0.0 (0%) |

In both scoring procedures for the Packet Tracer exercise, the means and the medians of the DT scores are closer to one another than those of the IToF scores. This result suggests more skewing among IToF than DT students. The skew is to the left in these cases, indicating more low- than high-scoring students relative to the group. This result agrees with a common expectation that more students will fall behind in group instruction (predominant in IToF instruction) than in individualized, tutorial instruction (predominant in DT instruction) (e.g., Corno and Snow 1986). These results for the

Packet Tracer exercises echo the differences between means and medians found for the Written Knowledge tests in both the April and November assessments.

## 4. Oral Reviews

These reviews provided an opportunity to examine the DT and IToF students in a less structured fashion based around the core set of topics listed earlier. Students were drawn for this review at random from both groups. This was a valuable exercise, particularly given the background and quality of the Oral Board members. Still, the Board members needed more time for the reviews to increase both the number of students examined and the depth with which the Board could examine them.

Table 19 shows results for the average total scores in the DT and IToF groups. The DT students scored statistically higher than the IToF students, with an effect size of 1.34, which would be considered very large.

As stated above, this was a blind review. However, differences between the two groups became clear. The response by one of the Board members summarizes an impression evidently shared by all three:

> It seemed comparatively unambiguous that the Digital Tutor students understood IT in a way that the other students did not, even though they had less than 7 weeks of exposure compared to the 16 [19] weeks the IT of the Future [students] spent prior to this event. The confidence of the digital tutor students and their clear knowledge was very considerable. This was further displayed when they provided correct answers or explanations quickly which resulted in further deeper dives for level of comprehension. All 3 panel members were impressed.

**Table 19. Average ranking of DT and IToF students from the Oral Reviews.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| Mean | Std Dev | *N* | Mean | Std Dev | *N* | | |
| 2.31 | 0.07 | 7 | 1.39 | 0.44 | 6 | *p* < .05 | 3.02 |

Results for the Satisfaction and Confidence scores are shown in Tables 20 and 21. In both cases, the results are statistically significant and show large effect sizes, but less so for Satisfaction than for Confidence, thereby reinforcing long-standing results from testing programs suggesting that confidence and competence are not closely related (e.g., Shuford and Brown, 1974).

These findings are promising, but because of the small number of students interviewed, they need additional verification.

**Table 20. Average Satisfaction ranking of DT and IToF students from the Oral Reviews.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| **Mean** | **Std Dev** | ***N*** | **Mean** | **Std Dev** | ***N*** | | |
| 3.26 | 0.98 | 7 | 1.90 | 0.56 | 6 | *p* < .01 | 1.81 |

**Table 21. Average Confidence ranking of DT and IToF students from the Oral Reviews.**

| Digital Tutor | | | IT of the Future | | | Probability | Effect Size (Sigma) |
|---|---|---|---|---|---|---|---|
| **Mean** | **Std Dev** | ***N*** | **Mean** | **Std Dev** | ***N*** | | |
| 3.27 | 0.88 | 7 | 1.80 | 0.73 | 6 | *p* < .01 | 1.80 |

# 5. Additional Notes

The impact of DT technology seems evident in DT student data where there appears to be a "dosage effect." Taking ILE graduates as a baseline, the effect size of DT training rose in these assessments from 2.81 after 4 weeks of DT training to 4.68 after 7 weeks of DT training. It may increase even more with the full 16 weeks of training planned for the tutor.

Correlations between the AFQT and Written Knowledge test scores were found to be 0.56 for DT, 0.45 for IToF, and 0.27 for ILE students. Ability as measured by the AFQT then accounted for 20%–30% of the variance in Written test scores among the DT and IToF students, but less than 10% of that variance among the ILE students.

