IDA

INSTITUTE FOR DEFENSE ANALYSES

Effectiveness of Intelligent Tutoring Systems

James A. Kulik
J.D. Fletcher

June 2012

IDA Document D-4664
Log: H 12-000988

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract W91WAW-11-C-0003, Task AI-2-3370, "Review and Assessment of Intelligent Tutoring Systems," for the Director, Human Performance, Training, and BioSystems Research Directorate, in the Office of the Assistant Secretary of Defense for Research and Engineering and the Office of Naval Research (ONR) Code 341. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Copyright Notice

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

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-4664

Effectiveness of Intelligent Tutoring Systems

James A. Kulik
J.D. Fletcher

Executive Summary

Introduction

This report reviews empirical evidence on the effectiveness of intelligent tutoring systems (ITSs). Although its findings may inform instructional and cognitive theory, its focus is primarily empirical and statistical. The review examined findings from 66 studies of ITSs: 45 system evaluations and 21 component evaluations. These systems employ computer-based tutoring as an instructional method based on (1) models of the subject area representing expert performance in solving problems and/or expert knowledge of underlying concepts, (2) dynamic models that learn from and evolve with the learner's developing skills and knowledge), and (3) application of these models to support and guide mixed-initiative tutorial dialogues derived from expert human tutoring.

Initially, this review identified about 550 reports as candidates for inclusion. These reports were drawn from databases maintained by the Educational Resources Information Clearinghouse (ERIC), the National Technical Information Service (NTIS), Comprehensive Dissertation Abstracts, and Google Scholar.

System evaluations and component evaluations were considered separately. Criteria for system evaluations were that they (1) compare an experimental group using an ITS with a comparison control group receiving conventional classroom instruction, (2) measure learning quantitatively and in the same way for both the experimental and control groups, and (3) contain no substantive methodological flaws. Criteria for component evaluations were that they compare two or more features that could influence the effectiveness of an ITS and that they satisfy system evaluation criteria (2) and (3).

Findings

System Evaluations

Of the candidate evaluations, 45 met criteria for inclusion in the system evaluations. Overall, the effect of intelligent tutoring in these evaluations was to raise student test scores by an average of 0.60 standard deviations over the test scores of conventionally taught students, which is roughly equivalent to an improvement from the 50th to the 73rd percentile.

Although eight system evaluations produced effect sizes of 1.00 standard deviations or higher, six evaluations resulted in effective sizes that were near or below zero. All but one of these latter results were for evaluations that were poorly aligned with the higher order instructional objectives targeted by the tutoring systems. When these studies were eliminated from the analysis, the average effect size was about 0.75 for the remaining 39 systems.

Another factor that influenced study results was the implementation fidelity of the tutoring program—the care and attention with which a program was implemented in the classroom. Programs that paid careful attention to implementation were significantly more effective than those that did not.

Component Evaluations

Of the candidate evaluations, 21 met criteria for inclusion in component evaluations. These studies sought to determine if the presence or absence of a specific tutoring feature affected learning, usually by comparing performance of two versions of a tutoring system, one with and the other without the feature. Findings from these component evaluations can be summarized as the following:

- **Interactive participation as opposed to passive or vicarious learning.** Average effect size for 11 assessments of this component was 0.31 in favor of including student-instructor interactions.
- **Support for self-explanation.** Although effect sizes in six assessments of this component ranged from 0.33 to –0.32, their average was 0.09 in favor of prompts encouraging students to reflect on their solutions, overall a null effect.
- **Flexible exploration.** One study found an effect size of 0.35 in favor of allowing learners flexibility to explore domains rather than limiting them to a fixed sequence of steps.
- **Game playing.** One study found an effect size of 0.28 in favor of an interface with the look and feel of a virtual-reality game over a more standard intelligent tutoring approach.
- **Spoken directions, feedback.** Two studies found effect sizes of 0.81 and 0.54 in favor of a pedagogical agent that appeared on screen and provided spoken directions and feedback compared with directions and feedback provided only as written text.
- **Presence of an animated pedagogical agent.** One study compared the value of narration delivered by an on-screen agent with the same narration delivered off-screen and found no difference between the two treatments.

Conclusion

Overall, this review suggests that ITSs may make substantial improvements over other instructional approaches by accelerating learning, expanding learners' basic problem-solving competencies, and developing the conceptual understanding that contributes to long-term retention and transfer. These computer-based systems provide many of the advantages of tutoring at scales that would be unaffordable if provided by human tutoring or classroom instruction. These systems demonstrated sufficient effectiveness and efficiencies to recommend continued investment in their research, development, and application.

Contents

1. Background

Developers of intelligent tutoring systems (ITSs) have long believed that their programs have strong, beneficial effects on student learning. They have estimated the increase in test performance from intelligent tutoring to be about one standard deviation, roughly an improvement in student test performance from the 50th to the 84th percentile (e.g., Anderson, Boyle, Corbett, & Lewis, 1990; VanLehn, 2011). An increase of this magnitude is much larger than the one-third standard-deviation improvement usually attributed to older forms of computer-based instruction (Fletcher, 2003; C. L. C. Kulik & Kulik, 1991; J. A. Kulik, 1994). It is smaller, however, than the two-standard-deviation effect that has been attributed to human tutoring (Bloom, 1984), which, roughly, would raise an average student's performance from the 50th to the 98th percentile.

Benjamin Bloom (1984) introduced the term two-sigma effect to describe the difference that human tutors seem to make in school programs. (Sigma commonly stands for standard deviation in statistical notation.) Bloom also used the term two-sigma problem to describe the search for other teaching approaches that could improve student performance to the same extent. In the years since Bloom first described the two-sigma problem, it has inspired instructional designers. Noting the impracticality of providing a single human tutor for every student, some researchers have sought to meet Bloom's challenge by capturing the behavior of human tutors in ways made affordable and readily accessible through computer technology (e.g., Corbett, 2001a; Fletcher, 1992; Woolf & McDonald, 1984). The hope was that these ITSs would soon match the success of human tutors and break the two-sigma barrier.

The picture has changed in recent years, however. For one thing, Bloom's findings on human tutors have come into question. Bloom based his conclusions about human tutors on evaluations of a single program by two of his students. Over the years, reviewers have produced at least five extensive meta-analyses of findings on human tutoring, and none of these analyses support Bloom's claims. The earliest meta-analysis (Hartley, 1977) examined findings on peer tutoring in 29 studies of mathematics learning in elementary and secondary schools. Hartley reported that the tutoring programs raised math test scores by an average of 0.60 standard deviations. P. A. Cohen, Kulik, and Kulik (1982) examined results of peer tutoring programs in elementary and secondary schools and reported an average effect size of 0.40 standard deviations in 65 studies. Mathes and Fuchs (1994) found an effect size of 0.36 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 an average effect size of 0.30 for 24 studies. Finally, VanLehn (2011) summarized results from 10 comparisons of performance of tutored and non-tutored students. He reported that the tutored students outperformed the non-tutored students by 0.79 standard deviations. The median effect size in the five meta-analyses is 0.40, far from Bloom's two-sigma effect.

These findings do not suggest that the effects of human tutoring are insignificant. To the contrary, the What Works Clearinghouse (August 2010) considers effect sizes of 0.25 and higher to be large enough to be of substantive importance in education. By this standard, the value of human tutoring programs has been amply demonstrated. However, the effects typically reported for human tutors are nowhere near as large as those found by Bloom's students (Bloom, 1984). If matching the success of typical human tutors remains a goal for developers of ITSs, the goalposts are not as distant as they once seemed.

