Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review

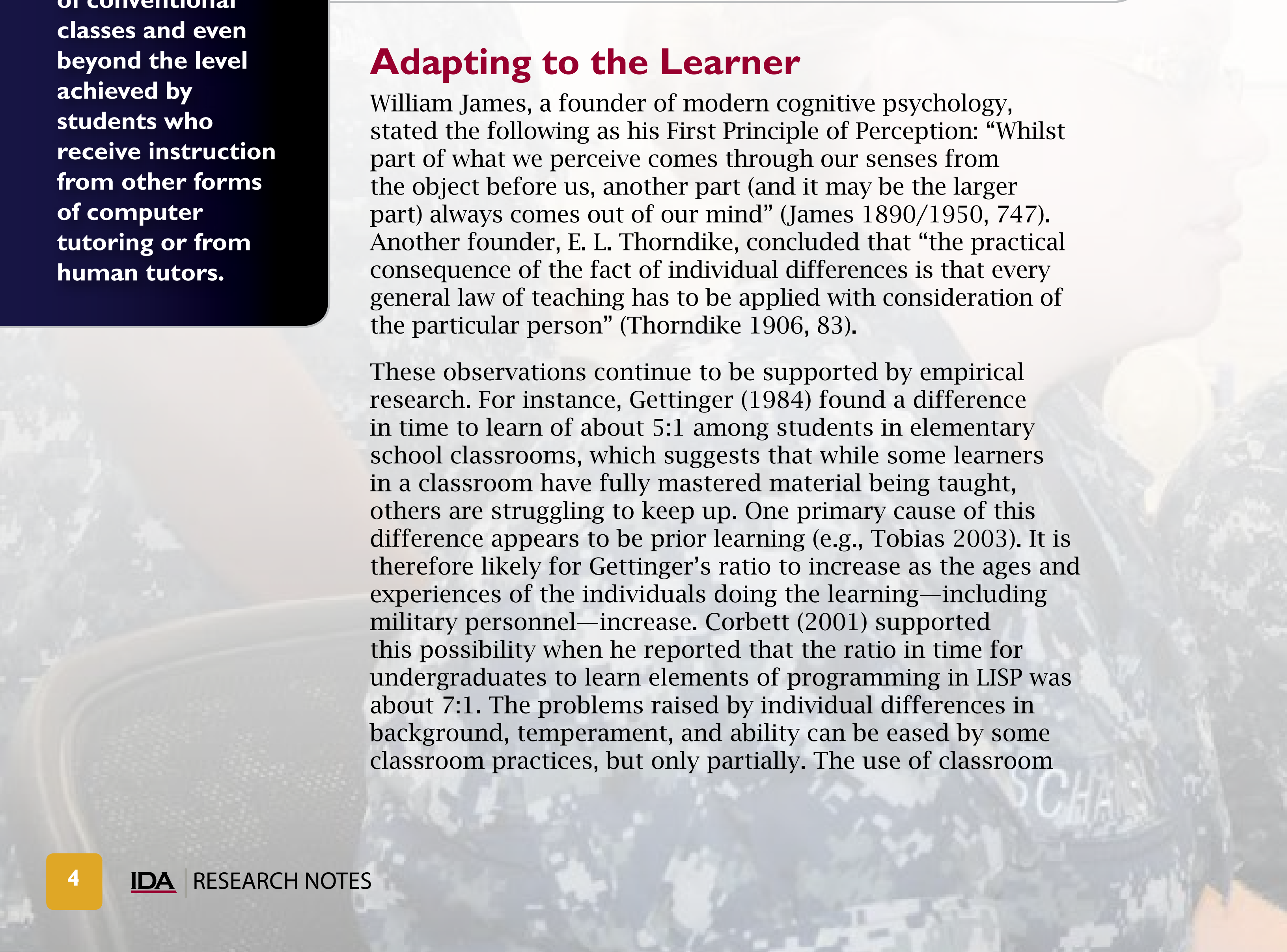
J. D. Fletcher and James A. Kulik

Military operations succeed or fail depending on the knowledge and skill of the soldiers, sailors, airmen, and marines who carry them out. However, the rapidly increasing technical complexity of military operations is raising the level of training needed to perform them. Research has found that one-on-one tutoring adapted to the specific needs, capabilities, and background of individual learners substantially increases learning well beyond that typically provided by classroom instruction. Unfortunately, training of this sort, delivered through the use of one-on-one human tutoring, is, except for rare instances, unaffordable. Nonetheless, it may become practicable through the use of computers employing machine intelligence to provide adaptive, individualized tutorial instruction. This article reviews efforts to build these intelligent computer-based systems and a recent meta-analysis to determine their effectiveness.

