Review of Educational Research
March 2016, Vol. 86, No. 1, pp. 42–78
DOI: 10.3102/0034654315581420

Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review

James A. Kulik
University of Michigan

J. D. Fletcher
Institute for Defense Analyses

This review describes a meta-analysis of findings from 50 controlled evaluations of intelligent computer tutoring systems. The median effect of intelligent tutoring in the 50 evaluations was to raise test scores 0.66 standard deviations over conventional levels, or from the 50th to the 75th percentile. However, the amount of improvement found in an evaluation depended to a great extent on whether improvement was measured on locally developed or standardized tests, suggesting that alignment of test and instructional objectives is a critical determinant of evaluation results. The review also describes findings from two groups of evaluations that did not meet all of the selection requirements for the meta-analysis: six evaluations with nonconventional control groups and four with flawed implementations of intelligent tutoring systems. Intelligent tutoring effects in these evaluations were small, suggesting that evaluation results are also affected by the nature of control treatments and the adequacy of program implementations.

Keywords: intelligent tutoring systems, computer-assisted instruction, tutoring, meta-analysis

Computer tutoring is a late development in the long history of tutoring in education. Whereas human tutoring has been used in schools for 2,500 years—or for as long as schools have existed—computer tutoring is largely a product of the past half century. The first computer tutoring systems to be used in school classrooms (e.g., R. C. Atkinson, 1968; Suppes & Morningstar, 1969) showed the influence of the programmed instruction movement of the time: They presented instruction in short segments or frames, asked questions frequently during instruction, and provided immediate feedback on answers (Crowder, 1959; Skinner, 1958). A different type of computer tutoring system appeared in research laboratories and classrooms during the 1970s and 1980s (e.g., Carbonell, 1970; Fletcher, 1985; Sleeman & Brown, 1982). Grounded in artificial intelligence concepts and cognitive theory, these newer systems guided learners through each step of a problem solution by creating hints and feedback as needed from expert-knowledge databases. The first-generation computer tutors have been given the retronym *CAI tutors* (for *computer-assisted instruction tutors*); the second-generation tutors are usually called *intelligent tutoring systems*, or *ITSs* (VanLehn, 2011).

VanLehn (2011) has summarized common beliefs about the effectiveness of different types of tutoring. According to VanLehn, CAI tutors are generally believed to boost examination scores by 0.3 standard deviations over usual levels, or from the 50th to the 62nd percentile. ITSs are thought to be more effective, raising test performance by about 1 standard deviation, or from the 50th to the 84th percentile. Human tutors are thought to be most effective of all, raising test scores by 2 standard deviations, or from the 50th to the 98th percentile.

These conventional views on tutoring effectiveness are based on research from decades ago. VanLehn (2011) attributed the belief that CAI tutors produce gains of around 0.3 standard deviations to a meta-analytic review of 165 studies (C.-L. C. Kulik & Kulik, 1991). He attributed the belief that ITSs produce 1–standard deviation gains to a widely cited article (Anderson, Corbett, Koedinger, & Pelletier, 1995) that summarized findings from several influential studies. The belief that human tutors raise student achievement levels by 2 standard deviations stems from an influential article by Bloom (1984), who coined the term *two-sigma problem*, to denote the search for other teaching approaches that are as effective as human tutoring.

More recent reviews support conventional beliefs about CAI tutoring effects. For example, a 1994 review aggregated results from 12 separate meta-analyses on computer-based instruction carried out at eight different research centers (J. A. Kulik, 1994). Each of the analyses yielded the conclusion that computer-based instruction improves student learning to a moderate degree. The median effect of computer-based instruction in the 12 meta-analyses was an increase in test scores of 0.38 standard deviations, or from the 50th to the 64th percentile. More recently, Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) reviewed results from 25 meta-analyses on instructional technology and learning. None of the analyses covered ITSs. Median effect of instructional technology in all 25 meta-analyses was an improvement in test scores of 0.35 standard deviations. Median effect in the 14 analyses that focused exclusively on CAI or computer-based instruction was an improvement of 0.26 standard deviations. Taken together, Tamim et al.'s (2011) and J. A. Kulik's (1994) reviews suggest that test score improvements of around one-third standard deviation are typical for studies of CAI tutoring.

It is much harder to find support for conventional beliefs about effects of human tutoring. Bloom (1984) based his claim for two-sigma tutoring effects on two studies carried out by his graduate students (Anania, 1981; Burke, 1980). Each of the studies compared performance of a conventionally taught control group with performance of two mastery learning groups, one taught with and one taught without the assistance of trained undergraduate tutors. Without tutoring, the mastery system raised test scores 1.2 standard deviations above control scores. Adding undergraduate tutors to the mastery program raised test scores an additional 0.8 standard deviations, yielding a total improvement of 2.0 standard deviations. This improvement is thus the combined effect of tutorial assistance plus special mastery learning materials and procedures. Neither Anania (1981) nor Burke (1980) evaluated the effects of tutoring alone. Because the studies confounded mastery and tutoring treatments, it is important to look beyond them for direct evidence on tutoring effects.

An early meta-analytic review (Hartley, 1977), which examined 29 studies of peer tutoring in elementary and secondary school mathematics, reported that tutoring programs raised math test scores by an average of 0.60 standard deviations. P. A. Cohen, Kulik, and Kulik (1982) reported an average improvement of 0.40 standard deviations in 65 studies of peer tutoring programs in elementary and secondary schools. Mathes and Fuchs (1994) found an improvement of 0.36 standard deviations in 11 studies of peer tutoring in reading for students with mild disabilities. G. W. Ritter, Barnett, Denny, and Albin (2009) examined the effectiveness of adult tutors in elementary schools and reported that tutoring improved student performance by 0.30 standard deviations in 24 studies. Finally, VanLehn (2011) summarized results from 10 studies of human tutoring, including Anania's (1981) study. The median effect of human tutoring in the 10 studies was a test score increase of 0.79 standard deviations. Without Anania's study, the median increase was 0.68 standard deviations. The median effect of human tutoring in the five meta-analyses was an improvement in performance of 0.40, far from Bloom's (1984) two-sigma effect.

Reviewers have not yet reached a consensus on the size of ITS effects on student learning. The most favorable conclusions come from early evaluations of Cognitive Tutor, the most widely used of all ITSs. Corbett, Koedinger, and Anderson (1997), for example, reported an average improvement in test scores of 1 standard deviation from early versions of Cognitive Tutor. They calculated this average from three sources: overall results reported by Anderson, Boyle, Corbett, and Lewis (1990) and Corbett and Anderson (1991); improvements on locally developed tests found in a study by Koedinger, Anderson, Hadley, and Mark (1997); and improvements for an experienced user of intelligent tutoring programs found in a study by Koedinger and Anderson (1993).

Four recent reviews have reported moderate effects from intelligent tutoring (Ma, Adesope, Nesbit, & Liu, 2014; Steenbergen-Hu & Cooper, 2014; U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2009; VanLehn, 2011). The What Works Clearinghouse review, the earliest of the four, focused on the use of Cognitive Tutor in middle school mathematics. The What Works evaluators found that only 1 of the 14 studies that they examined met their criteria for an acceptable evaluation. This study (S. Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007) reported that Cognitive Tutor improved student test scores by 0.38 standard deviations. What Works evaluators consider effects of 0.25 standard deviations and higher to be of substantive importance, so they classified this effect as a potentially important one.

The meta-analysis by VanLehn (2011) analyzed results from 54 comparisons of learning outcomes for ITSs and nontutored groups. The 54 comparisons were found in 28 separate evaluation studies. The average ITS effect in the 54 comparisons was an improvement in tests scores of 0.58 standard deviations. VanLehn classified the ITSs used in the 54 comparisons as either step based or substep based. Step-based tutoring provides hints and explanations on steps that students normally take when solving problems. Substep-based tutoring, which is a newer and more exacting approach, provides scaffolding and feedback at a finer level. Step-based tutoring, however, raised test scores by 0.76 standard deviations, whereas substep-based tutoring raised test scores by only 0.40 standard deviations. VanLehn's findings suggest, paradoxically, that older and simpler ITSs have strong effects on student performance, whereas newer and more sophisticated ITSs appear to be no more effective than "nonintelligent" CAI tutors.

The meta-analytic review by Steenbergen-Hu and Cooper (2014) examined 35 evaluations of ITS effectiveness in colleges. The researchers found that ITSs raised test scores overall by approximately 0.35 standard deviations, but they also reported that type of control group strongly influenced evaluation results. ITS scores were 0.86 standard deviations higher than control scores in evaluations where the control group received no instruction, 0.37 standard deviations higher in evaluations where the control group received conventional instruction, and 0.25 standard deviations lower than control scores in evaluations where the control group received human tutoring. Finally, the meta-analysis by Ma et al. (2014) analyzed 107 findings from 73 separate reports. The average ITS effect in the 107 comparisons was an improvement in test scores of 0.43 standard deviations. In addition, Ma et al. reported that ITS effects varied as a function of type of ITS used, nature of the control group in a study, outcome measure employed, and other factors.

Three recent reviews reported no real improvement in school performance due to the use of ITSs (Slavin, Lake, & Groff, 2009; Steenbergen-Hu & Cooper, 2013; U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2009). The reviews by the What Works Clearinghouse and Slavin et al. (2009) focused on Cognitive Tutor evaluations. The What Works reviewers examined 27 evaluations of Cognitive Tutor Algebra I in high schools. Only three of the evaluations met all the criteria for their analysis; three others met the criteria with reservations. Findings in the six evaluations were mixed, but the average effect was very near zero, a decrease in test scores from the 50th to the 49th percentile. Slavin et al. analyzed evaluations carried out in math courses in both middle and high schools. They located 13 evaluations, but only 7 of these met their requirements for acceptable studies. Cognitive Tutor raised student test scores by an average of 0.12 standard deviations in the seven evaluations. Slavin et al. considered this effect to be trivial. It was less than their cutoff (0.20 standard deviations) for effects of substantive importance.

The meta-analysis by Steenbergen-Hu and Cooper (2013) examined 26 reports on K–12 mathematics learning. Based on 34 comparisons described in the reports, Steenbergen-Hu and Cooper concluded that ITSs have very little or no overall effect on learning in these grades. Test scores of ITS and control students differed overall by around 0.05 standard deviations, a trivial amount. The researchers noted that ITS effects were positive and somewhat larger in studies that were less than 1 year in duration. Effects were decidedly negative, however, in two studies designed specifically to help students who were classified as lower achievers.

The lack of consensus about ITSs effectiveness is striking. Questions loom up on all sides. How effective are ITSs? Do they raise student performance a great deal, a moderate amount, a small amount, or not at all? If ITSs do have positive effects, has their effectiveness declined with the fine-tuning of the systems in recent years? What accounts for the striking differences in review conclusions about ITS effectiveness? This review uses meta-analytic methods to answer these questions.

Method

Glass, McGaw, and Smith (1981) identified four steps in a meta-analysis: (a) finding studies, (b) coding study features, (c) measuring study effects, and (d) statistically analyzing and combining findings.