Recent evaluation efforts have also challenged conventional beliefs about the effectiveness of ITSs. Specifically, several evaluators have suggested that the benefits of ITSs are much smaller than many researchers think. The evaluators have focused their attention on one specific ITS, the Carnegie Learning Corporation's Cognitive Tutor. As the culmination of years of pioneering work by researchers at Carnegie Mellon University, Cognitive Tutor is a computer-based system that teaches students to solve problems by simulating the behavior of expert human tutors. It is now used by 600,000 students in 44 states (Gabriel & Richtel, 2011). No other ITS has received such wide acceptance.

Recent reports, however, have suggested that Cognitive Tutor has little discernible effect on school performance. Campuzano, Dynarski, Agodini, and Rall (2009), for example, reported on results from a $15-million national study of reading and mathematics "software products." Cognitive Tutor Algebra I was 1 of 10 such products evaluated in the 2-year study, and like the 9 other software packages, it turned out to be no more effective than ordinary classroom teaching. None of the 10 software packages improved student test scores over the scores in conventional classes to any practical extent. Test scores from Cognitive Tutor were, in fact, slightly lower than the test scores from students in conventional classes, but the negative effect of the Cognitive Tutor program was too small to be considered statistically significant.

This finding was not the only blow to the reputation of ITSs. A synthesis of evaluation findings on programs to improve middle and high school mathematics examined results of seven evaluations of Cognitive Tutor Algebra I (Slavin, Lake, & Groff, 2009). To be included in the synthesis, the evaluation studies had to meet a set of exacting standards, and only 7 of 13 Cognitive Tutor evaluations examined by Slavin and his colleagues met these standards. Findings suggest that Cognitive Tutor raised student test performance by 0.12 standard deviations—a positive amount but less than the 0.25 standard deviations that the What Works Clearinghouse (August 2010) considers necessary for substantive importance in education.

The What Works Clearinghouse (August 2010) produced another important synthesis of findings on Cognitive Tutor Algebra I, setting up demanding criteria for adequate studies of Cognitive Tutor evaluations. They deemed that only 4 of the 24 high school evaluations they examined were worthy of serious consideration and reported that the average effect of Cognitive Tutor in these 4 studies was very near 0. Students who learned in classrooms that used Cognitive Tutor performed at about the same level as students who were taught in conventional classrooms.

The *New York Times* brought recent findings on Cognitive Tutor to the attention of the public in a front-page story (Gabriel & Richtel, 2011). Under the headline "Inflating the Software Report Card: School-Technology Companies Ignore Some Results," the article took Carnegie Learning to task for emphasizing the positive on its website and ignoring negative findings.

Our review and analysis present a somewhat different picture of the Cognitive Tutor. As will be seen from data in this analysis, the Cognitive Tutor's emphasis on conceptual understanding is poorly aligned with the mass measurement instruments that treat subjects broadly and provide a limited assessment of more substantive, conceptual objectives.

Another recently completely review also presents a different view of tutoring effectiveness. VanLehn (2011) searched journals and conference proceedings for studies that examined the effectiveness of human tutoring, intelligent tutoring, and other tutoring systems. Included in his review were 27 studies that compared intelligent tutoring to instruction that involved no tutoring. The studies examined two approaches to intelligent tutoring, which VanLehn called step-based tutoring and substep-based tutoring (more directive, fine-grained tutoring). VanLehn found an average effect size of 0.76 for step-based tutoring and 0.40 for substep-based tutoring. The overall effect of intelligent tutoring in the 27 studies was to raise test scores by 0.59 standard deviations (roughly an improvement from the 50th to the 72nd percentile). This effect falls short of the one standard deviation gains that many developers expect but is far greater than the zero gains reported elsewhere.

The stakes in the debate about ITSs are high. Educational costs are rising, student performance in many schools is far from satisfactory, and, on international tests of student performance, schools in many countries outperform schools in the United States. Pressures in the military and industrial training communities to raise quality, reduce costs, and compete internationally are also increasing. To improve the situation, educators, trainers, policy makers, and concerned citizens need solid information about the effects of proposed improvement strategies. They need to know which strategies are worth pursuing and which are not. Given these stakes, examining all the evidence on ITSs makes sense. The intent of this review is to provide as comprehensive an analysis of the evidence as possible.

2. Method

This review uses meta-analytic methodology to summarize findings on intelligent tutoring. Glass (1976) first described meta-analytic methods in his seminal presidential address to the American Educational Research Association 35 years ago. Glass's approach recognizes the technical and scientific imperative that any study be accountable and amenable to replication. It has at least four basic features:

- First, meta-analysts use objective methods to locate studies.
- Second, they describe the features of these studies in quantitative and quasi-quantitative terms.
- Third, they report study outcomes as effect sizes, which express the differences between experimental and control groups in standard-deviation units.
- Fourth, they use statistical techniques to examine relationships between study features and study outcomes.

This review focuses on two types of studies of ITSs: system evaluations and component evaluations. System evaluations compare results of ITSs with results of conventional classroom instruction. These evaluations thus measure the instructional effectiveness of ITSs. Component evaluations seek to determine if the presence or absence of a specific tutoring feature affects learning. A typical component evaluation, for example, would compare two different versions of a tutoring system, one with a specific feature and one without. Reviews of intelligent tutoring evaluations do not always distinguish between the two types of evaluations. Because each type of evaluation has its own contribution to make toward the understanding of intelligent tutoring, we analyze results of the two types of evaluations separately in this report.

A. Library Searches

To find studies for possible use in our analyses, we looked in four library databases:

- **ERIC, the digital database of the Educational Resources Information Clearinghouse.** This database covers more than 1 million education-related documents. We narrowed our search to documents tagged with the descriptor "ITS," along with one or more of the following descriptors: "instructional effectiveness," "comparative analyses," or "computer software evaluation." The search yielded 104 reports.

- **NTIS, the bibliographic database of the National Technical Information Service.** This database includes more than 2 million records on government-sponsored research and development efforts. We searched it for records that contained the string “ITSs” in the subject field. The search yielded 120 documents. Reports cleared for open publication and unlimited distribution in the Defense Technical Information Center (DTIC) are supposed to be sent automatically to NTIS. A spot check found that DTIC reports we covered in this review had been sent to NTIS, and we assumed that NTIS would suffice for the two databases.
- **Comprehensive Dissertation Abstracts.** This database holds information on more than 2 million dissertations and theses. We looked for records containing the strings “intelligent tutoring” and some form of the word “evaluate” in the title, abstract, or keywords fields. The search yielded 98 dissertations.
- **Google Scholar.** We restricted our Google search to documents that contained the strings “intelligent tutoring,” “evaluation,” “control group,” and “learning” somewhere in the text of the document. The search produced a list of 1570 reports that met this requirement. Google Scholar sorts reports by relevance, and all reports beyond the first 250 in the list appeared to be irrelevant for our analytic needs. We therefore restricted our examination to the first 250 documents listed by Google Scholar.

We tried using other terms, including the older term “intelligent computer-assisted instruction” (ICAI) in additional searches, but these searches did not increase the pool of studies for the meta-analysis.

We located additional reports by branching from reference lists in the studies we identified. Two reviews were especially helpful. Van Lehn’s (2011) review on tutoring systems cited 38 studies that compared learning gains from human tutoring, 3 forms of computer tutoring, and no-tutoring groups. The Carnegie Learning Corporation’s review of evaluations of cognitive tutors included a reference list of 33 evaluation studies. The bibliographies in other primary studies produced about 2 dozen additional leads to studies that were candidates for inclusion in our review.