Our evaluations found that digital tutors typically raise student performance well beyond the level of conventional classes and even beyond the level achieved by students who receive instruction from other forms of computer tutoring or from human tutors.

Adapting to the Learner

William James, a founder of modern cognitive psychology, stated the following as his First Principle of Perception: "Whilst part of what we perceive comes through our senses from the object before us, another part (and it may be the larger part) always comes out of our mind" (James 1890/1950, 747). Another founder, E. L. Thorndike, concluded that "the practical consequence of the fact of individual differences is that every general law of teaching has to be applied with consideration of the particular person" (Thorndike 1906, 83).



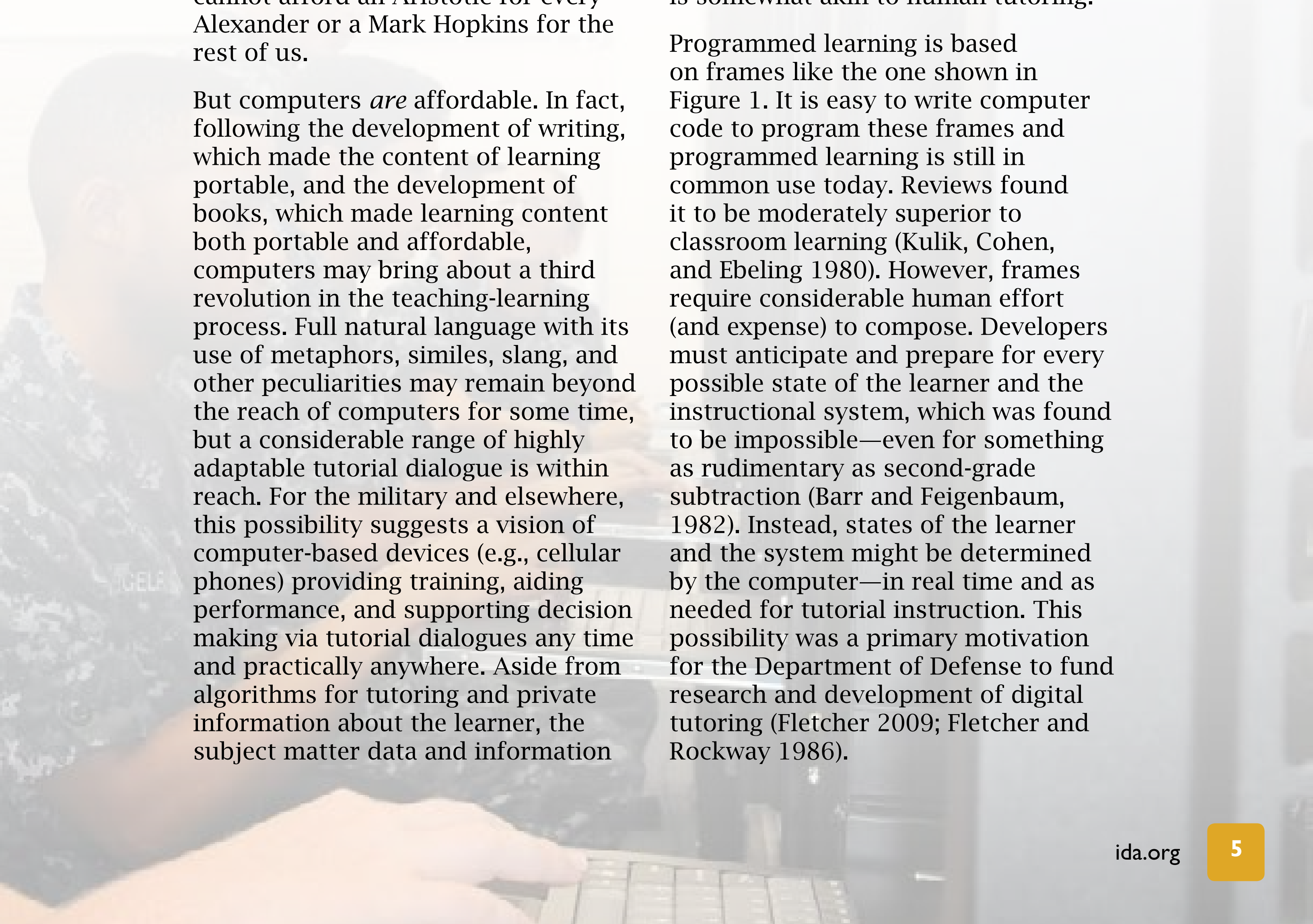
These observations continue to be supported by empirical research. For instance, Gettinger (1984) found a difference in time to learn of about 5:1 among students in elementary school classrooms, which suggests that while some learners in a classroom have fully mastered material being taught, others are struggling to keep up. One primary cause of this difference appears to be prior learning (e.g., Tobias 2003). It is therefore likely for Gettinger's ratio to increase as the ages and experiences of the individuals doing the learning—including military personnel—increase. Corbett (2001) supported this possibility when he reported that the ratio in time for undergraduates to learn elements of programming in LISP was about 7:1. The problems raised by individual differences in background, temperament, and ability can be eased by some classroom practices, but only partially. The use of classroom instruction continues to present unavoidable limits to efficiency and effectiveness in training and education.