Finding Studies

We used a two-stage procedure to find studies for this analysis. We first assembled a large pool of candidate reports through computer searches of electronic library databases. We then examined the candidate reports individually to determine whether they contained relevant data for a meta-analysis.

Candidate Reports

To find these, we carried out computer searches of databases from four sources: (a) the Educational Resources Information Clearinghouse (ERIC), (b) the National Technical Information Service, (c) ProQuest Dissertations and Theses, and (d) Google Scholar. We devised search strategies that took into account the characteristics of each of the databases:

1. The ERIC search focused on documents tagged with the descriptor *intelligent tutoring system* and one or more of the following descriptors: *instructional effectiveness, comparative analyses*, and *computer software evaluation*. The ERIC search yielded 104 reports.
2. The National Technical Information Service search focused on documents labeled with the text string *intelligent tutoring systems* in the subject field. This search yielded 120 documents.
3. The ProQuest Dissertations and Theses search targeted records containing both the text string *intelligent tutoring* and some form of the word *evaluate* in title, abstract, or keyword fields. The search yielded 98 dissertations.
4. The Google Scholar search focused on reports with the strings *intelligent tutoring, evaluation, control group*, and *learning* in the full document text. The search yielded 1,570 reports, which Google Scholar sorted by relevance to the search terms. We found many useful reports in the first documents listed by Google Scholar, but returns diminished quickly, and after 200 documents or so, Google Scholar stopped turning up useful new leads. We therefore added only the first 250 reports to our list of candidate reports.

We found additional candidate reports by branching from reference lists in reviews found in the four database searches. Two reviews were especially helpful:

VanLehn's (2011) review, which examined results of 28 ITS evaluations, and a Carnegie Learning (2011) reference list of 30 evaluations of its Cognitive Tutors. Taking into account the overlap in documents located in these searches, we estimate that our searches produced approximately 550 unique candidate reports for our analysis.

Final Data Set

After reviewing a small sample of candidate reports, we developed a list of requirements that evaluations had to meet to be considered acceptable for this meta-analysis. The most important requirement was that the treatment group actually received ITS instruction. CAI tutors continue to be developed, used, and evaluated, and it is possible to confuse these CAI systems with ITSs.

Carbonell (1970) was one of the first to draw a clear distinction between the two tutoring systems. According to Carbonell, computer tutoring systems are either frame oriented or information structure oriented. We now refer to Carbonell's frame-oriented tutors as CAI tutors; his information structure–oriented tutors are now known as ITSs. Frame-oriented tutors rely on frames, or prescripted blocks of material, to guide instruction. Information-structured tutors rely on organized knowledge databases, or information structures; computational and dialogue-generating tools extract relevant information from these structures to carry on tutorial interactions with learners. Carbonell thus emphasized two key defining features of intelligent tutors: (a) an information structure, or knowledge database, and (b) computational and dialogue-generating tools that extract relevant information from these structures.

Fletcher (1982, 1985) extended the definition of ITSs (then often called *intelligent computer-assisted instruction* or *ICAI*) to include three key features: (a) an explicit domain-knowledge model, which contains the foundations, concepts, and rules that experts understand and use in solving problems in the domain; (b) a dynamic student model, which keeps track of the student's state of knowledge with regard to the domain; and (c) a pedagogical module, which chooses tutoring strategies and actions to apply in specific situations for specific students. Anderson and his colleagues (Anderson et al., 1990; Anderson & Reiser, 1985) added a fourth defining feature: a user interface that students use to communicate flexibly with the system. For many years, these four structural characteristics were accepted as the defining features of ITS instruction.

VanLehn (2006) has noted that ITSs today come in different shapes, sizes, and designs, but whatever their structures, they all share common behavioral characteristics. To distinguish between CAI tutors and ITSs, VanLehn first described two types of tutoring behaviors: (a) *outer loop* behaviors, which give learners end-of-problem support, including appropriate feedback on their problem solutions and appropriate new problems to solve; and (b) *inner-loop* behaviors, which include prompting, hinting, and other support given while a student is working on a problem. In VanLehn's view, ITSs display both inner- and outer-loop behaviors, whereas CAI tutors display outer-loop behaviors only.

Although experts may differ on how to define intelligent tutoring, they usually agree on whether specific tutoring systems are intelligent or not. We therefore took a practical approach to the matter of identifying ITSs. We examined three factors before making final decisions. First, did the evaluator classify the computer tutor as an ITS? Second, do experts in the field also classify it as an ITS? Finally, does the computer tutor, like a human tutor, help learners while they are working on a problem and not just after they have recorded their solutions?

In addition to focusing on intelligent tutoring, evaluations had to meet seven other requirements:

1. Evaluations included in the meta-analysis could be either field evaluations or laboratory investigations, but all evaluations had to use an experimental or quasi-experimental design. Most of the 550 candidate reports found in the computer searches failed to meet this basic requirement. The pool of candidate reports included planning documents, reports on software development, impressionistic evaluations, case studies, review documents, and single-group studies. None of these provided results that could be used in our analysis.
2. Control groups had to receive conventional instruction. A control group could be either a conventional class or a specially constituted group that received instruction that closely approximated conventional teaching. Unacceptable for our analysis were evaluations in which control groups used materials that were extracted from ITS computer interactions, for example, *canned text* groups and vicarious-learning groups that studied script derived from ITS transcripts (e.g., Craig, Sullins, Witherspoon, & Gholson, 2006; Graesser et al., 2004; VanLehn et al., 2007). Also unacceptable were studies in which control groups were taught by human tutors or CAI tutors or received no relevant instruction.
3. Achievement outcomes had to be measured quantitatively and in the same way in both treatment and control groups. Results on both locally developed posttests and standardized tests were acceptable. Standardized tests included district, state, and national assessments, as well as published tests. School grades were not an acceptable outcome measure, because grades are often awarded on a different basis by different teachers in treatment and control classes. Also unacceptable for this meta-analysis were process measurements made during the course of a treatment.
4. The treatment had to cover at least one problem set or homework assignment, and the treatment duration had to be at least 30 minutes. Field evaluations were usually much longer in duration and easily met this requirement. Laboratory investigations usually covered only a small number of assignments or problem sets and were usually short in duration.
5. The treatment had to be implemented without major failures in the computer system or program administration. Excluded from this meta-analysis were results from implementations that were substantively disrupted by software or hardware failures.
6. Treatment and control groups had to be similar at the start of the evaluation. We eliminated from our data set any evaluation in which treatment and control groups differed by 0.5 standard deviations or more on pretests. Differences of this magnitude are too large to be adjusted by such

techniques as gain score or covariance analysis. Also eliminated were evaluations in which experimental and control groups were drawn from different populations (e.g., volunteers in the treatment group and nonvolunteers in the control group).

7. Overalignment of a study's outcome measure with treatment or control instruction was also a cause for excluding an evaluation from our analysis. Overalignment occurred, for example, when the outcome measure used test items that were included in the instructional materials for either the treatment or control group.

Only 50 of the 550 candidate reports described evaluations that met all of the above requirements and were thus qualified for use in the meta-analysis. Along with results from acceptable comparisons, a few of the 50 reports included results from unacceptable comparisons, for example, from comparisons with poorly implemented ITSs or inadequate control groups. Only results from the adequate comparisons were included in the meta-analysis.

Describing Evaluation Features

We used 15 variables to describe features of the evaluations (Table 1). Our selection of the 15 variables and coding categories was guided by our preliminary examination of the evaluations along with our examination of other reviews on intelligent tutoring and CAI tutoring. We originally coded some observations as continuous measurements (e.g., study year, sample size, and study length), but we later recoded the observations into ordered categories. The categorization helped solve analytic problems presented by skew, nonnormality, and presence of outliers in the continuous measurements.

Calculating Size of Effects

The *experimental effect size* is defined as the difference in posttest means for experimental and control populations, divided by the within-group population standard deviation (Glass et al., 1981). Meta-analysts estimate the population means and standard deviations from sample statistics included in research reports. We set up specific guidelines to help us choose the most appropriate sample statistics for estimating these population values.

Mean Differences

Whenever possible, we estimated mean differences from posttest means that were adjusted for pretreatment differences either by covariance or regression analysis. When studies did not report results from covariance or regression analysis, we estimated mean differences from pre–post gain scores of treatment and control groups. For studies that provided neither adjusted means nor gain score means, we estimated mean differences from raw posttest means. We set up these guidelines to maximize the precision of our estimates of treatment effects.

Standard Deviations

We used raw standard deviations, rather than adjusted ones, in calculating size of effect. Adjusted standard deviations include gain score and covariate-adjusted

TABLE 1

Fifteen study features and associated coding categories

Country (1 = United States, 2 = other)

Publication year (1 = up to 2000, 2 = 2001–2005, 3 = 2006 onward)

Grade level (1 = K–12, 2 = postsecondary)

Subject (1 = math, 2 = other)

Study type

1 = Experimental: short-term studies in which treatment and control groups work on the same assignments with or without intelligent tutoring

2 = Field evaluations: studies that compare performance in conventionally taught and intelligent tutoring classes

Sample size (1 = up to 80, 2 = 81–250, 3 = 251+)

Study duration (1 = up to 4 weeks, 2 = 5–16 weeks, 3 = 17+ weeks)

Intelligent tutoring system type (1 = step based, 2 = substep based)

Cognitive Tutor study

1 = No: not an evaluation of a current or earlier version of a Carnegie Learning Cognitive Tutor program

2 = Yes: evaluation of such software

Group assignment

1 = Intact groups: existing classes or groups assigned to treatment and control conditions

2 = Random: participants assigned randomly to conditions

Instructor effects

1 = Different instructors: different teachers taught treatment and control groups

2 = Same instructor: same teacher or teachers taught treatment and comparison groups

Pretreatment differences

1 = Unadjusted posttest: posttest means not adjusted for pretest differences

2 = Adjusted posttest: gain scores or posttest means adjusted by covariance or regression

Publication bias

1 = Published: study reported in a journal article, published proceedings, or book

2 = Unpublished: study reported in a dissertation or in a technical report

Test type

1 = Local: posttest was locally developed

2 = Standardized: posttest was a commercial, state, or district test

Test format

1 = Constructed-response items only: posttest was a problem-solving test, essay exam, etc.

2 = Both constructed-response and objective-test items: posttest included both constructed-response and objective-test items

3 = Objective items: posttest was a multiple-choice test or other test with a fixed alternative format

standard deviations as well as standard deviations derived from within-group variances in multifactor experimental designs. Experts usually caution against using such standard deviations in calculating size of effect (e.g., Borenstein, 2009; Glass et al., 1981). For reports that included only adjusted standard deviations, we estimated raw standard deviations using standard formulas and assuming a correlation of .60 between pretests and posttests. This is the median correlation in five studies in our data set that either reported pre–post correlations or presented data from which such correlations could be derived (Arnott, Hastings, & Allbritton, 2008; Fletcher, 2011; Pek & Poh, 2005; Suraweera & Mitrovic, 2002; VanLehn et al., 2007).