As might be expected, the studies identified by our various searches overlapped. Taking this overlap into account, we estimate that we located about 550 separate studies that were candidates for our analyses of system and component evaluations. To be used in the analyses, the candidate studies had to meet several additional criteria.

B. Criteria for Study Inclusion

We considered dissertations, government technical reports, and published conference papers, along with journal articles and book chapters to be fair game for this review.

For a study to be included in our analyses, the primary requirement was a focus on intelligent tutoring. We view ITSs as an instructional method, one that uses computers to simulate the teaching behavior of one instructor interacting with one student. Features that clearly distinguish ITSs from other forms of computer-based instruction are their use of (1) explicit models of the subject area, often with representations of expert performance and knowledge emphasizing underlying concepts, (2) dynamic models of student understanding, which evolve with the learner's developing skills and knowledge, and (3) models of tutorial strategies that support mixed-initiative dialogues between learners and tutors. Such dialogues allow either the learner or the computer-tutor to take the initiative in asking questions and are generated on demand as needed in solving problems, answering questions, and eliciting explanations.

ITSs can be contrasted with drill and practice programs. The latter methods were found to be quite effective in achieving lower level instructional objectives such as learning arithmetic facts (Suppes & Morningstar, 1972), grapheme-phoneme correspondences in beginning reading (Fletcher & Atkinson, 1973), and foreign language vocabulary and phonetics (Van Campen, 1981). Such items and objectives are found in initial learning of nearly all subjects. They generally consist of a collection of discrete items to be memorized and/or applied—not analyzed, evaluated, or synthesized—and are limited to objectives in the lower reaches of Bloom's hierarchy or the lower left-hand corner of Anderson and Krathwohl's (2001) learning space, where memorization of facts and rudimentary concepts, along with the application of simple procedures, is the targeted objectives. Early drill and practice programs employed models of individual learning and instructional prescriptions, some of which were quite sophisticated (e.g., Atkinson & Paulson, 1972; Suppes, Fletcher, & Zanotti, 1976), but they relied on pre-specified models of student states and static instructional methods and did not lend themselves well to learning at higher, more abstract conceptual levels. For material in this area, more flexible, dynamic, and highly adaptive models seem to be needed. ITSs tend to find their value in these areas.

One motivation, therefore, for the development of ITSs grew from the recognition that although computers could be used to teach effectively, pre-specifying all possible states of the learner and programming all possible instructional responses to these states were expensive. Application of dynamic information structures and mixed-initiative dialogue was to enable the computers to generate on demand at least some of the instructional interactions and thereby assume some of the burden and cost of providing adaptive, individualizing instruction (Carbonell, 1970; Fletcher, 2009).

ITSs typically contain four parts:

- An interactive interface for student-computer dialogue;

- A model of the knowledge and skills that form the objectives of the instruction (where we want to go);
- A dynamic model of the individual student's evolving knowledge, skills, and progress toward achieving the objectives of the instruction (where we are now); and
- Tutoring strategies that can be used to bridge the gap between the student's current knowledge and skills and the targeted instructional objectives (how to get from here to there).

We examined all studies to ensure that they contained these features.

ITSs use different techniques to develop and modify models of student knowledge and skills. Two common techniques are model tracing and knowledge tracing, which are described by Anderson et al. (1990), Anderson, Corbett, Koedinger, & Pelletier (1995), and other authors.

Model tracing begins with a model of a problem and with the steps taken (usually by an expert) to solve it and overlays the learner's steps in trying to solve the same or similar problem onto those of the expert. This approach identifies correct steps and missteps while allowing for some variation in step sequence. The tutor can then use the learner's missteps and errors to provide diagnostic feedback and assistance.

Knowledge tracing aims at deeper issues. It also begins with a model, often a concept map or diagram that breaks down subject matter knowledge into its many interrelated conceptual components (e.g., Hoffman, Shadbolt, Burton, & Klein, 1995; Novak & Cañas, 2008). This model can then be used to identify the basic knowledge and cognitive skills needed to solve a problem. Again, using an overlay technique, the computer tutor determines what concepts a learner applied or failed to apply in trying to solve a problem. In this way, it infers from the learner's actions and the concept map—or some other knowledge representation—what the learner understands or misunderstands. Evaluations of both model-tracing and knowledge-tracing systems are included in our analyses.

C. Criteria for System Evaluations

To be included in our analysis of system evaluations, a study had to meet the following additional criteria:

- The study had to compare the treatment group to a comparison group. Single-group pre-post comparisons do not provide an adequate basis for measuring treatment effectiveness and were not included in this review. Also ineligible were comparisons of treatment group results to norms or expected performance.

- The comparison group had to receive instruction that was representative of the instruction commonly provided in classrooms (e.g., lecture, recitation, homework, and perhaps laboratory exercises). The comparison group could be either a conventional class or an especially constituted group that received instruction that closely approximated conventional teaching. Studies in which comparison groups used materials that were specially developed for the treatment groups (e.g., "script" control groups) were not accepted. Also not accepted were studies in which comparison groups received no relevant instruction (i.e., no-instruction control groups).
- Achievement results had to be measured quantitatively and in the same way in the treatment group and the control group. School grades were not acceptable outcome measures, because grades could have been awarded on a different basis in different classes. Results from both locally developed posttests and tests developed for wider use (i.e., district, state, and national assessments and published tests) are included in the analyses.
- The study could not contain disqualifying methodological flaws such as significant pre-treatment differences between the treatment and control groups. Pretest differences of 0.50 standard deviations or more are considered too large for adjustment by regression techniques and are thus disqualifying (Slavin et al., 2009). Overalignment of the outcome measures with treatment or control treatments was also cause for exclusion. Overalignment occurs, for example, when the outcome measure uses test items that were specifically included in the instructional materials for either the treatment or control group. Also disqualifying was the use of groups drawn from different populations (e.g., volunteers in the treatment group and non-volunteers in the control group).

Forty-five of the approximately 550 studies that we located through database searching and branching met all the criteria for inclusion in the meta-analysis of system evaluations.

D. Criteria for Component Evaluations

To be included in our analysis of component evaluations, a study had to compare versions of a tutoring system with and without a specific feature of intelligent systems. For example, a component evaluation might examine the value of spoken feedback by comparing a tutoring system that provides spoken feedback with a version that provides only written feedback. When searching for system evaluations of tutoring effectiveness, we located 21 component evaluations, described in 11 separate reports, that other researchers (e.g., VanLehn, 2011) had included in their reviews on intelligent tutoring. This set of studies formed the data set for our analysis of component evaluations.

E. Study Features

We used 13 variables to describe study settings, treatments, validity threats, and outcome measures of the studies that we located (see Table 1). Definitions of these variables were guided by our knowledge of previous meta-analyses of instructional technology findings and by a preliminary examination of the studies located through library searches. Some variables originally recorded as continuous variables (e.g., study year, study size, study length) were split into ordered categories for the final analysis. This categorization helped solve the problems presented by skew, non-normality, and presence of outliers in the continuous measurements.

Table 1. Categories for Describing Study Settings, Treatments, Validity Threats, and Outcome Measures

Settings

- Country (1 = USA; 2 = other)
- Publication year (1 = Up to 2000; 2 = 2001–2005; 3 = 2006+)
- Grade level (1 = K–12; 2 = postsecondary)
- Subject (1 = math; 2 = other)

Treatments

- Study type
  - 1 = Experimental: short-term studies in which treatment and control groups work, usually in a computer laboratory, on the same assignments with or without intelligent tutor.
  - 2 = Field evaluations: studies that compare performance in conventionally taught and intelligent tutoring classes.
- Study size (1 = Up to 80; 2 = 81–250; 3 = 251+)
- Study duration (1 = Up to 4; 2 = 5–16; 3 = 17+ weeks)
- Cognitive Tutor study
  - 1 = No: Study does not evaluate a current or earlier version of a Carnegie Learning Cognitive Tutor program.
  - 2 = Yes: Study examines such software.