These observations are supported by continuing research and theory that emphasize the idiosyncrasy of perception, cognition, memory, and learning. Bloom's frequently cited article (1984) suggested that one instructor tutoring one learner is vastly more effective than classroom instruction. Subsequent research strongly supports this view, but individual instruction is not affordable except for sensitive and critical activities (e.g., brain surgery and fighter piloting). Military training cannot afford an Aristotle for every Alexander or a Mark Hopkins for the rest of us.

But computers *are* affordable. In fact, following the development of writing, which made the content of learning portable, and the development of books, which made learning content both portable and affordable, computers may bring about a third revolution in the teaching-learning process. Full natural language with its use of metaphors, similes, slang, and other peculiarities may remain beyond the reach of computers for some time, but a considerable range of highly adaptable tutorial dialogue is within reach. For the military and elsewhere, this possibility suggests a vision of computer-based devices (e.g., cellular phones) providing training, aiding performance, and supporting decision making via tutorial dialogues any time and practically anywhere. Aside from algorithms for tutoring and private information about the learner, the subject matter data and information needed for tutoring need not be stored locally. It can be collected as needed from the global information grid and tailored to the background, needs, evolving capabilities, and even interests of the individual learner.

In the context of teaching and learning, classroom instruction is a relatively recent technology. For the last 65,000 years or so, most instruction was provided in one-on-one tutorial dialogues. Like many innovations (e.g., wireless telegraph and horseless carriages), computer-assisted instruction began by layering one technology (programmed learning textbooks) onto another (computers) to provide interactive instruction that is somewhat akin to human tutoring.

Programmed learning is based on frames like the one shown in Figure 1. It is easy to write computer code to program these frames and programmed learning is still in common use today. Reviews found it to be moderately superior to classroom learning (Kulik, Cohen, and Ebeling 1980). However, frames require considerable human effort (and expense) to compose. Developers must anticipate and prepare for every possible state of the learner and the instructional system, which was found to be impossible—even for something as rudimentary as second-grade subtraction (Barr and Feigenbaum, 1982). Instead, states of the learner and the system might be determined by the computer—in real time and as needed for tutorial instruction. This possibility was a primary motivation for the Department of Defense to fund research and development of digital tutoring (Fletcher 2009; Fletcher and Rockway 1986).