Correlations between AFQT and Troubleshooting scores of the DT students were 0.32, but 0.64 between their Written Knowledge test and Troubleshooting scores, indicating that, AFQT ability aside, the DT students were able to put the IT knowledge they had acquired to practical use. Similar comparison of correlations for IToF students cannot be made because of the limited dynamic range of the Practical Troubleshooting test for these students—the correlation between their Written test and Troubleshooting scores was 0.07, as was the correlation between their AFQT and Troubleshooting scores.

Correlations for DT students between AFQT and unweighted Packet Tracer scores were 0.33; between the AFQT and weighted Packet Tracer scores they were 0.31. Similar correlations for the IToF students were 0.41 for unweighted Packet Tracer and 0.51 for weighted Packet Tracer scores.

Correlations for DT students between Written test and unweighted Packet Tracer scores were 0.73; for the Weighted Packet Tracer scores they were 0.68. Similar correlations for the IToF students were 0.63 for unweighted Packet Tracer and 0.76 for weighted Packet Tracer scores. These correlations suggest that both groups were able to apply the IT knowledge they had acquired to problems in the Packet Tracer exercise.

It appears then that the ability measured by the AFQT helps both DT and IToF students acquire IT knowledge, but instruction in applying that knowledge remains of considerable importance in training for practical IT duties—echoing a result typically found in research on problem-solving (e.g., Mayer and Wittrock 1996). These findings also indicate the importance of a spiral approach—learn a little then apply it, learn a little more then apply it to more difficult problems, and so on—in this and in other training.

Along with most computer-assisted instruction, which can substitute technology for human labor, DT training may realize greater economies of scale than can IToF, which relies heavily on human instructors and classroom learning. The same should be true for ILE training, which also uses computer-assisted instruction, but which is providing students with considerably less capability than either IToF or DT training. As Clark (1983) emphasized in a widely noted article, technology, or any other medium, does not by itself guarantee high-quality instruction. Good design of the instruction and its proper implementation remain essential.

IToF instruction provides students with more IT skill and knowledge than ILE instruction, but not as much as the DT, and it takes more time to do so. It may increase manpower costs by holding students longer in training (19 weeks) than either DT (eventually 16 weeks) or ILE (average of 8 weeks), while producing less skilled ITs for Fleet IT duty than the DT.

DT training is expensive to produce. However, moderate- to large-scale training is far more sensitive to delivery costs than development costs (Fletcher and Chatham 2010). The perennial problem of up-front versus life-cycle costs remains. Funding to develop DT training is minor compared with the substantial savings and operational effectiveness to be gained from it (Cohn and Fletcher 2010). Unfortunately, these savings occur almost entirely in the Fleet and in overall operational effectiveness, while development costs may have to borne by the Navy’s residential training establishment, which operates with far fewer financial resources.

Finally, and in brief, it appears that the DT training will:

- Provide students with the knowledge and practical skills necessary to perform IT duties required by CID training and needed to perform Navy operational IT duties.
- Provide students with substantially more IT skill and knowledge in considerably less time than IToF or ILE training.

# References

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

Brown, J. S., R. R. Burton, and A. G. Bell. 1975. “SOPHIE: A Step Toward Creating a Reactive Learning Environment.” *International Journal of Man-Machine Studies* 7: 675–96.

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

Clark, R. E. 1983. “Reconsidering Research on Learning from Media.” *Review of Educational Research* 53: 445–59.

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

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

Corno, L., and R. E. Snow. 1986. “Adapting Teaching to Individual Differences Among Learners.” In *Handbook of Research on Teaching,* 3rd. ed., edited by M. C. Wittrock, 605–29. New York, NY: Macmillan Publishing.

Fletcher, J. D. 1990. “Individualized Systems of Instruction.” In *Encyclopedia of Educational Research,* 6th ed., edited by M. C. Alkin, 613–20. New York, NY: Macmillan.

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

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

Fletcher, J. D., and R. E. Chatham. 2010. “Measuring Return on Investment in Military Training and Human Performance.” In *Human Performance Enhancements in High-Risk Environments,* edited by J. Cohn and P. O’Connor, 106–28. Santa Barbara, CA: Praeger/ABC-CLIO.