Validity Threats

- 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.
- Pre-treatment differences
  - 1 = Unadjusted posttest: posttest means not adjusted for pretest differences.
  - 2 = Adjusted posttest: gain scores or posttest means adjusted by covariance or regression.

Outcome Measures

- Test source
  - 1 = Local: Posttest was a locally developed test.
  - 2 = Regional: Posttest was a commercial, state, or district test.
- Test format
  - 1 = Constructed-response items only: Posttest was a problem-solving test, essay exam, and so forth.
  - 2 = Both constructed-response and objective-test items: Posttest measures included both item formats.
  - 3 = Objective items: Posttest was a multiple-choice test or other test with a fixed-alternative format.

F. Effect Sizes

Effect sizes in comparisons of average performance are often calculated as the difference between treatment and control posttest means, divided by the posttest standard deviation of the control group (Glass, McGaw, & Smith, 1981), which is the approach used in this review. Some reports included this kind of effect size, along with other statistical data. When these reports did not, we calculated this kind of effect size, whenever possible, from means and standard deviations provided in the reports. Occasionally, it was necessary to retrieve effect sizes from reported test statistics rather than from means and standard deviations. We used standard formulas and techniques (e.g., Cooper, Hedges, & Valentine, 2009; Glass et al., 1981) to retrieve effect sizes from *t*-statistics and *F*-statistics when means and variances were not reported.

A few guidelines are useful when calculating effect sizes from means and standard deviations. First, posttest means adjusted by pre-instruction measures for prior knowledge usually provide a better estimate of population treatment effects than unadjusted posttest means do. Pre-post gain scores are a good example of adjusted means, but covariate-adjusted means and regression estimates of treatment effects are even better estimators. In calculating or recalculating effect sizes, we therefore established explicit priorities. We gave highest priority to covariance-adjusted means and regression estimates of treatment effects, next highest priority to gain scores, and lowest priority to simple posttest means.

Although using adjusted rather than raw means is preferable when estimating treatment effects, using the standard deviations of the adjusted measures to standardize treatment effects (i.e., as the denominator in effect size calculations) is not appropriate in meta-analysis. We used pre-adjusted, raw standard deviations instead. Standard deviations of gain scores or covariance-adjusted scores are smaller than raw standard deviations, and effect sizes based on these reduced standard deviations are inflated. These inflated effect sizes cannot be interpreted simply (e.g., in terms of percentile scores or standard scores) and cannot be aggregated with effect sizes that are calculated with raw standard deviations.

Although effect sizes based on reduced standard deviations are not appropriate for meta-analyses, such effect sizes sometimes appear in the educational literature. When we found reports that had these inflated effect sizes, we made corrections whenever possible by recalculating the effect sizes from other statistics supplied in the reports. When a report did not present additional statistics from which effect sizes could be recalculated, we assumed a correlation of 0.60 between pretest and posttest scores and adjusted the effect sizes accordingly. This default correlation was the median value in five studies that reported either pre-post correlations or statistics from which the correlations could be derived (Arnott, Hastings, & Allbritton, 2008; Fletcher, 2011; Pek & Poh, 2005; Suraweera & Mitrovic, 2002; VanLehn et al., 2007).

Finally, a common but not universal practice is to use control-group rather than pooled standard deviations for standardizing treatment effects for meta-analysis. Control-group standard deviations are not affected by experimental treatments, whereas treatment group standard deviations may be. As a result, control-group standard deviations usually provide a better estimate of variation in the general population. Some researchers, however, report only pooled standard deviations for their measurements. When researchers failed to report separate posttest standard deviations for treatment and control groups, we used pooled standard deviations instead of control-group standard deviations in effect-size calculations. A meta-analysis of interactive, computer-based videodisc instruction by Fletcher (1989) reported 151 effect sizes calculated using control group and pooled standard deviations. Although some sizable differences emerged, overall, the average difference in effect size using the two approaches was 0.007. In 91 cases, the effect sizes based on pooled standard deviations were larger. In the remaining 60 cases, the effect sizes based on control-group standard deviations were larger.

3. Results

This section presents results from two sets of analyses. The first set of results comes from analyses of the 45 system evaluations. The second set comes from analyses of the 21 component evaluations.

A. System Evaluations

The 45 system evaluations that satisfied our inclusion criteria constitute a diverse group (see Table 2). The publication dates of the studies span nearly three decades. The earliest study dates from 1985, and the most recent study dates from 2011. The studies come from six countries: the United States, Germany, Singapore, New Zealand, Croatia, and Serbia. The content covered in the studies varies from "borrowing" in third-grade subtraction problems to solving analytic problems of the sort that appear on the Law School Admissions Test (LSAT). The studies took place in elementary schools, high schools, colleges, and military training sites. The shortest study lasted less than 1 hour, and the longest study lasted three semesters.

1. Overall Effects

In 41 of the 45 studies (91%), the students who received intelligent tutoring outperformed the control students on posttests. In the remaining four studies (9%), the conventionally taught students had higher averages. Although these box-score results look good as a won-loss record, they provide little information about the strength and consistency of the intelligent tutoring effects. Effect-size analysis provides a more complete picture.

The strongest positive effect of tutoring in the 45 studies was to raise posttest scores by 1.97 standard deviations (Fletcher, 2011). The largest negative effect was to reduce scores by 0.34 standard deviations (Hategekimana, 2008). The median effect size in the 45 studies is 0.63. The average effect size for the 45 studies is 0.60. The effect sizes for the Fletcher and Hategekimana studies are outlier values, where an outlier is defined as a value that is at least 1.5 interquartile ranges above the 75th percentile or a value that is at least 1.5 interquartile ranges below the 25th percentile. The 5% trimmed mean, which is calculated from all values in this data set except the highest and lowest, is 0.60.

Cohen (1988) defined rough guidelines for the interpretation of effect sizes, calling effect sizes of 0.20 small, 0.50 medium-size, and 0.8 large. By these standards, the

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Anderson, Boyle, Corbett, & Lewis (1990); also Anderson, Corbett, Koedinger, & Pelletier (1995)	Geometry course at a high school in Pittsburgh, Pennsylvania, 1986–1987	1 quarter	Geometry Tutor	Local	5 classes	1.00
Arbuckle (2005)	Algebra I course in Edmond Public Schools, Edmond, Oklahoma, Grades 9–11, 2003–2004	10 weeks	Cognitive Tutor	Local	2 classes (1T, 1C); 111 students (83T, 28C)	0.74
Arnott, Hastings, & Allbritton (2008)	Research methods in psychology course at DePaul University, Chicago, Illinois, Winter 2007	5 weeks	Research Methods Tutor (RMT)	Local	5 classes (3T, 2C); 125 students (73T, 52C)	0.60
Atkinson (2007)	Training in reading comprehension in three high schools and a technical center in Phoenix, Arizona (metro area), 2004–2006	8–12 weeks, approximately 36 hours	Gradations, STAR, and Read On!	Regional	7 groups (6T, 1C); 159 students (139T, 20C)	0.25
Burns (1993)	Arithmetic exercises in a public school in Westchester County, New York, Grade 3	6 weeks, one 20-minute session per week	MEADOW	Local	4 classes; 56 student volunteers (19T, 37C)	0.75