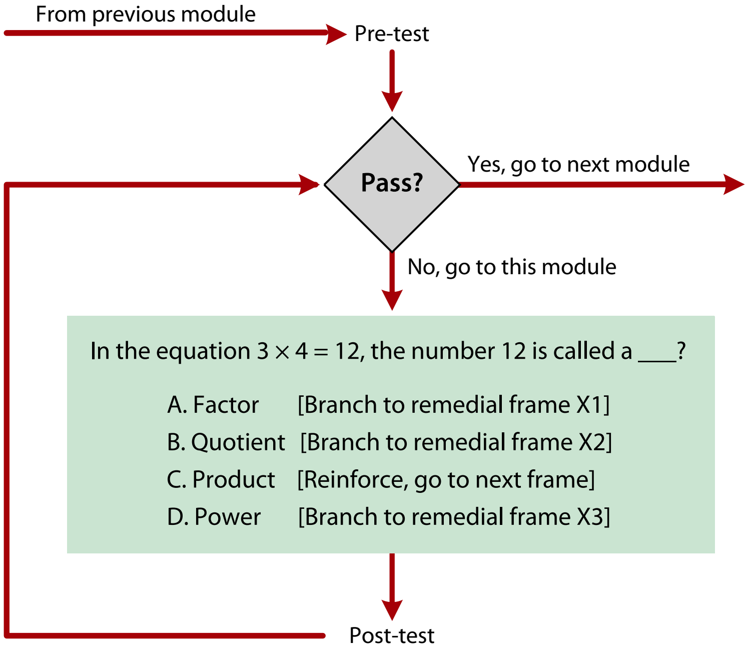
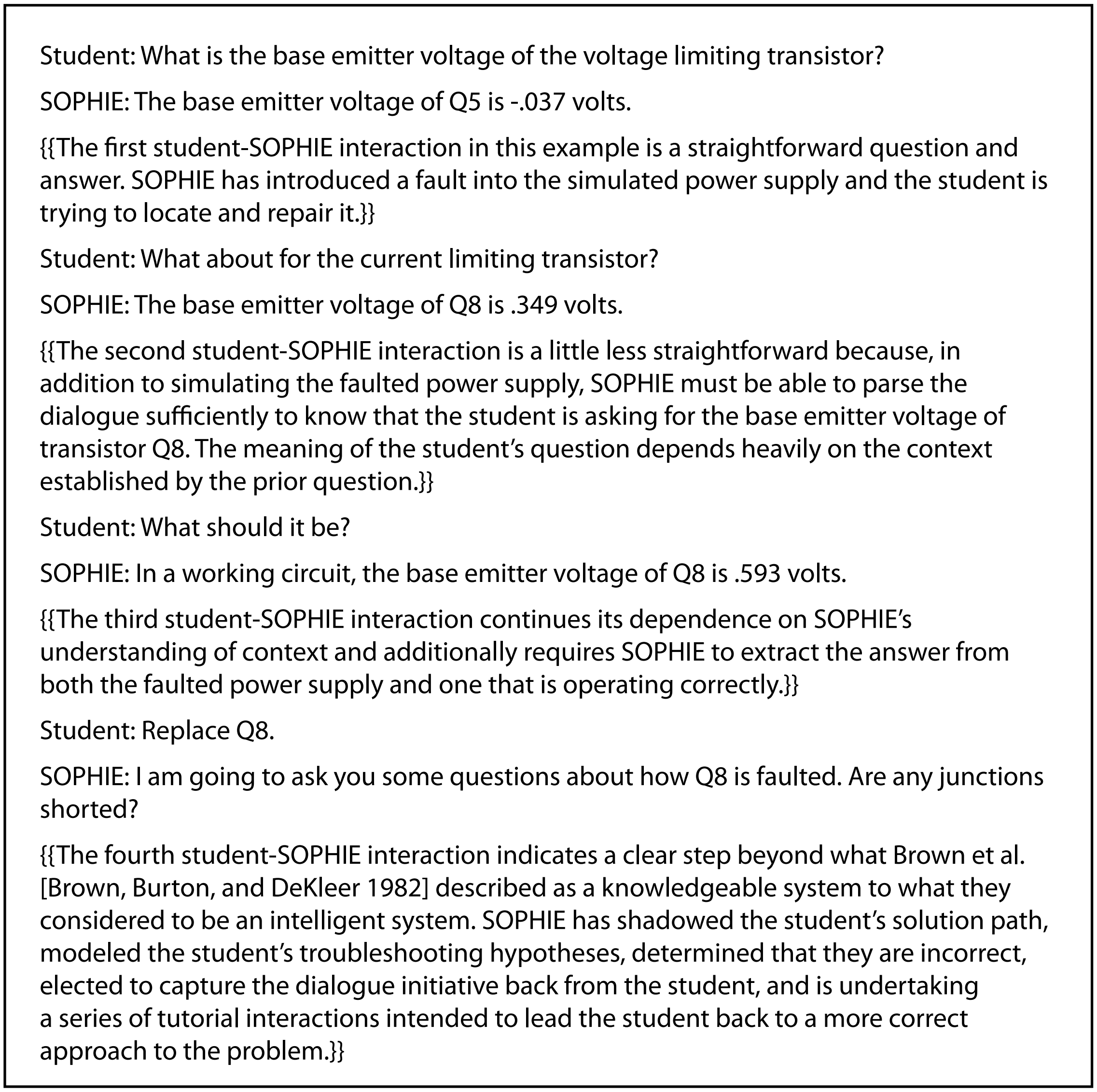