Glass's ES *and Hedges's* g

Tamim et al. (2011) reported that Hedges's *g* and Glass's *ES* were the two estimators of size of effect most often used in 25 meta-analyses on instructional technology conducted during the past four decades. Ten of the 25 meta-analyses, covering a total of 239 studies, used Hedges's *g* exclusively to report size of effects, whereas 6 meta-analyses, covering 505 studies, used Glass's *ES* exclusively. The remaining meta-analyses used either a different estimator of size of effect (e.g., a correlation coefficient), an unspecified estimator, or a combination of estimators.

Glass's *ES* and Hedges's *g* measure treatment effects in different ways. Glass's *ES* measures effects in control group standard deviations (Glass et al., 1981); Hedges's *g* uses pooled treatment and control standard deviations (Hedges & Olkin, 1985). We calculated both Hedges's *g* and Glass's *ES*, whenever possible, for studies in our data set and found a very high correlation between the two estimators (.97). The average Hedges's *g*, however, was 0.05 standard deviations lower than the average Glass's *ES* in the 36 studies for which we could make both estimates; median *g* was 0.08 standard deviations lower than the median *ES*. In addition, the values of the two estimators diverged more substantially in those cases where treatment and control standard deviations were significantly different (e.g., Fletcher, 2011; Fletcher & Morrison, 2012; Gott, Lesgold, & Kane, 1996; Hastings, Arnott-Hill, & Allbritton, 2010; Le, Menzel, & Pinkwart, 2009; Naser, 2009; Reif & Scott, 1999).

Using pooled standard deviations makes a great deal of sense when treatment and control standard deviations can be assumed to be equal. Pooling standard deviations is less justifiable when the two standard deviations are significantly different. For example, pooling is probably the wrong choice when a highly effective treatment brings all or almost all members of a heterogeneous population up to a uniformly high level of posttest performance. Such highly effective treatments can reduce standard deviations significantly below normal levels. Pooling standard deviations is probably also the wrong choice when a treatment affects different students very differently, for example, by greatly improving the performance of some while hampering the performance of others. Such treatments can raise standard deviations above normal levels.

We report results with both Glass's *ES* and Hedges's *g*, but we give primary emphasis to Glass's *ES* and treat Hedges's *g* as an important supplementary measure. Our preference for Glass's *ES* is based primarily on our reluctance to make a blanket assumption that control and treatment variances are equal in the studies, the assumption that is usually made when standard deviations are pooled. We found too many instances of unequal treatment and control variances for us to be comfortable with an assumption of no treatment effect on variance.

Statistical Analysis

A fundamental choice in any meta-analysis is whether to use weighted or unweighted means when combining estimators of size of effect. Glass et al. (1981) recommend using unweighted means. Hedges and Olkin (1985) recommend using weighted ones. The weights that Hedges and Olkin assign are different for fixed-effect and random-effects analyses. In fixed-effect analyses, where all studies can be assumed to share a common population effect, they weight the observed estimators of size of effect by the inverse of their standard errors, which is roughly equivalent to weighting by sample size. In random-effects analyses, where an assumption of a common underlying population effect is untenable, they use a more complex weighting system.

The high correlation between sample size and other important variables in our data set makes us cautious about weighting means fully or in part by sample size. For example, almost all of the large studies in our data set evaluated a single software program, Cognitive Tutor, and measured learning gains on off-the-shelf standardized tests rather than local tests tailored to local curricula. In addition, the large studies in our data set were longer in length and probably lower in implementation quality than small studies. If we used weighted means exclusively in our analyses, our conclusions would be very heavily influenced by a few large-scale evaluations of Cognitive Tutor. For example, if we assigned weights for a fixed-effect analysis, the largest evaluation in our data set (with 9,840 students) would receive nearly 750 times the weight of the smallest (with 24 students). With the weights assigned in a random-effects analysis, the largest evaluation receives about 5 times the weight of the smallest. Without weighting, each evaluation study receives the same weight.

We calculated unweighted means for our primary analyses, but we also calculated weighted means for our supplementary analyses. We calculated the weighted means using the procedures that Hedges and his colleagues developed for random-effects analyses (Comprehensive Meta-Analysis, Version 2.2.064). We do not include any results from fixed-effect analyses, because a fixed-effect model does not fit our data set. A fixed-effect model would not accurately represent the uniqueness and diversity of the individual treatments and measures used in the evaluations. In our experience, fixed-effect models are seldom if ever appropriate for meta-analytic data sets in education and the social sciences.

Another decision in meta-analysis involves the treatment of evaluation reports with multiple findings. Some meta-analysts report a single value for size of effect for each evaluation study; some report as many values as there are independent groups in the study. We used both approaches. We carried out our primary analyses with each study represented by a single value for size of effect, and we carried out supplemental analyses with each evaluation report represented by as many independent groups as were included in the evaluation.

Results

The 50 reports located for this meta-analysis are a diverse group (Table 2). They describe evaluations that were carried out on four continents over the course of nearly three decades. The content taught ranged from "borrowing" in third-grade subtraction to solving analytic problems from the Law School Admissions Test. The evaluations took place in elementary schools, high schools, colleges, and military training institutions. The shortest of the evaluations provided less than 1 hour of intelligent tutoring; the longest provided intelligent tutoring for three semesters, or 48 weeks.

Overall Effects

For our primary analysis, we used Glass's *ES* as the estimator of size of effect, evaluation study as the unit of analysis, and unweighted means to represent combined effects. Supplementary analyses used Hedges's *g* as the estimator of size of effect, both evaluation study and evaluation finding as units of analysis, and both weighted means and unweighted means to represent overall effects.

Primary Analysis

Students who received intelligent tutoring outperformed control students on posttests in 46 (or 92%) of the 50 studies. In 39 (or 78%) of the 50 studies, tutoring gains were larger than 0.25 standard deviations, or large enough to be considered of substantive importance by the standards of the What Works Clearinghouse (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2013). Thus, the vast majority of studies found ITS effects that were not only positive but also large enough to be important for instruction.

The strongest effects in the 50 evaluations were produced by the DARPA Digital Tutor, an ITS developed to teach U.S. Navy personnel the knowledge and skills needed by information systems technicians in duty station settings. The DARPA Digital Tutor was evaluated in two separate summative evaluations (Fletcher, 2011; Fletcher & Morrison, 2012). Each of the evaluations compared end-of-course test scores from a Digital Tutor course with scores from a standard classroom course. In the first evaluation, the Digital Tutor course lasted 8 weeks, and the classroom course, 17 weeks. In the second evaluation, the Digital Tutor course lasted 16 weeks, and the classroom course, 35 weeks. Both of the evaluations measured outcomes on locally developed, third-party tests: a 4-hour written test and a half-hour oral examination given by a review board. The first evaluation also included two tests of individual problem solving; the second evaluation included measurement of troubleshooting skills of three-member teams that responded to actual requests for shore-based assistance. Average *ES* in the first evaluation was 1.97 (Fletcher, 2011); average *ES* in the second evaluation was 3.18 (Fletcher & Morrison, 2012).

Both *ES*s are outliers, the only ones in the data set, where an outlier is defined as a high value that is at least 1.5 interquartile ranges above the 75th percentile or a low value that is at least 1.5 interquartile ranges below the 25th percentile. To keep these extreme values from having an undue influence on results, we formed a 90% Winsorized data set by substituting the value at the 95th percentile for these

(Text continues on p. 60.)

TABLE 2
Descriptive information and ES*s for 50 ITS evaluations*

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Anderson, Boyle, Corbett, and Lewis (1990); also Anderson, Corbett, Koedinger, and Pelletier (1995)	Geometry course, high school in Pittsburgh, PA, 1986–1987	5 classes	Geometry Tutor	SBT	1 quarter	Local	1.00
Arbuckle (2005)	Algebra I, public schools, Edmond, OK, Grades 9–11, 2003–2004	111 students (83 T, 28 C)	Cognitive Tutor	SBT	10 weeks	Local	0.74
Arnott, Hastings, and Allbritton (2008)	Research methods, DePaul University, Winter 2007	125 students (73 T, 52 C)	Research Methods Tutor	SSBT	5 weeks	Local	0.60
Arroyo, Royer, and Woolf (2011)	Grades 7 and 8 math classes	172 students (81 T, 91 C)	Wayang Outpost	SBT	4 days	Standardized	0.23
R. K. Atkinson (2007)	Reading comprehension, 3 high schools, and technical center, Phoenix metro area, AZ, 2004–2006	159 students (139 T, 20 C)	Gradations, STAR, and Read On!	SBT	8–12 weeks, about 36 hours	Standardized	0.25
Burns (1993)	Arithmetic, public school, Westchester County, NY, Grade 3	56 students (19 T, 37 C)	MEADOW	SBT	6 weeks, 1 20-minute session per week	Local	0.75
Cabalo and Vu (2007)	Algebra I, 5 high schools and 1 community college, Maui, HI, 2005–2006	345 students (182 T, 163 C)	Cognitive Tutor Algebra I	SBT	6 months	Standardized	0.03
Campuzano, Dynarski, Agodini, and Rall (2009)	Algebra I, 11 schools in 4 districts, Grades 8–9, 2004–2006	775 students (440 T, 315 C)	Cognitive Tutor Algebra I	SBT	1 school year, 24 weeks, 2,149 minutes	Standardized	−0.06
Carlson and Miller (1996)	Writing, 2 high schools, San Antonio, TX, 1993	852 students (429 T, 423 C)	Fundamental Skills Training Project's R-WISE 1.0	SBT	1 term, 9 sessions, 8 hours total	Local	0.78

(*continued*)

TABLE 2 (CONTINUED)

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Chang, Wang, Dai, and Sung (1999)	Computing course, universities in Taiwan	48 students (24 T, 24 C)	Web-Soc individualized tutoring	SBT	1 session, 50 minutes	Local	0.95
Corbett (2001b)	Prealgebra, North Hills Junior High academic classes, Grade 7, Pittsburgh, PA, 2000–2001	50 students (10 T, 40 C)	Cognitive Tutor Pre-Algebra	SBT	1 school year	Both local and standardized	0.46
Corbett (2002)	Prealgebra, Chartiers Valley Middle School academic classes, Grades 8–9, Pittsburgh, PA, 2001–2002	173 students	Cognitive Tutor Pre-Algebra	SBT	1 school year	Both local and Standardized	0.21
Corbett and Anderson (2001); also Corbett (2001a)	College course in LISP computer programming	40 students (30 T, 10 C)	ACT Programming Tutor	SBT	5 lessons, average of 7 sessions and 12 total hours	Local	1.00
Fletcher (2011)	Information technology systems, U.S. Navy Center for Information Dominance, Corry Station, Pensacola, FL, Fall 2010	40 students (20 T, 20 C)	DARPA Digital Tutor	SBT	8 weeks	Local	1.97
Fletcher and Morrison (2012)	Information technology systems, San Diego Naval Base, Winter 2011–2012	24 students (12 T, 12 C)	DARPA Digital Tutor	SBT	16 weeks	Local	3.18
Gott, Lesgold, and Kane (1996)	Electronic maintenance, 3 U.S. Air Force bases	41 students (18 T, 23 C)	Sherlock 2	SBT	—	Local	0.85
Graesser, Jackson, et al. (2003); reanalyzed in Graesser et al. (2004)	College physics, Universities of Memphis and Mississippi, and Rhodes College	29 students (21 T, 8 C)	Why/AutoTutor	SSBT	1 week, 2 sessions, 2–3 hours each	Local	0.78
Graesser, Moreno, et al. (2003); reanalyzed in Graesser et al. (2004)	Lesson in computer literacy at University of Memphis	81 students	AutoTutor 1.0 and 2.0	SSBT	1 session, 45–55 minutes	Local	0.17