Glass, G. V., and B. McGaw. 1980. “Choice of the Metric for Effect Size in Meta-Analysis.” *American Educational Research Journal* 17: 325–27.

Graesser, A. C., and N. K. Person. 1994. “Question Asking During Tutoring.” *American Educational Research Journal* 31: 104–37.

Graesser, A. C., S. K. D’Mello, and W. Cade. 2010. “Instruction Based on Tutoring.” In *Handbook of Research on Learning and Instruction,* edited by R. E. Mayer and P. A. Alexander. New York: Routledge Press.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Kiszely, J. 2007. *Post-Modern Challenges for Modern Warriors.* The Shrivenham Papers, number 5. Shrivenham, UK: Defence Academy of the United Kingdom.

Luckin, R., K. R. Koedinger, and J. Greer, eds. 2007. *Artificial Intelligence in Education*. Amsterdam: IOS Press.

Mayer, R. E., and M. C. Wittrock. 1996. “Problem-Solving Transfer.” In *Handbook of Educational Psychology,* edited by D. C. Berliner, and R. C. Calfee, 47–62. New York: Macmillan.

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

Rosenthal, R. L. 1991. *Meta-Analytic Procedures for Social Research.* Newbury Park, CA: Sage.

Scriven, M. 1975. “Problems and Prospects for Individualization.” In *Systems of Individualized Education,* edited by H. Talmage, 199–210. Berkeley, CA: McCutchan.

Shuford, E. H., and T. A. Brown. 1974. *Rationale of Computer-Administered Admissible Probability Assessment.* R-1371-ARPA. Santa Monica, CA: RAND Corporation.

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

Wisher, R. A., M. A. Sabol, and J. A. Ellis. 1999. *Staying Sharp: Retention of Military Knowledge and Skills.* ARI Special Report 39. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (http://www.ari.army.mil).

# Abbreviations

| | |
|---|---|
| AFQT | Armed Forces Qualification Test |
| AOR | Area of Responsibility |
| CID | Center for Information Dominance [Navy] |
| DARPA | Defense Advanced Research Projects Agency |
| DT | Digital Tutor |
| FSET | Fleet Systems Engineering Team |
| ILE | Integrated Learning Environment |
| IT | Information Systems Technology [rating] |
| IToF | IT of the Future |

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

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

| | | |
|---|---|---|
| 1. REPORT DATE<br>February 2011 | 2. REPORT TYPE<br>Final | 3. DATES COVERED *(From–To)*<br>January 2011 – January 2011 |

4. TITLE AND SUBTITLE

DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor Assessments

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

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

J.D. Fletcher

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 NS D-4260
Log: H11-000117

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)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited. (4 March 2011)

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report presents findings from two assessments of the Digital Tutor (DT) being developed by the Education Dominance Program, which is sponsored by the Defense Advanced Research Projects Agency. The DT is intended to provide initial specialized skill training ("A" school and some additional "C" school training) for the Navy's Information Systems Technology (IT) rating. These assessments measured knowledge and skills of DT students and compared them with those of IT students trained by other means. Even though the DT was only partly finished, it was found to have produced substantially greater IT capabilities in less time than other training programs currently in use. Some of these differences exceeded two standard deviations in magnitude, and in one case, the difference was well above four standard deviations.

15. SUBJECT TERMS

Intelligent Tutoring Systems; Information Systems Technology; Training; Education; Computer-Based Instruction

| | | | | | |
|---|---|---|---|---|---|
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON<br>LTC William Casebeer |
| a. REPORT<br>Uncl. | b. ABSTRACT<br>Uncl. | c. THIS PAGE<br>Uncl. | SAR | 31 | 19b. TELEPHONE NUMBER *(include area code)*<br>703-526-4163 |