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Study	Interactive Condition	Non-Interactive Condition	Effect Size	Study	Interactive Condition	Non-Interactive Condition
Cabalo & Vu (2007)	Algebra I course at five high schools and one community college in Maui, Hawaii, 2005–2006	6 months	Cognitive Tutor Algebra I	Regional	22 classes (11T, 11C); 345 students (182T, 163C)	0.03
Campuzano, Dynarski, Agodini, & Rall (2009)	Algebra I course in 11 schools in 4 districts, Grades 8 and 9, 2004–2006	1 school year, 24 weeks, approximately 36 hours	Cognitive Tutor Algebra I	Regional	29 teachers (15T, 14C); 775 students (440T, 315C)	–0.10
Carlson (1996)	Instruction in writing at two high schools in San Antonio, Texas, second semester, 1993	1 semester, 9 sessions, 8 hours total	Fundamental Skills Training Project's R-WISE 1.0 (Reading and Writing in a Supportive Environment)	Local	48 classes (26T, 22C); 852 students (429T, 423C)	0.78
Corbett & Anderson (2001); also Corbett (2001a)	College course in LISP computer programming	5 lessons, average of 7 sessions and 12 total hours	Standard ACT Programming Tutor (APT)	Local	20 paid student volunteers (10T, 10C)	1.00
Corbett (2001b)	Pre-algebra course at North Hills Junior High School, Grade 7 academic classes, Pittsburgh, Pennsylvania, 2000–2001	1 school year	Cognitive Tutor Pre-Algebra	Both local and regional	9 classes (1T, 8C); 175 students	0.46 (0.74 local test, 0.19 regional test)

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Corbett (2002)	Pre-algebra course at Chartiers Valley Middle School, Grades 8 and 9 academic classes, Pittsburgh Pennsylvania, 2001–2002	1 school year	Cognitive Tutor Pre-Algebra	Both local and regional	5 classes (1T, 4C); 173 students	0.21 (0.29 local test, 0.14 regional test)
Fletcher (2011)	Information technology systems at U.S. Navy Center for Information Dominance (CID), Corry Station, Pensacola, Florida, Fall 2010	8 weeks	Digital Tutor vs. Information Technology of the Future class	Local	40 students (20T, 20C)	1.97
Gott, Lesgold, & Kane (1996)	Electronic maintenance at three U.S. Air Force Bases (AFBs) (Langley AFB, Virginia; Nellis AFB, Nevada; Eglin AFB, Florida)	–	Sherlock 2	Local	41 students (18T, 23C)	0.85
Graesser, Jackson, et al. (2003); reanalyzed in Graesser et al. (2004)	College physics course at Ole Miss, Rhodes College, and University of Memphis	1 week, 2 sessions, 2–3 hours each	Why/AutoTutor	Local	29 students (21T, 8C)	0.78

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Graesser, Moreno, et al. (2003); also Graesser et al. (2004)	College physics course in computer literacy	1 session	AutoTutor	Local	81 students	0.17 (0.34 deep learning, 0.00 shallow learning)
Grubišić, Stankov, & Žitko (2006)	Introduction to computer science at University of Split, Split, Croatia, 2005–2006	14 weeks	eXtended Tutor-Expert System (xTex-Sys)	Local	80 students (40T, 40C)	0.79
Grubišić, Stankov, Rosić, & Žitko (2009)	Introduction to computer science at University of Split, Split, Croatia, 2006–2007	14 weeks	eXtended Tutor-Expert System (xTex-Sys)	Local	39 students (20T, 19C)	1.23
Hastings, Arnott-Hill, & Allbritton (2010)	Research methods in psychology course at Chicago State University, Chicago, Illinois	5 weeks, 2–4 hours total	Research Methods Tutor (RMT)	Local	2 classes (1T, 1C); 87 students (56T, 31C)	1.21
Hategekimana (2008)	Picture-editing software lessons at Iowa State University, Ames, Iowa, Fall 2007	1 week, 2 sessions, total of 90 minutes	Locally developed ITS	Local	50 paid student volunteers (26T, 24C)	–0.34
Jeremic, Jovanovic, & Gasevic (2009)	Upper division software course at Military Academy, Belgrade, Serbia, Spring 2006	5 months	Design Patterns Teaching Helping System (DEPTHS)	Local	3 classes (1T, 2C); 42 students (14T, 28C)	0.62

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Johnson, Flesher, Jehng, & Ferej (1993)	Electrical troubleshooting course at Institute of Aviation, University of Illinois at Urbana Champaign, 1990–1991	12 weeks, average of 5.25 total hours	Technical Troubleshooting Tutor	Local	2 classes (1T, 1 C); 34 students (18T, 16C)	0.80
Koedinger & Anderson (1993)	Geometry theorem-proving lessons at high school in Pittsburgh, Pennsylvania, 1992	4–5 weeks, 25 class periods, 44 minutes each	ANGLE with experienced teacher (A) and inexperienced teachers (B&C)	Local	8 classes (4T, 4C); 62 students	0.35 (0.96 stronger implementation, –0.23 weaker implementation)
Koedinger, Aleven, Heffernan, McLaren, & Hockenberry (2004)	Solving LSAT analytic problems at a college in the northeastern United States	1 session, 1 hour	LSAT Analytic Logic Tutor	Local	30 students (15T, 15C)	0.78
Koedinger, Anderson, Hadley, & Mark (1997)	Algebra course at three high schools in Pittsburgh, Pennsylvania, Grade 9, 1993–1994	1 year	Practical Algebra Tutor (PAT) and Pittsburgh Urban Math Project (PUMP)	Both local and regional	25 classes (20T, 5C); 590 students (470T, 120C)	0.68 (0.99 local tests, 0.36 regional tests)
Le, Menzel, & Pinkwart (2009)	Computer programming at the University of Hamburg, Germany, Department of Informatics	1 session, 1 hour	INCOM	Local	35 students (18T, 17C)	0.31 (0.28 stronger implementation, 0.04 weaker implementation)
Mendicino & Heffernan (2007)	Algebra lesson at a high school in rural area, 2004–2006	1 session, 30–45 minutes	ITS	Local	121 students	0.63

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Mendicino, Razzaq, & Heffernan (2009)	Mathematics lesson at elementary school in rural area, Grade 5	1 session	ITS	Local	28 students	0.55
Naser (2009)	C programming at Al-Azhar University of Gaza, Palestine, Faculty of Engineering and Information Technology, freshman class	1 month	C Intelligent Tutoring System (CPP-Tutor)	Local	62 students (31T, 31C)	0.77
Pane (2010)	Geometry course at eight high schools in Baltimore County, Maryland, 2005–2008	1 semester	Cognitive Tutor Geometry	Regional	38 classes (19T, 19C); 699 students (348T, 351C)	–0.19 (0.04 stronger implementation, –0.42 weaker implementation)
Parvez & Blank (2007)	Object-oriented programming at Lehigh University, Bethlehem, Pennsylvania, summer program for high school students, spring and summer 2007	–	DesignFirstITS	Local	32 students (16T; 16C)	0.9
Pek & Poh (2005)	Engineering mechanics lesson at School of Mechanical and Manufacturing Engineering, Singapore Polytechnic	1 session, about 80 minutes	iTutor	Local	33 students (16T, 17C)	1.17