(continued)

TABLE 2 (CONTINUED)

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Grubišic, Stankov, Rosic, and Žitko (2009)	Introduction to computer science, University of Split, Croatia, 2006–2007	39 students (20 T, 19 C)	xTex-Sys (eXtended Tutor-Expert System)	SBT	14 weeks	Local	1.23
Grubišic, Stankov, and Žitko (2006)	Introduction to computer science, University of Split, Croatia, 2005–2006	80 students (40 T, 40 C)	xTex-Sys (eXtended Tutor-Expert System)	SBT	14 weeks	Local	0.79
Hastings, Arnott-Hill, and Allbritton (2010)	Research methods, Chicago State University	87 students (56 T, 31 C)	Research Methods Tutor	SSBT	5 weeks, 2–4 hours total	Local	1.21
Hategekimana, Gilbert, and Blessing (2008)	Picture-editing, Iowa State University, Fall 2007	50 students (26 T, 24 C)	ITS	SBT	1 week, 2 sessions, total of 90 minutes	Local	−0.34
Jeremic, Jovanovic, and Gasevic (2009)	Upper-division software course, Military Academy, Belgrade, Serbia, Spring, 2006	42 students (14 T, 28 C)	DEPTHS (Design Patterns Teaching Helping System)	SBT	5 months	Local	0.62
Johnson, Flesher, Jehng, and Ferej (1993)	Electrical troubleshooting, University of Illinois at Urbana-Champaign, 1990–1991	34 students (18 T, 16 C)	Technical Troubleshooting Tutor	SBT	12 weeks, average of 5.25 total hours	Local	0.80
Koedinger, Aleven, Heffernan, McLaren, and Hockenberry (2004)	LSAT analytic problems, college in northeastern U.S.	30 students (15 T, 15 C)	LSAT Analytic Logic Tutor	SSBT	1 session, 1 hour	Local	0.78
Koedinger and Anderson (1993)	Geometry theorem proving, high school in Pittsburgh, PA, 1992	31 students	ANGLE	SBT	4–5 weeks, 25 class periods, 44 minutes each	Local	0.96
Koedinger, Anderson, Hadley, and Mark (1997)	Algebra, 3 high schools, Pittsburgh, PA, Grade 9, 1993–1994	590 students (470 T, 120 C)	PAT (Practical Algebra Tutor) and Pittsburgh Urban Math Project	SBT	1 year	Both local and standardized	0.68

(continued)

TABLE 2 (CONTINUED)

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Le, Menzel, and Pinkwart (2009)	Computer programing, University of Hamburg, Germany	35 students (18 T, 17 C)	INCOM	SBT	1 session, 1 hour	Local	0.31
Mendicino and Heffernan (2007)	Algebra, high school, rural area, 2004–2006	121 students	ITS	SSBT	1 session, 30–45 minutes	Local	0.63
Mendicino, Razzaq, and Heffernan (2009)	Mathematics, elementary school, rural area, Grade 5	28 students	ITS	SSBT	1 session	Local	0.55
Naser (2009)	C programming, Al-Azhar University of Gaza, Palestine	62 students (31 T, 31 C)	CPP-Tutor (C Intelligent Tutoring System)	SBT	1 month	Local	0.77
Pane, Griffin, McCaffrey, and Karam (2013)	Middle and high schools in 7 states	9840 students (4296 T, 5544 C)	Cognitive Tutor Algebra I	SBT	1 year	Standardized	0.20
Pane, McCaffrey, Slaughter, Steele, and Ikemoto (2010)	Geometry, 8 high schools, Baltimore County, MD, 2005–2008	699 students (348 T, 351 C)	Cognitive Tutor Geometry	SBT	1 term	Standardized	−0.19
Parvez and Blank (2007)	Object-oriented programming, Lehigh University summer high school program, Spring and Summer 2007	32 students (16 T, 16 C)	DesignFirstITS	SBT	—	Local	0.90
Pek and Poh (2005)	Engineering mechanics, Singapore Polytechnic	33 students (16 T, 17 C)	iTutor	SBT	1 session, about 80 minutes	Local	1.17
Person, Bautista, Graesser, Mathews, and The Tutoring Research Group (2001); also Graesser et al. (2004)	Computer literacy, University of Memphis	60 students	AutoTutor 1.1 and 2.0	SSBT	1 session, 45–55 minutes	Local	0.21
Phillips and Johnson (2011)	Financial Accounting, University of Saskatchewan	139 students	ITS	SBT	1 homework assignment	Local	0.39

(continued)

TABLE 2 (CONTINUED)

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Reif and Scott (1999)	Introductory physics, Carnegie Mellon University, Fall 1996	30 students (14 T, 16 C)	PAL (Personal Assistant for Learning)	SBT	1 week, 5 sessions, 7.5 hours	Local	0.78
Reiser, Anderson, and Farrell (1985); also Anderson et al. (1990)	LISP computer programming course, Carnegie Mellon University, 1984	20 students (10 T, 10 C)	GREATERP LISP Tutor	SBT	6 weeks	Local	1.00
S. Ritter, Kulikowich, Lei, McGuire, and Morgan (2007)	Algebra I, 3 junior high schools, Moore, OK, 2000–2001	257 students (153 T, 102 C)	Cognitive Tutor Algebra I	SBT	1 year	Standardized	0.40
Shneyderman (2001)	Algebra I, 6 high schools, Miami, FL, 2000–2001	777 students (325 T, 452 C)	Cognitive Tutor Algebra I	SBT	1 year	Standardized	0.22
Smith (2001)	Algebra I, 6 suburban high schools, Virginia City Beach, VA, 1999–2000	445 students (229 T, 216 C)	Carnegie Algebra Tutor	SBT	3 terms	Standardized	−0.07
Stankov, Glavinic, and Grubišic (2004)	Computer science, University in Croatia, 2004–2005	22 students (11 T, 11 C)	DTex-Sys (Distributed Tutor Expert System)	SBT	1 term, 15 weeks, 2 hours weekly	Local	1.16
Stankov, Rosic, Žitko, and Grubišic (2008)	Science, University of Split (1 study) and primary schools (8 studies), Split, Croatia, Grade 2 to first-year college, 2005–2007	380 students (190 T, 190 C)	xTex-Sys (eXtended Tutor-Expert System)	SBT	5 to 14 weeks	Local	0.74
Steuck and Miller (1997)	Scientific inquiry in ecology and biology, 15 junior and senior high schools in 5 states, Grades 7, 9, and 10, 1995–1996	1553 students (765 T, 788 C)	Fundamental Skills Training Project's ISIS (Instruction in Scientific Inquiry Skills)	SBT	36 weeks, 18 sessions, 18 hours total tutoring time	Local	0.37
Suraweera and Mitrovic (2002)	Database design, University of Canterbury, Christchurch, New Zealand, August 2001	62 students	KERMIT	SBT	1 session, about 1 hour	Local	0.56

(continued)

TABLE 2 (CONTINUED)

Publication	Subject and setting	Participants	Treatment	Steps	Duration	Posttest	*ES*
Timms (2007)	Lessons on force, motion, and speed, middle school science classes	131 students (101 T, 31 C)	Full FOSS tutor	SBT	Several days	Local	0.63
VanLehn et al. (2007), Experiment 2	Physics, University of Memphis and 3 other universities	48 students (32 T, 16 C)	Why2-AutoTutor and Why2-Atlas	SSBT	2 sessions, average of 126 total minutes	Local	0.70
VanLehn et al. (2005)	Introductory physics, U.S. Naval Academy, 2000–2003	912 students (89 T, 823 C)	Andes	SBT	1 term	Local	0.25
Wheeler and Regian (1999)	Word problems, 7 high schools, Texas, New Mexico, and Ohio, Grade 9, 1992–1993	493 students (409 T, 84 C)	Fundamental Skills Training Project's Word Problem Solving tutor	SBT	1 school year, one 50-minute session per week	Local	0.40

Note. T = treatment group; C = control group; ITS = intelligent tutoring systems; SBT = step-based tutoring; SSBT = substep-based tutoring; *ES* = Glass's estimator of effect size.

two outlier values and also substituting the value at the 5th percentile for the two lowest observed values. We report averages for both the original data set and the 90% Winsorized data set below.

The median *ES* in the original data set is 0.66. The mean *ES* is 0.65; the standard deviation is 0.56. In the Winsorized data set, median is 0.66, mean is 0.61, and standard deviation is 0.38. An improvement in test scores of 0.66 standard deviations over conventional levels is equivalent to an improvement from the 50th to the 75th percentile. According to J. Cohen (1988), an effect of 0.20 standard deviations is small, 0.50 standard deviations is medium size, and 0.8 standard deviations is large. By these standards, the average *ES* for intelligent tutoring is moderate to large.

Supplementary Analyses

We calculated the same statistics for the 63 independent comparisons included in the 50 studies. Results were affected very little by this change in unit of analysis. For example, the median *ES* is 0.63 in the data set of 63 independent comparisons. Mean *ES* is 0.62 without Winsorization and 0.59 with Winsorization. In 58 (or 92%) of the 63 comparisons, the ITS group scored higher than the control group; and in 49 (or 78%) of the comparisons, the improvement due to ITS use was substantively important, or more than 0.25 standard deviations.

Results were only slightly different when we calculated the same statistics for Hedges's *g* without weighting means. With evaluation study as the unit of analysis, the median *g* is 0.64 for the 50 cases. The mean *g* is 0.62 without Winsorization and 0.60 with Winsorization. With independent comparison as the unit of analysis, the median *g* is 0.61 for the 63 comparisons. The mean *g* is 0.59 without Winsorization and 0.57 with Winsorization. Results changed, however, when we used weighted means in the analysis. With evaluation report as the unit of analysis and weighting based on a random-effects model, the average $g = 0.50$, 95% CI [0.40, 0.59], $p < .001$. With evaluation finding as the unit of analysis and weighting based on a random-effects model, the average $g = 0.49$, 95% CI [0.40, 0.58], $p < .001$.

Evaluation Features and Effects Overall

Although ITSs most often improved learning by moderately large amounts, their effects were very large in some studies and near zero in others. To determine whether study features were related to the variation in results, we carried out a series of univariate analyses of variance (ANOVAs) with study feature as independent variable and size of effect as dependent variable.