From previous module
Pre-test
Pass?
Yes, go to next module
No, go to this module
In the equation 3 × 4 = 12, the number 12 is called a ___?
A. Factor [Branch to remedial frame X1]
B. Quotient [Branch to remedial frame X2]
C. Product [Reinforce, go to next frame]
D. Power [Branch to remedial frame X3]
Post-test

Figure 1. Typical Intrinsic Programming Frame

Digital Tutoring

With support from the Office of Naval Research, Wallace Feurzeig determined that computers could do more than simply mimic programmed textbooks. He developed a computer language and a program (both called Mentor) to prepare learners to perform medical examinations (Feurzeig 1969). It was based on what Carbonell (1970) called information-structure-oriented instruction, using machine intelligence, as opposed to ad hoc frame-oriented instruction based on programmed learning techniques. Further, and in contrast to programmed learning approaches, Mentor allowed mixed initiative dialogues, which meant either the computer or the learner could initiate questions and lines of inquiry. This approach was later the basis for developing the Sophisticated Instructional Environment (SOPHIE) system to train electronic technicians (Brown, Burton, and DeKleer 1982). Example interactions between students and SOPHIE are provided in Figure 2.

Because tutorial dialogues typically include mixed initiative dialogue and information-structure orientation, we treated them as defining characteristics of instructional tutoring systems in our meta-analysis. Instructional tutoring systems typically apply four explicit models: (1) a model of the knowledge and skills to be acquired, (2) a dynamically evolving model of each learner and the learner's acquisition of these knowledge and skills, (3) a model of the instructional techniques that may be used by a specific learner to develop the knowledge and skills derived from the first two models, (4) a model for communication between the learner and the instructional system (Fletcher 1975).



Student: What is the base emitter voltage of the voltage limiting transistor?

SOPHIE: The base emitter voltage of Q5 is -.037 volts.

{{The first student-SOPHIE interaction in this example is a straightforward question and answer. SOPHIE has introduced a fault into the simulated power supply and the student is trying to locate and repair it.}}

Student: What about for the current limiting transistor?

SOPHIE: The base emitter voltage of Q8 is .349 volts.

{{The second student-SOPHIE interaction is a little less straightforward because, in addition to simulating the faulted power supply, SOPHIE must be able to parse the dialogue sufficiently to know that the student is asking for the base emitter voltage of transistor Q8. The meaning of the student's question depends heavily on the context established by the prior question.}}

Student: What should it be?

SOPHIE: In a working circuit, the base emitter voltage of Q8 is .593 volts.

{{The third student-SOPHIE interaction continues its dependence on SOPHIE's understanding of context and additionally requires SOPHIE to extract the answer from both the faulted power supply and one that is operating correctly.}}

Student: Replace Q8.

SOPHIE: I am going to ask you some questions about how Q8 is faulted. Are any junctions shorted?

{{The fourth student-SOPHIE interaction indicates a clear step beyond what Brown et al. [Brown, Burton, and DeKleer 1982] described as a knowledgeable system to what they considered to be an intelligent system. SOPHIE has shadowed the student's solution path, modeled the student's troubleshooting hypotheses, determined that they are incorrect, elected to capture the dialogue initiative back from the student, and is undertaking a series of tutorial interactions intended to lead the student back to a more correct approach to the problem.}}

Source: Foster and Fletcher (2002, 6-15–6-16).

Figure 2. Example of a Digital Tutoring Dialogue from SOPHIE

Meta-Analysis

Meta-analysis is a systematic, statistical technique for reviewing, combining, and summarizing quantitative results from many sources. It is frequently used in medicine and instruction to review the capabilities of a particular technique and provide an overall assessment of its effectiveness. Typically, it calculates statistical probabilities and effect sizes that compare one procedure with another. Statistical results determine the probability that a procedure will be superior in these comparisons (e.g., that a particular medical procedure will cure an ailment or that a particular instructional approach will produce more learning than another). Statistical results follow well-known procedures and identify differences that may be considered probabilistically significant.