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Person, Bautista, Graesser, Mathews, & The Tutoring Research Group (2001); also Graesser et al. (2004)	Computer literacy lesson at the University of Memphis	1 session, 45–55 minutes	AutoTutor	Local	60 students	0.16 (0.30 deep learning, 0.03 shallow learning)
Reif & Scott (1999)	Introductory physics lessons at Carnegie Mellon University, Fall 1996	1 week, 5 sessions, 7.5 hours	Personal Assistant for Learning (PAL)	Local	30 students (14T, 16C)	0.78
Reiser, Anderson, & Farrell (1985); Also Anderson, Boyle, Corbett, & Lewis (1990)	LISP computer programming course at Carnegie Mellon University, 1984	6 weeks	GREATERP LISP Tutor (Goal-Restricted Environment for Tutoring and Educational Research on Programming)	Local	1 class; 40 students (20T, 20C)	1.00
Ritter, Kulikowich, Lei, McGuire, & Morgan (2007)	Algebra I at three junior high schools in Moore, Oklahoma, 2000–2001	1 year	Cognitive Tutor Algebra I	Regional	257 students (153T, 102C)	0.40
Shneyderman (2001)	Algebra I at six high schools in Miami, Florida, 2000–2001	1 year	Cognitive Tutor Algebra I	Regional	24 classes (12T, 12C); 777 students (325T, 452C)	0.12
Smith (2001)	Algebra I at six suburban high schools in Virginia City Beach, Virginia, 1999–2000	3 semesters	Carnegie Algebra Tutor	Regional	12 classes (6T, 6C); 445 students (229T, 216C)	–0.07

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
Stankov, Glavinic, & Grubišić (2004)	Introductory computer science course at a university in Croatia, 2004–2005	1 semester, 15 weeks, 2 hours weekly	Distributed Tutor Expert System (DTex-Sys)	Local	22 students (11T, 11C)	1.16
Stankov, Rosić, Žitko, & Grubišić (2008)	Courses in several science areas at University of Split (one study) and primary schools (eight studies) in Split, Croatia, Grade 2 to first-year college, 2005–2007	5 to 14 weeks	eXtended Tutor-Expert System (xTex-Sys)	Local	380 students (190T, 190C)	0.74
Steuck & Miller (1997)	Scientific inquiry skills in ecology and biology at 15 junior and senior high schools in 5 states, Grades 7, 9, and 10, 1995–1996	36 weeks, 18 sessions, 18 hours total tutoring time	Fundamental Skills Training Project's Instruction in Scientific Inquiry Skills (ISIS)	Local	1,553 students (765T, 788C)	0.37
Suraweera & Mitrovic (2002)	Database design lesson at University of Canterbury, Christchurch, New Zealand, August 2001	1 session, about 1 hour	KERMIT	Local	62 student volunteers	0.56
Timms (2007)	Lessons on force, motion, and speed at middle school science classes	Several days	Full Option Science System (FOSS) tutor	Local	131 students (101T, 31C)	0.63

Table 2. Descriptive Information and Effect Sizes for 45 Studies of ITSs (Continued)

Publication	Subject and Setting	Treatment Duration	Treatment	Posttest	Sample Size[a]	Effect Size
VanLehn et al. (2005)	Introductory physics course at U.S. Naval Academy, 2000–2003	1 semester	Andes	Local	396 students (282T, 114C)	0.44 (0.95 deep learning, –0.08 shallow learning; 0.57 stronger implementation, 0.21 weaker implementation)
VanLehn et al. (2007), Experiment 2	Lessons on physics principles and misconceptions at Universities of Memphis, Mississippi, and, and Pittsburgh and Rhodes College	2 sessions, average of 126 total minutes	Why2-AutoTutor and Why2-Atlas	Local	48 students (32T, 16C)	0.70
Wheeler & Regian (1999)	Course on solving word problems at seven high schools in Texas, New Mexico, and Ohio, Grade 9, 1992–1993	1 school year, one 50-minute session per week	Fundamental Skills Training Project's Word Problem Solving (WPS) tutor	Local	40 classes (32T, 8C); 493 students (409T, 84C)	0.34

[a] T = treatment group; C = comparison group

average effect size of intelligent tutoring is moderate to large in size. Effects in 10 (27%) of the studies are positive but small in size, effects in 17 (38%) of the studies are positive and moderate in size, and effects in 11 (24%) of the studies are positive and large in size. Effects in 6 of the remaining studies (13%) are trivial in size. One study (Hategekimana, 2008) reported an effect that was negative and small in size.

More recently, the What Works Clearinghouse (August 2010) concluded that effect sizes of 0.25 and higher are large enough to be of substantive importance for instruction. By this standard, in 36 of the 45 studies (80%), tutoring gains were large enough to be of substantive importance. Only one study (Hategekimana, 2008) had a negative effect that was large enough to be of substantive importance.

2. Study Features and Effect Sizes

Although intelligent tutoring, on average, improved learning by a moderate-to-large amount, effects were very large in some studies and near zero in others. To determine whether the variation in study results was related to the ways in which the studies were carried out, we calculated average effect sizes for different categories of studies (see Table 3), and we also calculated correlations between study features and effect sizes (see Table 4). We carried out these analyses with the full sample of 45 studies and the trimmed sample of 43 studies (i.e., all studies but the one with the largest positive effect size and the one with the largest negative effect size). Results were similar in the two analyses, but, for the sake of simplicity, we present only the results for the trimmed sample in the tables.

Table 3. Average Effect Sizes by Study Features

Study Feature	Number	Mean	Standard Deviation
Country			
USA	34	0.54	0.35
Other	9	0.82	0.32
Publication year			
Up to 2000	11	0.70	0.24
2001 through 2005	14	0.56	0.39
After 2006	18	0.56	0.39
Grade level			
Elementary and high school	21	0.43	0.33
Postsecondary	22	0.76	0.31

Table 3. Average Effect Sizes by Study Features (Continued)

Study Feature	Number	Mean	Standard Deviation
Subject			
Mathematics	16	0.38	0.34
Computer science	12	0.72	0.37
Science	7	0.70	0.26
Other	8	0.75	0.27
Study type			
Experimental study	14	0.65	0.29
Field evaluation	29	0.57	0.39
Study size			
Up to 80 participants	21	0.76	0.28
81 through 250 participants	9	0.54	0.32
More than 250 participants	12	0.30	0.33
Study duration			
Up to 4 weeks	12	0.61	0.29
5 through 16 weeks	15	0.75	0.38
17 weeks or more	11	0.29	0.26
Cognitive Tutor study			
No	29	0.68	0.29
Yes	14	0.41	0.42
Control for group assignment			
Quasi-experimental design	12	0.50	0.40
Random control trial	28	0.65	0.35
Control for instructor effects			
Different instructors	13	0.48	0.37
Same instructor	26	0.64	0.37
Control for pre-treatment differences			
Unadjusted posttest	10	0.66	0.31
Adjusted posttest	32	0.58	0.38
Test source			
Local	33	0.72	0.28
Local and regional	3	0.45	0.24
Regional	7	0.08	0.22
Test format			
Constructed-response items only	12	0.74	0.23
Constructed-response and objective	12	0.54	0.36
Objective items only	15	0.53	0.47

Table 4. Correlations Between Study Features and Effect Sizes in 43 System Evaluations

	Correlation	
Study Feature	***r***	**Sig**
Country (1 = USA; 2 = other)	.32	.035
Publication year (1 = up to 2000; 2 = 2001–2005; 3 = 2006+)	–.17	.269
Grade level (1 = K–12; 2 = college and postsecondary)	.47	.002
Subject (1 = math; 2 = other)	.47	.001
Study type (1 = experimental study; 2 = field evaluation)	–.11	.490
Study size (1 = up to 80; 2 = 81–250; 3 = 251+)	–.56	.000
Study duration (1 = up to 4; 2 = 5–16; 3 = 17+ weeks)	–.33	.041
Cognitive Tutor study (1 = no; 1 = yes)	–.36	.017
Group assignment (1 = intact groups; 2 = random)	–.19	.241
Instructor effects (1 = different; 2 = same instructors)	.20	.212
Pre-treatment differences (1 = unadjusted; 2 = adjusted)	–.09	.560
Test source (1 = local; 3 = regional; 2 = both)	–.67	.000
Test format (1 = constructed response; 3 = objective; 2 = both)	–.22	.168

The analyses showed that six study features are significantly related to effect size: the country in which the study was conducted, the grade level of the participants, the subject that was taught, the study sample size, the test source, and the intelligent tutoring program used. Specifically, effect sizes are smaller in studies (1) from the United States, (2) with younger participants, (3) with larger sample sizes, (4) with math as the subject matter, (5) with outcomes measured on regional tests, and (6) with Cognitive Tutor as the ITS.