Primary Analysis

The dependent variable in the primary analysis was Glass's *ES*, evaluation study was the unit of analysis, and the Winsorized data set was used to keep outliers from having an inordinate influence on the analysis. Results show that test type is the study feature most strongly related to *ES* (Table 3). *ES*s are large in evaluations that used local tests as outcome measures (average *ES* = 0.73), small in evaluations that used standardized tests (average *ES* = 0.13), and intermediate in evaluations that used a combination of the two (average *ES* = 0.45). Five additional study features are also strongly related to *ES*: sample size, grade level of

TABLE 3
Relationship between study features and study effects

Study feature	*r* with *ES*	Categories	Category *ES*		
			N	*M*	*SD*
Test type	−.63***	1 = Local	38	0.73	0.32
		2 = Local and standardized	3	0.45	0.24
		3 = Standardized	9	0.13	0.17
Sample size	−.55***	1 = Up to 80 participants	26	0.78	0.34
		2 = 81 through 250 participants	10	0.53	0.31
		3 = More than 250 participants	13	0.30	0.30
Grade level	.41**	1 = Elementary and high school	23	0.44	0.33
		2 = Postsecondary	27	0.75	0.36
Subject	.41**	1 = Mathematics	18	0.40	0.34
		2 = Other	32	0.72	0.35
Test item format	−.33*	1 = Constructed response only	15	0.84	0.26
		2 = Constructed and objective	14	0.47	0.36
		3 = Objective only	17	0.53	0.44
Cognitive Tutor study	−.28*	1 = No	35	0.68	0.34
		2 = Yes	15	0.45	0.42
Country	.26	1 = United States	39	0.56	0.38
		2 = Other	11	0.79	0.31
Publication bias	−.25	1 = Published	35	0.67	0.35
		2 = Unpublished	15	0.46	0.42
Publication year	−.21	1 = Up to 2000	12	0.78	0.21
		2 = 2001 through 2005	14	0.55	0.40
		3 = After 2006	24	0.56	0.41
Pretreatment differences	−.17	1 = Unadjusted posttest	14	0.72	0.39
		2 = Adjusted posttest	35	0.58	0.37
Group assignment	−.15	1 = Intact groups	32	0.67	0.38
		2 = Random assignment	13	0.55	0.38
Study duration	−.13*	1 = Up to 7 weeks	22	0.64	0.33
		2 = 8 weeks or more	25	0.54	0.42
Instructor effects	.10	1 = Different instructors	16	0.55	0.43
		2 = Same instructor	30	0.63	0.37
Tutoring steps	.02	1 = Step based	41	0.60	0.39
		2 = Substep based	9	0.63	0.31
Study type	.01	1 = Experimental study	15	0.60	0.33
		2 = Field evaluation	35	0.61	0.40

Note. *ES* = Glass's estimator of effect size.
**p* < .05. ***p* < .01. ****p* < .001.

participants, subject taught, test item format, and the tutoring system used in the evaluation. Specifically, study effects are smaller when (a) outcomes are measured on standardized rather than local tests, (b) sample size is large, (c) participants are at lower grade levels, (d) the subject taught is math, (e) a multiple-choice test is used to measure outcomes, and (f) Cognitive Tutor is the ITS used in the evaluation.

Supplementary Analyses

We carried out three parallel series of ANOVAs with the following estimators of effect magnitude and units of analysis: (a) Glass's *ES* as estimator and evaluation finding as unit of analysis, (b) Hedges's *g* as the estimator and evaluation study as the unit, and (c) Hedges's *g* as the estimator and evaluation finding as the unit. We used the 90% Winsorized data set in each of the analyses to keep outlier values from having an inordinate influence on results. The results of these ANOVAs are similar to the results in Table 3. Each set of analyses showed that test type was the study feature most strongly related to size of effect, and each found that the five other study features mentioned above were strongly related to size of effect.

We also carried out two supplementary analyses of the 90% Winsorized data set that used Hedges's *g* as the estimator of size of effect and Hedges's homogeneity procedures as the analytic method. Evaluation study was the unit in one analysis; evaluation finding was the unit in the other. Overall, these analyses confirmed the main ANOVA findings. As in other analyses, test type was the study feature most strongly related to study result. For example, in the homogeneity analysis of evaluation study results, average *g* was 0.62 when outcomes were measured on local tests, 0.09 when they were measured on standardized tests, and 0.46 when they were measured on both. In addition, all five of the other study features that were significantly related to *ES* in the analyses of variance were significantly related to *g* in these homogeneity analyses. However, the homogeneity analyses also detected significant but smaller relationships between Hedges's *g* and five other study features, including the method of assigning participants to treatment and control groups, the country in which the evaluation was conducted, the year in which it was conducted, the duration of the evaluation in weeks, and whether the evaluation report was published or not.

Key Study Features

It is important to note that many of the features that are significantly related to size of effect (Table 3) are highly intercorrelated. For example, standardized tests were used almost exclusively in large-scale evaluations of Cognitive Tutor Algebra in middle schools and junior high schools in the United States, and as a consequence, test type is highly correlated with sample size, subject taught, and grade level. The correlation is .60 between test type and sample size, .62 between test type and subject taught, and −.59 between test type and grade level.

A small number of underlying influences—perhaps a single factor—could easily account for many of the significant relationships between study features and size of effects in Table 3. To identify fundamental influences, we examined effects not only for different categories of studies but also for different conditions within studies. In addition, we examined findings in a few studies that could not be used in our main meta-analysis. We found that at least three factors had a substantive influence on evaluation findings: (a) the type of posttest used in a study, (b) the type of control group in the study, and (c) the fidelity of the ITS implementation.

Test Type

This is the study feature that distinguished most clearly between studies with large and small effects in both our primary and supplementary analyses. An early study by Koedinger et al. (1997) sheds light on the way that test type can influence evaluation results. The study examined effects of the Practical Algebra Tutor, an early version of Cognitive Tutor, on two types of posttests: locally developed tests that were aligned with the problem-solving objectives stressed in the program and standardized multiple-choice tests that did not stress problem solving. The researchers found large effects on the locally developed tests (mean *ES* = 0.99) and smaller effects on the standardized ones (mean *ES* = 0.36). They concluded that Practical Algebra Tutor was very effective in teaching the higher order skills it was designed to teach and that it did not negatively affect performance on standardized tests.

Later studies of Cognitive Tutor found the same pattern of results. For example, Corbett (2001b, 2002) examined the effects of Cognitive Tutor both on locally developed problem-solving tests and on multiple-choice tests consisting of released questions on international, national, and state assessments. For Grade 7 students, effects were large on the locally developed problem-solving tests (mean *ES* = 0.71) and trivial on the multiple-choice questions (mean *ES* = 0.18). For Grade 8 students, effects were small (mean *ES* = 0.28) on local problem-solving tests but even smaller on the multiple-choice questions (mean *ES* = 0.13).

The pattern holds up in the full set of 15 studies of Cognitive Tutor (Table 4). Overall, Cognitive Tutor raised student performance on locally developed tests significantly and substantially but neither helped nor hindered student performance on standardized tests. The mean *ES* on the standardized tests in the Cognitive Tutor evaluations is 0.12, whereas the mean *ES* on locally developed tests is 0.76. Median *ES* on standardized tests is 0.16; median *ES* on local tests is 0.86. That is, Cognitive Tutor boosted performance on locally developed problem-solving tests that were well aligned with its curricular objectives, but it did not boost performance on multiple-choice standardized tests that emphasized recognition skills.

We also conducted several analyses to determine whether study features were related to size of effect when type of test was held constant. We carried out these analyses with the 90% Winsorized sample to keep outliers from having an undue influence on results. We found that study features were not related to size of effect with test type held constant. There were no significant relationships between study features and effect magnitude in the 38 evaluation reports that measured outcomes on local tests, nor were there any in the 9 evaluations that measured outcomes on standardized tests. This was true whether the estimator of effect size was Glass's *ES* or Hedges's *g*. It also made no difference whether standard ANOVAs, correlations, or Hedges's homogeneity procedures were used to study the relationships.

Control Condition

In addition to examining studies with conventional control groups, we examined results in 6 reports, covering 11 separate experiments, with nonconventional control groups (Table 5). The nonconventional control groups were of two types. Control students in the first type of experiment read special materials that were

TABLE 4

Effects by test type in 15 Cognitive Tutor studies

	ES		
Publication	Local	Standardized	Overall
Anderson, Boyle, Corbett, and Lewis (1990)	1.00		1.00
Arbuckle (2005)	0.74		0.74
Cabalo and Vu (2007)		0.03	0.03
Campuzano, Dynarski, Agodini, and Rall (2009)		−0.10	−0.10
Corbett (2001b)	0.71	0.18	0.45
Corbett (2002)	0.28	0.13	0.21
Corbett and Anderson (2001)	1.00		1.00
Koedinger and Anderson (1993)	0.35		0.35
Koedinger, Anderson, Hadley, and Mark (1997)	0.99	0.36	0.68
Pane, McCaffrey, Slaughter, Steele, and Ikemoto (2010)		−0.19	−0.19
Pane, Griffin, McCaffrey, and Karam (2013)		0.20	0.20
Reiser, Anderson, and Farrell (1985)	1.00		1.00
S. Ritter, Kulikowich, Lei, McGuire, and Morgan (2007)		0.40	0.40
Shneyderman (2001)		0.22	0.22
Smith (2001)		−0.07	−0.07
Mdn	0.86	0.16	0.35

Note. *ES* = Glass's estimator of effect size.

derived from ITS computer interactions. The instructional material used by the control group therefore overlapped with ITS material. Graesser et al. (2004) referred to such control material as *textbook-reduced*; VanLehn et al. (2007) called it *canned text*. Control students in the second type of experiment simply viewed the recorded tutoring sessions of other students. The control students therefore received the same explanations and feedback as ITS students did but only for problems missed by paired, or *yoked*, students in the ITS group.

Effects are small in most of these studies. The strongest positive effect of tutoring in the six reports is an increase in posttest scores of 0.50 standard deviations; the largest negative effect was a reduction of −0.36 standard deviations. The median of the six *ES*s is 0.24, and the mean is 0.18. The mean *ES* is substantially lower than the mean *ES* (0.60) in evaluations with conventional control groups. We carried out several supplementary analyses of the data that varied the unit of

TABLE 5

Descriptive information and ES*s for six studies of intelligent tutoring systems without conventional control groups*

Publication	Subject and setting	Participants	Treatments	Steps	Duration	Posttest	*ES*
Craig, Driscoll, and Gholson (2004), Experiments 1 and 2	Computer literacy, University of Memphis	230 students (88 T, 142 C)	AutoTutor vs. yoked and vicarious tutoring controls	SSBT	1 session, 37 minutes	Local	0.46
Craig, Sullins, Witherspoon, and Gholson (2006), Experiments 1 and 2	Computer literacy, University of Memphis	267 students (61 T, 206 C)	AutoTutor vs. vicarious conditions	SSBT	1 session, 37 minutes	Local	0.50
Gholson et al. (2009)	Computer literacy, Grades 8 and 10; Newtonian physics, Grades 9 and 11	342 students (112 T, 230 C)	AutoTutor vs. 2 vicarious conditions	SSBT	1 session, 37 minutes	Local	0.04
Lane and VanLehn (2005)	Computer programming, University of Pittsburgh, Spring 2004	25 students (12 T, 13 C)	ProPl vs. reading same content	SSBT	6 weeks	Local	0.33
VanLehn et al. (2007), Experiments 1, 3, 5, and 6	Physics, University of Memphis and several other universities	290 students (188 T, 102 C)	Atlas and AutoTutor vs. canned text	SSBT	1 or 2 sessions, 2–4 hours total time	Local	0.14
Weerasinghe and Mitrovic (2006)	Database design, University of Canterbury, Christchurch, New Zealand, July 2002	94 students (35 T, 59 C)	KERMIT-SE vs. cutback KERMIT	SSBT	2 weeks (99 minutes for T, 105 for C)	Local	−0.36

Note. T = treatment group; C = control group; SBT = step-based tutoring; SSBT = substep-based tutoring; *ES* = Glass's estimator of effect size.

analysis and the estimator of effect magnitude. The supplementary analyses produced results that were similar to those in the primary one.