However, it is not uncommon for one training procedure to have significant probability of being superior to another, but the difference between the two is so small it has little practical effect. Effect sizes provide a measure, in standard deviations or fractions of standard deviations, of practical significance—how far apart the results from two different approaches are from each other. Effect size is calculated by dividing the difference in results by an estimate of the standard deviation of the population, but discussion over how best to calculate effect size continues. For example, should the estimate be obtained from the standard deviations of all the samples, or should it consider the control group standard deviation alone? Effect sizes reported here are based on pooled standard deviations adjusted for sample size. In the parlance for effect sizes, this measure is known as Hedges's g.

Interpretations of effect sizes vary. A set of interpretations for training and education effect sizes is provided in Table 1. It suggests, in accord with the U.S. Department of Education, that an effect size should exceed 0.25 standard deviations to be worthy of consideration. Bloom (1984) stated that the ultimate goal for effect sizes in education and training research should be 2.00 standard deviations, but researchers in training and education properly celebrate finding an effect size of 0.80.

Results

As in all research, meta-analyses need to leave behind a sufficiently detailed trail of experimental procedures to allow replication. Four steps must be taken and reported clearly in specific detail: (1) identify procedures used to find relevant reports; (2) follow explicit procedures for coding findings from these reports; (3) compile and organize available measures of effectiveness; and (4) use statistical analysis and techniques for combining findings from the reports. Our meta-analysis assembled well over 500 candidate

Table 1. Overview of Effect Size

Effect Size (ES)	Suggested Designation[a]	50th Percentile (Roughly) Raised To . . .
ES < 0.25	Negligible[b]	60th percentile
0.25 < ES < 0.40	Small	60th–66th percentile
0.40 < ES < 0.60	Moderate	66th–73rd percentile
0.60 < ES < 0.80	Large	73rd–79th percentile
ES > 1.00	Very large	80th percentile and up
ES > 2.00	Bloom's challenge[c]	98th percentile and up

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

[b] What Works Clearinghouse (2010).

[c] Bloom (1984).

reports and found that 50 of them met the requirements for inclusion that we had established.

Findings in our meta-analysis of effectiveness of instructional tutoring systems ranged from –0.34 to 3.18. Effect sizes of the larger magnitude were found by Fletcher and Morrison (2014) for the Defense Advanced Research Projects Agency (DARPA) Digital Tutor, which may represent a breakthrough for digital tutoring technology. In 16 weeks, the DARPA Digital Tutor produced U.S. Navy Information System Technicians who scored much higher on tests of both knowledge and troubleshooting skill than other new sailors who had received 35 weeks of classroom training and experienced sailors who averaged 9 years of U.S. fleet experience. The monetary value of avoiding many years of on-the-job training is substantial. The operational value is likely to be larger, but it is more difficult to quantify—the loss of a Navy ship due to information technology failure is conceivable. Results of the DARPA Digital Tutor assessment were outliers for the meta-analysis and were Winsorized—a method to adjust for the statistical effect of extreme data points—by setting the values for its two upper outliers at the 95th percentile and setting the values for its two lowest outliers at the 5th percentile.

With our Winsorized data set, the median effect size was 0.66 overall, and the average effect size was 0.61. Roughly, this suggests an improvement of 50th percentile students to the 75th percentile. These findings are comparable to those of other reviews of digital tutoring techniques (e.g., VanLehn 2011). Our analysis suggests that instructional tutoring systems can provide unusually effective instruction. Students who received intelligent tutoring outperformed students from conventional classes in 46 (92 percent) of the 50 controlled evaluations. The improvement in learning was large enough to be considered statistically significant in 39 (78 percent) of the 50 studies.

Our evaluations found that digital tutors typically raise student performance well beyond the level of conventional classes and even beyond the level achieved by students who receive instruction from other forms of computer tutoring or from human tutors. Kulik and Kulik (1991) found an average effect size of 0.31 in 165 studies of computer-assisted instruction that did not at the time include digital tutoring. Digital tutoring gains are about twice that. Digital tutoring systems may also produce more learning than human tutoring, which typically raise student test scores about 0.40 standard deviations over control level (Cohen, Kulik, and Kulik 1982).

In conclusion, our meta-analytic findings, especially recent results showing effect sizes in excess of 3.00 with the DARPA Digital Tutor, suggest substantial improvements in the ability to provide education and training for military personnel and others. By accelerating learning and the acquisition of expertise, such improvements are likely to yield substantial monetary (Cohn and Fletcher 2010) and operational benefits.