These six features are highly intercorrelated. Their influences are not independent. For example, test source (local vs. regional) is the study feature most strongly related to effect size, and, when its influence is controlled statistically by partial correlation or regression techniques, none of the remaining study features are significantly correlated with effect size. This result suggests that the zero-order correlations between study features and effect sizes should not be taken at face value. One or more underlying influences may be behind all of the significant zero-order correlations.

3. Test Alignment

We carefully examined the differences in effect sizes between studies and within studies to identify the fundamental influences. This closer analysis began with the observation that almost all the studies with trivial or very small effect sizes used standardized or regional posttests that were poorly aligned with the higher order instructional objectives emphasized in the ITSs. Many of the studies with poorly aligned posttest measures were Cognitive Tutor evaluations, but we also found instances of poor alignment with other tutoring systems.

a. Cognitive Tutor Studies

In a pioneering evaluation of the Practical Algebra Tutor (PAT), an early version of a Cognitive Tutor program, Koedinger, Anderson, Hadley, and Mark (1997) noted that this ITS was developed to support a new math curriculum and that standardized tests available at the time were poorly aligned with the objectives of the new curriculum. Specifically, PAT taught such problem-solving skills as analysis of complex problem situations, selection of solution methods, and application of these methods to find answers. However, standardized tests available at the time were not designed to measure such higher order curricular objectives. Instead, these multiple-choice tests measured recognition skills taught in standard curricula.

Koedinger and his colleagues therefore included two types of tests in their evaluation of PAT: locally developed tests and standardized tests. The locally developed tests measured problem-solving skill by requiring students to construct answers for the test problems. The standardized tests were multiple-choice measures of recognition skills. The researchers found large effects on the locally developed problem-solving tests (average effect size = 0.99) and small effects on the regional multiple-choice tests (average effect size = 0.36). Koedinger and his colleagues were encouraged by these results. The problem-solving tests showed that this version of Cognitive Tutor was very effective in teaching the higher order skills that it was designed to teach. Results on standardized tests showed that the problem-solving benefits came without negative effects on poorly aligned tests that did not stress problem solving.

Other studies of Cognitive Tutor found the same pattern of results. For example, Corbett (2001b, 2002) examined the effects of Cognitive Tutor on locally developed problem-solving tests and multiple-choice tests consisting of released questions on regional—international, national, and state—tests. For Grade 7 students, the effects for locally developed problem-solving tests were large (average effect size = 0.71), and the effects for regional, multiple-choice tests were trivial (average effect size = 0.18). For Grade 8 students, effects on problem-solving tests were small (average effect size = 0.28), and the effects for regional tests were trivial (average effect size = 0.13).

The pattern holds up in the full set of 14 studies of Cognitive Tutor (see Table 5). Overall, Cognitive Tutor significantly and substantially raised student performance on locally developed tests but neither helped nor hindered student performance on regional tests. The average effect size on locally developed tests is 0.72, whereas the average effect size on the regional tests in the 14 Cognitive Tutor evaluations is 0.10. That is, Cognitive Tutor boosted performance on tests that were well aligned with curricular objectives but did not lower performance on tests that were less clearly aligned.

Table 5. Effect Sizes by Test Source for 14 Studies of Algebra, Geometry, and LISP Cognitive Tutors

	Effect Size		
Publication	Overall	Local	Regional
Anderson et al. (1990)	1.00	1.00	–
Arbuckle (2005)	0.74	0.74	–
Cabalo & Vu (2007)	0.03	–	0.03
Campuzano et al. (2009)	–0.10	–	–0.10
Corbett (2001b)	0.45	0.71	0.18
Corbett (2002)	0.21	0.28	0.13
Corbett & Anderson (2001)	1.00	1.00	–
Koedinger & Anderson (1993)	0.35	0.35	–
Koedinger et al. (1997)	0.68	0.99	0.36
Pane et al. (2010)	–0.19	–	–0.19
Reiser et al. (1985)	1.00	1.00	–
Ritter et al. (2007)	0.40	–	0.40
Shneyderman (2001)	0.12	–	0.12
Smith (2001)	–0.07	–	–0.07
Average	0.36	0.72	0.10

b. Alignment Effects

Additional evidence for the importance of the alignment of instructional and test objectives comes from two studies of the AutoTutor system (Graesser, Jackson, et al., 2003; Person, et al., 2001) and one study of the Andes tutoring system (VanLehn, et al., 2005). Like the evaluators of Cognitive Tutor, Graesser and VanLehn and their colleagues found that the size of tutoring effects depended on the degree of test alignment with the higher order objectives of their programs. Specifically, effects were large on tests of conceptual or deep understanding but small on tests of factual information or more shallow learning, as shown in Table 6.

Table 6. Effect Sizes for Three Studies With Separate Measures of Deep and Shallow Learning

Publication	Deep Learning	Shallow Learning
Graesser, Moreno, et al. (2003)	0.34	0.00
Person et al. (2001)	0.30	0.03
VanLehn et al. (2005)	0.95	–0.08
Average	0.62	–0.02

For the earliest of the studies, Person et al. (2001) had hoped to use a sample of questions included with the course textbook as a posttest in their evaluation, but the researchers found that all the textbook questions were at the information level in Bloom's hierarchy. Then, they asked experts at Carnegie Mellon University to write conceptual test items based on the textbook material, and they used these questions as the measure of deep learning in their evaluation. Person et al. (2001) found an effect size of 0.30 for conceptual or deep learning and an effect size 0.03 for informational or shallow learning. A subsequent evaluation by Graesser and his colleagues (2003) examined the effects of AutoTutor on shallow and deep learning in a computer literacy course. They found the same pattern of results in this study. AutoTutor raised scores by an average of 0.34 on the conceptual test and an average of zero on the information items.

An evaluation by VanLehn et al. (2005) further confirmed this pattern of findings. The researchers looked at the effects of Andes on several different tests in a physics course. They reported that effect sizes were high on the tests that measured conceptual learning but low on the tests that measured informational learning. The average effect size for the measures of deep learning and for measures of shallow learning was –0.08.

c. Overall Importance of Test Alignment

The inclusion of poorly aligned outcome measures in these evaluations affected the overall meta-analytic results. Results from these tests depressed the overall average effect size for ITSs, inflated the variability in study findings, and created the illusion that many different study features influenced the findings. When we eliminated poorly aligned tests from our analysis, the pool of studies became smaller (i.e., 39 studies rather than 45), but the average effect size went up from 0.60 to 0.73, the consistency of results increased, and the results seemed more robust. Average effect size in the reduced sample of 39 studies is 0.73, the median effect size is 0.75, and the 5% trimmed mean is 0.72. The interquartile range is 0.40. Approximately half the effect sizes fall between 0.55 and 0.95. In addition, no study feature is related significantly to effect size in the reduced data set. The effects of intelligent tutoring on aligned tests therefore seem to be robust and not susceptible to slight changes in subjects, tutoring features, or design features.