It should be noted that all six of the studies with nonconventional control conditions evaluated substep-based tutoring; none examined step-based tutoring. Two variables are thus confounded in the six studies: type of control condition and type of intelligent tutoring. Which of these is responsible for the depressed *ES*s in these studies? The six studies by themselves do not provide an answer, but we can answer the question by looking back at step-based and substep-based studies with conventional control groups (see Table 3). The mean *ES* in 41 studies of step-based tutoring with conventional control groups is 0.60, and the mean *ES* in 9 studies of substep-based tutoring with conventional control groups is 0.63. It therefore seems safe to conclude that the lower *ES*s in the six studies listed in Table 5 are attributable to the control conditions in the studies, not the type of ITS evaluated.

Implementation Adequacy

The adequacy of intelligent tutoring implementations also affects the strength of evaluation findings. Evidence on this point comes from four studies that reported data from both weaker and stronger implementations of an ITS. The median *ES* for the stronger implementations is 0.44; the median *ES* for the weaker implementations is −0.01. The evaluators who carried out these evaluations did not directly manipulate implementation adequacy in their studies. The variation in implementation adequacy resulted instead from technical or training weaknesses that affected part but not all of the experiments. The evaluators reported results in sufficient detail so that effects of the weaker and stronger parts of the experiments could be contrasted.

Koedinger and Anderson (1993), for example, compared results achieved by an experienced ITS teacher with results achieved by two teachers who were new to ITS instruction. In the hands of the experienced teacher, the ITS improved performance 0.96 standard deviations. In the hands of teachers with little prior experience with intelligent tutoring, the ITS had a negative effect on student performance; *ES* was −0.23. Teachers with limited experience treated the ITS as a replacement for the teacher, and they graded papers and worked on similar tasks while the students were working on the computer. The experienced teacher, on the other hand, thought that the ITS provided an opportunity for him to give more individualized help to students. When students were working with the ITS, he circulated around the classroom giving extra help to those who needed it and challenging other students with additional questions. When they did interact with students, the teachers with limited experience tended to focus on design features of the instructional technology, whereas the experienced teacher moved students quickly past the technology interface and directed their attention instead to the geometry content.

Le et al. (2009) examined the effects of a single 1-hour session of intelligent tutoring on student's logic programming skills. The intelligent tutoring session was held on two separate days. On the first day, the intelligent tutoring implementation was poor. Technical problems created long delays in the computer tutor's responses (e.g., 1-minute delays). The average *ES* for intelligent tutoring on the first day was 0.01. Technical problems were resolved by the second day of the experiment, and the average *ES* for intelligent tutoring rose to 0.31.

Pane, Griffin, McCaffrey, and Karam (2013) found significantly different effects during the first and second years of an implementation of Cognitive Tutor Algebra I. Nearly 10,000 Algebra I students were included in the evaluation during the first year of the Cognitive Tutor program, and another 10,000 students were included during the second year. Pane et al. reported that Cognitive Tutor had no significant effect on student test scores when teachers were using it for the first time (mean *ES* = −0.06), but it had a small but highly significant positive effect when teachers used it for a second time (mean *ES* = 0.20).

Finally, VanLehn et al. (2005) reported results from 5 years of use of the Andes tutoring system at the U.S. Naval Academy. In the first year, the Andes system presented students with relatively few physics problems and the program contained a relatively large number of bugs. In the first year of the program, *ES* for hour exams was 0.21. In the second through fifth years of the program, the number of physics problems was increased, and bugs were fixed. Average *ES* for hour exams for these 5 years was 0.57.

Discussion

This meta-analysis shows that ITSs can be very effective instructional tools. Students who received intelligent tutoring outperformed students from conventional classes in 46 (or 92%) of the 50 controlled evaluations, and the improvement in performance was great enough to be considered of substantive importance in 39 (or 78%) of the 50 studies. The median *ES* in the 50 studies was 0.66, which is considered a moderate-to-large effect for studies in the social sciences. It is roughly equivalent to an improvement in test performance from the 50th to the 75th percentile.

This is stronger than typical effects from other forms of tutoring. C.-L. C. Kulik and Kulik's (1991) meta-analysis, for example, found an average *ES* of 0.31 in 165 studies of CAI tutoring. ITS gains are about twice as high. The ITS effect is also greater than typical effects from human tutoring. As we have seen, programs of human tutoring typically raise student test scores about 0.4 standard deviations over control levels. Developers of ITSs long ago set out to improve on the success of CAI tutoring and to match the success of human tutoring. Our results suggest that ITS developers have already met both of these goals.

ITS effects are also robust. The 50 controlled evaluations we reviewed took place at different times, in different places, and in different educational settings. Although the settings were diverse, moderately strong ITS effects were the rule. For example, the 50 evaluations in our meta-analysis were carried out in nine countries on four continents. A total of 39 (or 78%) of the studies were done in the United States, where ITSs were first developed, and 11 (or 22%) were done outside the United States. The average *ES* found in studies conducted within the United States was 0.56; the average *ES* in studies conducted outside the United States was 0.79. It appears therefore that ITSs have not only traveled far from their country of origin but also traveled well. They appear to be just as effective abroad as they are at home.

We found one important exception to the rule of moderately strong positive effects in the 50 controlled evaluations. Although effects were moderate to strong in evaluations that measured outcomes on locally developed tests, they were much smaller in evaluations that measured outcomes on standardized tests. Average *ES* on studies with local tests was 0.73; average *ES* on studies with standardized tests was 0.13. This discrepancy is not unusual for meta-analyses that include both local and standardized tests. A meta-analysis by Rosenshine and Meister (1994), for example, found that reciprocal teaching systems raised student performance 0.88 standard deviations on local tests but only 0.32 standard deviations on standardized tests. A meta-analysis by C.-L. C. Kulik, Kulik, and Bangert-Drowns (1990) found mastery learning systems boosted student performance by 0.57 standard deviations on local tests but by only 0.29 standard deviations on standardized tests.

Which kind of test should we trust? Both local and standardized tests have their champions. Some evaluators prefer local tests, because local tests are likely to align well with the objectives of specific instructional programs. Off-the-shelf standardized tests provide a looser fit. Evaluators who prefer standardized tests, on the other hand, usually praise them for being free of bias. Unlike local tests, which may be written by developers or supporters of an experimental program, standardized tests are almost always third-party affairs. The authors of standardized tests can hardly slant them to favor one group or another in future evaluation studies.

Our own belief is that both local and standardized tests provide important information about instructional effectiveness, and when possible, both types of tests should be included in evaluation studies. We think that Koedinger et al. (1997) were on the right track when they included both standardized and local tests in their pioneering ITS evaluation. They found strong ITS effects on local tests that were aligned with the curriculum and smaller effects on standardized tests that were not. The ITS thus improved the problem-solving skills it was designed to teach, and the improvement in problem solving came at no cost to the recognition skills emphasized on standardized tests. We suspect that the same conclusion may be appropriate for ITSs in general. Only the wider use of both standardized and local tests in ITS evaluations will provide conclusive evidence.

Another factor that affects ITS evaluation results is the type of control group used in a study. Specifically, results are different for studies with conventional and nonconventional control groups. Median *ES* is 0.66 in studies with conventional control groups. Median *ES* is 0.28 in studies with nonconventional control groups that were taught with materials derived from the ITS interactions. Studies with nonconventional control groups can be useful in determining how ITSs work, but they do not give a useful answer to the question of overall ITS effectiveness.

A third factor that can influence results of an intelligent tutoring program is the adequacy of the program implementation. Very few ITS evaluations measured implementation adequacy directly, but four studies suggested that intelligent tutoring effects are stronger when programs are carefully implemented and weaker when programs are not implemented expertly or when technical problems affect implementations. It is not clear whether implementation adequacy affected other studies in our data set beyond these four. On the one hand, we did not include in our main analyses findings from implementations with reported inadequacies, so the effect might be small. On the other hand, ITSs were a novelty to teachers in some large studies included in our analyses, and the teacher's limited experience with ITSs may have affected results in their classrooms.

Our meta-analytic findings shed light on some otherwise puzzling conclusions reached in other reviews of ITS findings. Reviews of Cognitive Tutor evaluations, for example, have drawn contradictory conclusions about its effectiveness. Early reviews reported strong improvements in student performance due to Cognitive Tutor (e.g., Corbett et al., 1997), but recent reviews have reported that Cognitive Tutor has little or no consistent effect on student learning (e.g., Slavin et al., 2009; U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2013). We found that review findings depend on the proportion of reviewed studies that used locally developed tests. Early reviews, which reported strong positive improvements, based their conclusions entirely on findings from local tests. Recent reviews that reported little or no positive improvements from Cognitive Tutor based their conclusions entirely on results from standardized tests. We found a median *ES* of 0.86 on local tests used in Cognitive Tutor evaluations, a median *ES* of 0.16 on standardized tests, and a median *ES* of 0.35 for all tests used in Cognitive Tutor evaluations.

Our analysis also sheds light on an unexpected finding in VanLehn's (2011) review on tutoring effects. Specifically, VanLehn found an average size of effect of 0.76 for an older and less exacting form of ITS, which he called step-based tutoring. He found an average size of effect of only 0.40 for substep-based ITSs, a newer and more rigorous approach. We found similar effects for step-based and substep-based ITSs in studies with conventional control groups. However, we found smaller effects in studies of substep-based tutoring with nonconventional control groups. We excluded studies with nonconventional control groups from our meta-analysis, but VanLehn included them in his analyses. The low average size of effect that he reported for substep-based tutoring thus seems to be due more to the type of control groups in VanLehn's studies than to substep-based tutoring itself.