References

Barr, A., and E. Feigenbaum, eds. 1982. "BUGGY." In *Handbook of Artificial Intelligence, Volume 2*, 279–282. Stanford: HeurisTech Press.

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, no. 6: 4–16.

Brown, J. S., R. R. Burton, and J. de Kleer. 1982. "Pedagogical, Natural Language and Knowledge Engineering in SOPHIE I, II, and III." In *Intelligent Tutoring Systems,* edited by D. Sleeman and J. S. Brown, 227–282. New York: Academic Press.

Carbonell, J. R. 1970. "Al in CAI: An Artificial Intelligence Approach to Computer-Assisted Instruction." *IEEE Transactions on Man-Machine Systems* 11, no. 4: 190–202.

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

Cohen, P. A., J. A. Kulik, and C.-L. C. Kulik. 1982. "Educational Outcomes of Tutoring: A Meta-Analysis of Findings." *American Educational Research Journal* 19: 237–248.

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.

Corbett, A. T. 2001. "Cognitive Computer Tutors: Solving the Two-Sigma Problem." In *User Modeling 2001: 8th International Conference, UM 2001, Sonthofen, Germany, July 13–17, 2001*, *Proceedings*, edited by M. Bauer, P. J. Gmytrasiewicz, and J. Vassileva, 137–147. Berlin: Springer-Verlag.

Feurzeig, W. 1969. *Computer Systems for Teaching Complex Concepts* (BBN Report 1742). Cambridge: Bolt Beranek & Newman, Inc. http://www.dtic.mil/get-tr-doc/pdf?AD=AD0684831.

Fletcher, J. D. 1975. "Modeling the Learner in Computer-Assisted Instruction." *Journal of Computer-Based Instruction* 3, no. 1: 118–126.

———. 2009. "Education and Training Technology in the Military." *Science* 323, no. 5910: 72–75. https://doi.org/10.1126/science.1167778.

Fletcher, J. D., and J. E. Morrison, 2014. *Accelerating Development of Expertise: A Digital Tutor for Navy Technical Training* Alexandria: Institute for Defense Analyses. Draft Final, Document D-5358.

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

Foster, R. E., and J. D. Fletcher. 2002. "Computer-Based Aids for Learning, Job Performance, and Decision Making in Military Applications: Emergent Technology and Challenges." Presented at the RTO HFM Symposium, *The Role of Humans in Intelligent and Automated Systems*, Warsaw, Poland, October 7–9, 2002, 6-1–6-24. RTO-MP-088.

Gettinger, M. 1984. "Individual Differences in Time Needed for Learning: A Review of Literature." *Educational Psychologist* 19, no. 1: 15–29. https://doi.org/10.1080/00461528409529278.

James, W. 1890/1950. *Principles of Psychology: Volume I*. New York: Dover Press.

Kulik, C.-L. C., and J. A. Kulik. 1991. "Effectiveness of Computer-Based Instruction: An Updated Analysis." *Computers in Human Behavior* 7, nos. 1–2: 75–94. https://doi.org/10.1016/0747-5632(91)90030-5.

Kulik, J. A., P. A. Cohen, and B. J. Ebeling. 1980. "Effectiveness of Programmed Instruction in Higher Education: A Meta-Analysis of Findings." *Educational Evaluation and Policy Analysis* 2, no. 6: 51–64.

Thorndike, E. L. 1906. *Principles of Teaching*. New York: A. G. Seiler & Company.

Tobias, S. 2003. "Extending Snow's Conceptions of Aptitudes." [Review of the book *Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow*, edited by L. J. Cronbach]. *Contemporary Psychology* 48, no. 3: 277–279. https://doi.org/10.1037/00078.

VanLehn, K. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46, no. 4: 197–221. https://doi.org/10.1080/00461520.2011.611369.

What Works Clearinghouse. 2010. *WWC Intervention Report: High School Math, Carnegie Learning Curricula and Cognitive Tutor Software*. Washington, DC: U.S. Department of Education.

J. D. (Dexter) Fletcher (left, with IDA President David S.C. Chu) is a Research Staff Member in the Science and Technology Division of IDA's Systems and Analyses Center. He holds a doctorate in educational psychology from Stanford University.

James Kulik (not pictured), an IDA consultant, holds a doctorate in psychology from the University of California at Berkeley.

The original Welch Award–winning publication was published in *Review of Educational Research* 86, no. 1 (March 2016): 42–78, https://doi.org/10.3102/0034654315581420.