Our findings are clearly different from those of Steenbergen-Hu and Cooper (2013), who reported that ITSs had no real effect on K–12 math performance. They found an average effect of about 0.05 standard deviations in the 26 studies included in their meta-analysis. In contrast, we found an average *ES* of 0.40 in 18 studies of ITS effectiveness in elementary and high school mathematics. The average *ES* was 0.72 in seven studies that measured outcomes on local tests, 0.45 in three studies that measured outcomes on both standardized and local tests, and 0.10 in eight studies that measured outcomes only on standardized tests.

No single factor is responsible for the difference in findings of our meta-analysis and Steenbergen-Hu and Cooper's (2013), but it is important to note that the two meta-analyses defined ITSs differently. Steenbergen-Hu and Cooper defined ITSs as "self-paced, learner-led, highly adaptive, and interactive learning environments operated through computers" (p. 983). This broad definition led them to include in their meta-analysis a number of computer systems that are not ordinarily considered to be ITSs. Specifically, their meta-analysis included evaluations of such CAI systems as iLearnMath, Larson Pre-Algebra, Larson Algebra, Plato Algebra, Plato Achieve Now, and an online remediation system used in a study by

Biesinger and Crippen (2008). These systems are not classified as ITSs by the developers of the systems, and they would not be considered to be ITSs by most experts on intelligent tutoring. To use VanLehn's terminology, these systems are answer-based CAI tutors. They can provide feedback on student answers but not on the thinking that goes into individual answers. We therefore excluded evaluations of these and other CAI systems from our meta-analysis.

It is also important to note that Steenbergen-Hu and Cooper (2013) had looser requirements than we did for acceptable control groups, and they included in their meta-analysis a number of evaluations without adequate control groups. For example, their meta-analysis included evaluations by Beal, Walles, Arroyo, and Woolf (2007); Plano (2004); and Walles (2005) in which treatment and control groups differed substantially in pretest scores. The difference was equivalent to 0.81 standard deviations in Beal's study, 1.09 standard deviations in Plano's, and 0.76 standard deviations in Walles's. Also included in Steenbergen-Hu and Cooper's review were studies with no-instruction controls (Beal, Arroyo, Cohen, & Woolf, 2010; Biesinger & Crippen, 2008; Radwan, 1997) and studies that provided no evidence of baseline equivalence of groups (Carnegie Learning Inc., 2001; Corbett, 2002; Koedinger, 2002; Sarkis, 2004). We excluded these studies from our analysis, because they did not appear to provide a fair baseline for assessing the contributions that ITSs might make.

Overall, the message from what we judge to be fair comparisons of ITS and conventional instruction seems clear. The evaluations show that ITSs 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. Although a small minority of ITS studies found no significant difference in performance of ITS and control students, most of these studies were weak in design or execution. Some measured outcomes solely on off-the-shelf tests that were poorly aligned with the higher order curricular objectives emphasized in ITS programs. Other studies used nonconventional control groups that studied special materials that were derived from ITS interactions. Still other studies suffered from poorly implemented ITS treatments. When results from such questionable comparisons are left out of the mix, the message from ITS evaluations is clear, consistent, and positive.

It is hard to predict the exact shape that computer tutoring will take in the future. In effect, we may be at the "wireless telegraph" phase, with radio yet to be developed. Advances are surely coming on a number of fronts—in computer hardware, software, networking, and cognitive science—and these advances will likely affect both the appearance and structure of future tutoring systems. It remains to be seen whether tomorrow's computer tutors will produce the two-sigma improvements that have so far eluded most ITS developers, but the available evidence suggests that today's ITSs can serve as a sound foundation for future work.

Note

The research was carried out with support from the Office of Naval Research.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Anania, J. (1981). *The effects of quality of instruction on the cognitive and affective learning of students* (Doctoral dissertation). Available from ProQuest Dissertations & Theses database. (Order No. T-28171)

*Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, *42*, 7–49. doi:10.1016/0004-3702(90)90093-F

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*, 167–207. doi:10.1207/s15327809jls0402_2

Anderson, J. R., & Reiser, B. J. (1985). The LISP tutor. *Byte*, *10*, 159–175. Retrieved from http://www.psychology.nottingham.ac.uk/staff/com/c8clat/resources/TheLISPTutor.pdf

*Arbuckle, W. J. (2005). *Conceptual understanding in a computer-assisted Algebra 1 classroom* (Doctoral dissertation). Available from Proquest Dissertations & Theses database. (Order No. 3203318)

*Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods*, *40*, 694–698. doi:10.3758/BRM.40.3.694

*Arroyo, I., Royer, J. M., & Woolf, B. P. (2011). Using an intelligent tutor and math fluency training to improve math performance. *International Journal of Artificial Intelligence in Education*, *21*, 135–152. doi:10.3233/JAI-2011-020

Atkinson, R. C. (1968). Computerized instruction and the learning process. *American Psychologist*, *23*, 225–239. doi:10.1037/h0020791

*Atkinson, R. K. (2007). *An experimental evaluation of three computer-based reading comprehension tutors* (Final Report ONR N00014-05-1-0129). Tempe: Division of Psychology in Education, Arizona State University.

Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning*, *9*, 64–77.

Beal, C. R., Walles, R., Arroyo, I., & Woolf, B. P. (2007). On-line tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning*, *6*, 43–55.

Biesinger, K., & Crippen, K. (2008). The impact of a state-funded online remediation site on performance related to high school mathematics proficiency. *Journal of Computers in Mathematics and Science Teaching*, *27*, 5–17.

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*(6), 4–16. doi: 10.3102/0013189X013006004

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–236). New York, NY: Russell Sage.

Burke, A. J. (1980). *Students' potential for learning contrasted under tutorial and group approaches to instruction*. (Doctoral dissertation). Available from ProQuest Dissertations & Theses database. (Order No. T-28810)

*Burns, L. M. (1993). *MEADOW: An integrated system for intelligent tutoring of subtraction concepts and procedures*. (Doctoral dissertation). Available from Proquest Dissertations & Theses database. (Order No. 9333735)

*Cabalo, J., & Vu, M. (2007). *Comparative effectiveness of Carnegie Learning's Cognitive Tutor Algebra I curriculum: A report of a randomized experiment in the Maui School District*. Palo Alto, CA: Empirical Education. Available from ERIC database. (ED538963)

*Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts* (Report No. NCEE 2009-4041). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Available from ERIC database. (ED504657)

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. *IEEE Transactions on Man-Machine Systems*, *11*, 190–202. doi:10.1109/TMMS.1970.299942

*Carlson, P. A., & Miller, T. M. (1996). *Beyond word processing: Using an interactive learning environment to teach writing* (Report No. AL/HR-TR-1996-0090). Brooks AFB, TX: Human Resources Directorate, Technical Training Research Division. Available from DTIC Online database. (ADA319034)

Carnegie Learning. (2001). *Report of results from Canton, Ohio* (Cognitive Tutor Research Report OH-01-01). Pittsburgh, PA: Author. Retrieved from https://www.carnegielearning.com/research-results/whitepapers-reports/

Carnegie Learning. (2011). *Cognitive tutor evaluation*. Unpublished manuscript, Carnegie Learning, Pittsburgh, PA. Retrieved from https://www.carnegielearning.com/research-results/whitepapers-reports/references/cognitive-tutor-evaluation

*Chang, K.-E., Wang, K.-Y., Dai, C.-Y., & Sung, T.-C. (1999). Learning recursion through a collaborative Socratic dialectic process. *Journal of Computers in Mathematics and Science Teaching*, *18*, 303–315.

Cohen, J. (1988). *Statistical power analysis for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237–248. doi:10.3102/00028312019002237

Comprehensive Meta-Analysis (Version 2.2.064) [Computer software]. Englewood, NJ: Biostat.

Corbett, A. T. (2001a). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User modeling 2001: Proceedings of the eighth international conference, UM 2001* (pp. 137–147). Berlin, Germany: Springer-Verlag.

*Corbett, A. T. (2001b). *Cognitive Tutor results report: 7th grade*. Unpublished manuscript, Carnegie Learning, Pittsburgh, PA.

*Corbett, A. T. (2002). *Cognitive Tutor results report: 8th & 9th grade*. Unpublished manuscript, Carnegie Learning, Pittsburgh, PA.

Corbett, A. T., & Anderson, J. R. (1991, April). *Feedback control and learning to program with the CMU LISP tutor*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Retrieved from http://repository.cmu.edu/psychology/28

*Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko & A. Sears (Eds.), *Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems* (pp. 245–252). New York, NY: ACM.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 849–874). Amsterdam, Netherlands: Elsevier Science.

Craig, S. D., Driscoll, D. M., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia*, *13*, 163–184.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction*, *24*, 565–591. doi:10.1207/s1532690xci2404_4

Crowder, N. A. (1959). Automatic tutoring by means of intrinsic programming. In E. Galanter (Ed.), *Automatic teaching: The state of the art* (pp. 109–116). New York, NY: Wiley.

Fletcher, J. D. (1982). Training technology: An ecological point of view. In R. A. Kasschau, R. Lachman, & K. R. Laughery (Eds.), *Psychology and society: Information technology in the 1980s* (pp. 166–191). New York, NY: Holt, Rinehart & Winston.

Fletcher, J. D. (1985). Intelligent instructional systems in training. In S. A. Andriole (Ed.), *Applications in artificial intelligence* (pp. 427–451). Princeton, NJ: Petrocelli Books.

*Fletcher, J. D. (2011). *DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor assessments* (IDA Document D-4260). Alexandria, VA: Institute for Defense Analysis. Available from DTIC Online database. (ADA542215)

*Fletcher, J. D., & Morrison, J. E. (2012). *DARPA digital Tutor: Assessment data* (IDA Document D-4686). Alexandria, VA: Institute for Defense Analyses.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., . . .Craig, S. D. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instructional Science*, *37*, 487–493. doi:10.1007/s11251-008-9069-2

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

*Gott, S. P., Lesgold, A., & Kane, R. S. (1996). Tutoring for transfer of technical competence. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 33–48). Englewood Cliffs, NJ: Educational Technology.

*Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., . . .Louwerse, M. M. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the twenty-fifth Annual Conference of the Cognitive Science Society* (pp. 474–479). Mahwah, NJ: Erlbaum. Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2003/pdfs/103.pdf

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*, 180–192. doi:10.3758/BF03195563

*Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., Person, N., & The Tutoring Research Group. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies* (pp. 47–54). Amsterdam, Netherlands: IOS Press.

*Grubišic, A., Stankov, S., Rosic, M., & Žitko, B. (2009). Controlled experiment replication in evaluation of e-learning system's educational influence. *Computers & Education*, *53*, 591–602. doi:10.1016/j.compedu.2009.03.014

*Grubišic, A., Stankov, S., & Žitko, B. (2006). An approach to automatic evaluation of educational influence. In S. Impedovo, D. Kalpic, & Z. Stjepanovic (Eds.), *DIWEB 06 Proceedings of the 6th WSEAS International Conference on Distance Learning and Web Engineering* (pp. 20–25). Stevens Point, WI: World Scientific and Engineering Academy and Society. Retrieved from http://bib.irb.hr/datoteka/259289.DIWEB2006_Grubisic_Stankov_Zitko.pdf

Hartley, S. S. (1977). *Meta-analysis of the effects of individually paced instruction in mathematics* (Doctoral dissertation). Available from ProQuest Dissertations & Theses database. (Order No. 7729926)

*Hastings, P., Arnott-Hill, E., & Allbritton, D. (2010). Squeezing out gaming behavior in a dialog-based ITS. In V. Aleven, H. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems 2010* (pp. 204–213). Berlin, Germany: Springer-Verlag.

*Hategekimana, C., Gilbert, S., & Blessing, S. (2008). Effectiveness of using an intelligent tutoring system to train users on off-the-shelf software. In K. McFerrin, R. Weber, R. Carlsen, & D. A. Willis (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2008* (pp. 414–419). Chesapeake, VA: AACE.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

*Jeremic, Z., Jovanovic, J., & Gasevic, D. (2009). Evaluating an intelligent tutoring system for design patterns: The DEPTHS experience. *Educational Technology & Society*, *12*, 111–130. Available from ERIC database. (EJ836295)

*Johnson, S. D., Flesher, J. W., Jehng, J. C. J., & Ferej, A. (1993). Enhancing electrical troubleshooting skills in a computer-coached practice environment. *Interactive Learning Environments*, *3*, 199–214. doi:10.1080/1049482930030303

Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of PMENA XXII (North American Chapter of the International Group for the Psychology of Mathematics Education*; Vol. *1*, pp. 21–49). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education. Available from ERIC database. (SE066887)

*Koedinger, K. R., Aleven, V., Heffernan, N., McLaren, B., & Hockenberry, M. (2004). Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. In J. C. Lester, R. M. Vicario, & F. Paraguacu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 162–173). Berlin, Germany: Springer-Verlag. Retrieved from http://repository.cmu.edu/hcii/158

*Koedinger, K. R., & Anderson, J. R. (1993). Effective use of intelligent software in high school math classrooms. In S. P. Brna, S. Ohlsson, & H. Pain (Eds.), *Proceedings of the World Conference on AI in Education, 1993* (pp. 241–248). Charlottesville, VA: Association for the Advancement of Computing in Education. Retrieved from http://repository.cmu.edu/hcii/4

*Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30–43. Retrieved from http://telearn.archives-ouvertes.fr/hal-00197383/

Kulik, C.-L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, *7*, 75–94. doi:10.1016/0747-5632(91)90030-5

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*, 265–299. doi:10.3102/00346543060002265

Kulik, J. A. (1994). Meta-analytic studies of findings on computer-based instruction. In E. L. Baker & H. F. O'Neil Jr. (Eds.), *Technology assessment in education and training* (pp. 9–33). Hillsdale, NJ: Erlbaum.

Lane, H. C., & VanLehn, K. (2005). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education*, *15*, 183–201. doi:10.1080/08993400500224286

*Le, N. T., Menzel, W., & Pinkwart, N. (2009). Evaluation of a constraint-based homework assistance system for logic programming. In H. Leung, R. Li, R. Lau, & Q. Li (Eds.), *Proceedings of the 6th International Conference on Web-based Learning, Edinburgh, UK, 2007* (pp. 367–379). Berlin, Germany: Springer. Retrieved from http://www.icce2009.ied.edu.hk/pdf/C1/proceedings051-058.pdf

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*, 901–918. doi:10.1037/a0037123

Mathes, P. G., & Fuchs, L. S. (1994). The efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review*, *23*, 59–80. Available from ERIC database. (ED344352)

*Mendicino, M., & Heffernan, N. (2007). *Comparing the learning from intelligent tutoring systems, non-intelligent computer-based versions, and traditional classroom instruction*. Unpublished manuscript, West Virginia University, Morgantown.

*Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education*, *41*, 331–359. Available from ERIC database. (EJ835243)

*Naser, S. (2009). Evaluating the effectiveness of the CPP-Tutor, an intelligent tutoring system for students learning to program in C++. *Journal of Applied Sciences Research*, *5*, 109–114. Retrieved from http://www.aensiweb.com/jasr/jasr/2009/109-114.pdf

*Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). *Effectiveness of Cognitive Tutor Algebra I at Scale* (Working Paper No. WR-984-DEIES). Santa Monica, CA: Rand Corporation. Retrieved from http://www.rand.org/pubs/working_papers/WR984.html

*Pane, J. F., McCaffrey, D. F., Slaughter, M. E., Steele, J. L., & Ikemoto, G. S. (2010). An experiment to evaluate the efficacy of Cognitive Tutor geometry. *Journal of Research on Educational Effectiveness*, *3*, 254–281. doi:10.1080/19345741003681189

*Parvez, S. M., & Blank, G. D. (2007). A pedagogical framework to integrate learning style into intelligent tutoring systems. *Journal of Computing Sciences in Colleges*, *22*, 183–189. Retrieved from http://dl.acm.org/citation.cfm?id=1181849.1181886

*Pek, P.-K., & Poh, K.-L. (2005). Making decisions in an intelligent tutoring system. *International Journal of Information Technology & Decision Making*, *4*, 207–233. doi:10.1142/S0219622005001489

*Person, N. K., Bautista, L., Graesser, A. C., Mathews, E. C., & The Tutoring Research Group. (2001). Evaluating student learning gains in two versions of AutoTutor. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-ED in the wired and wireless future* (pp. 286–293). Amsterdam, Netherlands: IOS Press.

*Phillips, F., & Johnson, B. G. (2011). Online homework versus intelligent tutoring systems: Pedagogical support for transaction analysis and recording. *Issues in Accounting Education*, *26*, 87–97. doi:10.2308/iace.2011.26.1.87

Plano, G. (2004). *The effects of the Cognitive Tutor Algebra on student attitudes and achievement in a 9th-grade algebra course* (Doctoral dissertation). Available from Proquest Dissertations and Theses database. (Order No. 3130130)

Radwan, Z. R. (1997). *Evaluation of the effectiveness of a computer-assisted intelligent tutoring system model developed to improve specific learning skills of special needs students* (Doctoral dissertation). Available from Proquest Dissertations and Theses database. (Order No. 9729551)

*Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics*, *67*, 819–831. doi:10.1119/1.19130

*Reiser, B. J., Anderson, J. R., & Farrell, R. G. (1985). Dynamic student modelling in an intelligent tutor for LISP programming. In A. K. Joshi (Ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 8–13). San Francisco, CA: Morgan Kaufmann. Available from ACM Digital Library database. (1623611)

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research*, *79*, 3–38. doi:10.3102/0034654308325690

*Ritter, S., Kulikowich, J., Lei, P. W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In T. Hirashima, H. U. Hoppe, & S.-C. Young (Eds.), *Supporting learning flow through integrative technologies* (pp. 13–20). Amsterdam, Netherlands: IOS Press.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*, 479–530.

Sarkis, H. (2004). *Cognitive Tutor Algebra 1 program evaluation: Miami-Dade County Public Schools*. Lighthouse Point, FL: The Reliability Group. Retrieved from https://www.carnegielearning.com/research-results/whitepapers-reports/

*Shneyderman, A. (2001). *Evaluation of the Cognitive Tutor Algebra I program*. Unpublished manuscript, Miami–Dade County Public Schools, Office of Evaluation and Research, FL.

Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977. doi:10.1126/science.128.3330.969

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*, 839–911. doi:10.3102/0034654308330968

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. New York, NY: Academic Press.

*Smith, J. E. (2001). *The effect of the Carnegie Algebra Tutor on student achievement and attitude in introductory high school algebra* (Doctoral dissertation). Available from ProQuest Dissertations & Theses database. (Order No. 3065460)

*Stankov, S., Glavinic, V., & Grubišic, A. (2004). What is our effect size: Evaluating the educational influence of a web-based intelligent authoring shell. In S. Nedevschi & I. J. Rudas (Eds.), *Eighth IEEE International Conference on Intelligent Engineering Systems* (pp. 545–550). Cluj-Napoca, Romania: Faculty of Automation and Computer Science, Technical University of Cluj-Napoca.

*Stankov, S., Rosic, M., Žitko, B., & Grubišic, A. (2008). TEx-Sys model for building intelligent tutoring systems. *Computers & Education*, *51*, 1017–1036. doi:10.1016/j.compedu.2007.10.002

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, *105*, 970–987. doi:10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*, 331–347. doi:10.1037/a0034752

*Steuck, K., & Miller, T. M. (1997, March). *Evaluation of an authentic learning environment for teaching scientific inquiry skills*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Available from ERIC database. (ED409217)

Suppes, P., & Morningstar, M. (1969). Computer-assisted instruction. *Science*, *166*, 343–350. doi:10.1126/science.166.3903.343

*Suraweera, P., & Mitrovic, A. (2002). KERMIT: A constraint-based tutor for database modeling. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Lecture notes in Computer Science: Vol. 2363. Intelligent tutoring systems, 6th International Conference, ITS 2002* (pp. 377–387). Berlin, Germany: Springer-Verlag.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning a second-order meta-analysis and validation study. *Review of Educational Research*, *81*, 4–28. doi:10.3102/0034654310393361

*Timms, M. J. (2007). Using item response theory (IRT) to select hints in an ITS. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (pp. 213–221). Amsterdam, Netherlands: IOS Press.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2009, July). *Middle school math intervention report; Cognitive Tutor Algebra I*. Retrieved from http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=87

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, January). *High school mathematics intervention report; Carnegie Learning Curricula and Cognitive Tutor*. Retrieved from http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=88

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*, 227–265. Retrieved from http://iospress.metapress.com/content/AL6R85MM7C6QF7DR

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*, 197–221. doi:10.1080/00461520.2011.611369

*VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62. doi:10.1080/03640210709336984

*VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*, 147–204. Retrieved from http://iospress.metapress.com/content/4QH80UBFDFT0G4YR

Walles, R. L. (2005). *Effects of web-based tutoring software on math test performance: A look at gender, math-fact retrieval ability, spatial ability and type of help* (Unpublished master's thesis). University of Massachusetts, Amherst.

Weerasinghe, A., & Mitrovic, A. (2006). Facilitating deep learning through self-explanation in an open-ended domain. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *10*, 3–19. Retrieved from http://iospress.metapress.com/content/6124YN5QFY99W83N

*Wheeler, J. L., & Regian, J. W. (1999). The use of a cognitive tutoring system in the improvement of the abstract reasoning component of word problem solving. *Computers in Human Behavior*, *15*, 243–254. doi:10.1016/S0747-5632(99)00021-7

Authors

JAMES A. KULIK, PhD, is research scientist emeritus at the Office of Evaluation and Examinations, University of Michigan, 500 S. State Street, Ann Arbor, MI 48109; e-mail: *jimkulik@umich.edu*. His research interests include research synthesis, instructional methods, and teaching evaluation.

J. D. FLETCHER, PhD, is a senior research staff member at the Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311; e-mail: *fletcher@ida.org*. The Institute for Defense Analyses performs studies and analyses on scientific and technical matters for the Office of the Secretary of Defense. Fletcher's work includes assessment of human performance and the value of advanced technologies for education and training, including their monetary and operational return on investment.