Also notable is that all but one of the studies in the reduced data set found an effect size of 0.25 or more. This cutoff point is the one used by the What Works Clearinghouse (August 2010) to separate results of no practical significance from educational results that are important. We therefore conclude that with well-aligned outcome measures, the effects of intelligent tutoring were large enough to be considered substantive in 38 (97%) of the 39 studies.

d. Implementation Fidelity

Along with test alignment, implementation fidelity appears to have a substantial effect on study findings. Implementation fidelity refers to the degree to which the implementation of a tutoring system meets developer specifications. Fidelity is high when developer specifications are met and low when they are not. Fidelity depends, in part, on the orientation and training given to the schools and teachers implementing a tutoring system, but it also depends on the proper technical operation of the system. Implementation fidelity, although obviously an important concern, is seldom studied experimentally in evaluations of tutoring systems. The available evidence on its effects comes instead from natural experiments in which a weak implementation resulted from a technical or training failure in part but not all of an experiment. Although the weaker and stronger implementations were not planned, some researchers documented procedures and results sufficiently well to permit conclusions about the impact of implementation fidelity.

Koedinger and Anderson (1993) reported two sets of intelligent tutoring findings: one set for a teacher who was very experienced with intelligent tutoring programs and the other set for teachers who had far less experience with intelligent tutoring. The overall effect of intelligent tutoring in the study was to raise posttest scores by 0.35 standard deviations, but the effect sizes were very different for the experienced and inexperienced teachers. The effect size for the experienced teacher, based on a comparison of performance in his intelligent tutoring and conventionally taught classes, was 0.96. The effect size for the less experienced teachers, based on a comparison of their intelligent tutoring and conventionally taught classes, was –0.23. Students of the teachers who were not familiar with the tutoring program gained nothing from intelligent tutoring. Observations showed that the experienced teacher spent his laboratory time in content-related discussions with his students, whereas the inexperienced teachers were more likely to focus on technical problems and advice about the computer interface.

Le, Menzel, and Pinkwart (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 (e.g., 1-minute delays) in the computer tutor’s responses. This implementation of intelligent tutoring produced negligible effects on learning. The average effect size was 0.01. Technical problems were resolved by the second tutoring day. Students who received intelligent tutoring on this day showed a significant positive effect of intelligent tutoring. The average effect size was 0.28.

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

Pane et al. (2010) found a negative effect of Cognitive Tutor Geometry in a large field evaluation. Posttest scores of Cognitive Tutor classrooms were 0.19 standard deviations lower than posttest scores in control classrooms, but the researchers also found a correlation between implementation fidelity and effect sizes. Posttest scores from classes where teachers implemented the new curriculum faithfully were almost one-half standard deviation higher than posttest scores from classes where teachers implemented the curriculum less faithfully.

Overall, the average effect size for strong implementations in the four studies was 0.46, and the average effect size for weak implementations was –0.10, as shown in Table 7. This finding suggests that implementation fidelity may affect results of intelligent tutoring evaluations. However, adjusting overall results to take into account failures in implementation was impossible because so few reports contained information about implementation fidelity.

Table 7. Effect Sizes for Four Studies With Separate Measures From Stronger and Weaker Implementations of Intelligent Tutoring

	Effect Size	
Publication	Stronger Implementation	Weaker Implementation
Koedinger & Anderson (1993)	0.96	–0.23
Le et al. (2009)	0.28	0.04
VanLehn, et al. (2005)	0.57	0.21
Pane et al. (2010)	0.04	–0.42
Average	0.46	–0.10

B. Component Evaluations

In this section, we summarize results from 21 component evaluations, which were described in 11 separate reports. We excluded these evaluations from our meta-analysis of system evaluations because they do not include conventionally taught control groups. These evaluations, therefore, lack a baseline from which to measure the potential contributions of an alternative teaching system. However, component evaluations do provide useful information for system design. Through them, researchers can identify features that increase the effectiveness of a system so that developers can set priorities for revising it. The 21 evaluations examined three aspects of intelligent tutoring: interactivity, self-explanation, and interface features.

1. Interactivity

Student and tutor interactions play a key role in tutoring. Human tutoring employs mixed-initiative interactions. Tutors ask questions and students respond, or students ask questions and tutors respond. Carbonell's (1970) seminal paper cited mixed-initiative dialogue as a distinctive feature of intelligent computer-assisted systems. It remains a critical feature of ITSs, although the full, free-form interactivity of human tutoring still eludes us. Therefore, ITS researchers must, to some degree, limit student and tutor dialogue in their systems. The value of various dialogue capabilities remains a matter of interest, at least until computer natural language processing improves even more than it recently has.

In any case, researchers sometimes severely limit natural language interactions in ITSs for experimental purposes. What is left is usually a skeletal computer system that poses problems for students but does not provide adaptive hints or scaffolding to help students answer correctly. Instead, these stripped-down, limited-interactive systems usually give all students the same explanations and feedback. They do not try to "understand" individual student answers or misconceptions. The six studies of interactivity that we located differed substantially in the way they approached the question. For this reason, we describe the main features of each study separately below. Table 8 summarizes their results.

Table 8. Interactivity Effects in 11 Component Evaluations

Study	Interactive Condition	Non-Interactive Condition	Effect Size
Craig, Driscoll, and Gholson (2004), Experiment 1	Standard AutoTutor instruction	Students viewed the tutorial session of another "yoked" student	0.49
Craig et al. (2004), Experiment 2	Standard AutoTutor instruction	Students viewed the tutorial session of another "yoked" student	0.44
Craig, Sullins, Witherspoon & Gholson (2006), Experiment 1	Standard AutoTutor instruction	Students viewed the tutorial session of another "yoked" student	0.65
Craig et al. (2006), Experiment 2	Standard AutoTutor instruction	Students viewed the tutorial session of another "yoked" student	0.34
Gholson et al. (2009)	Standard AutoTutor instruction	Students viewed a monologue presentation of ideal problem solutions	0.04
Lane & VanLehn (2005)	ProPL (pronounced Pro-PELL) ITS	Read the same content	0.33

Table 8. Interactivity Effects in 11 Component Evaluations (Continued)

Study	Interactive Condition	Non-Interactive Condition	Effect Size
Moreno, Mayer, Spires, & Lester (2001)	Students solved problems with immediate feedback from computer tutor	Students read answers to problems without working on the problems	0.69
VanLehn et al. (2007), Experiment 1	Why2-Atlas and Why2-AutoTutor	Canned text remediation	–0.18
VanLehn et al. (2007), Experiment 3	Why2-Atlas and Why2-AutoTutor	Canned text remediation	0.28
VanLehn et al. (2007), Experiment 5	Why2-Atlas and Why2-AutoTutor	Canned text remediation	0.17
VanLehn et al. (2007), Experiment 6	Why2-Atlas and Why2-AutoTutor	Canned text and canned text remediation	0.11

Craig, Driscoll, & Gholson (2004) studied interactivity in two experiments with students at the University of Memphis. Each of the laboratory experiments provided 30 to 40 minutes of instruction in computer literacy. Students in the interactive conditions of the experiments interacted with the AutoTutor ITS in the normal fashion. Students in the non-interactive condition viewed recorded tutoring sessions of other students. The estimated effect size for Experiment 1 was 0.49, when the interactive condition is contrasted with the non-interactive vicarious one. The estimated effect size for interactivity in Experiment 2 was 0.44.

Craig, Sullins, Witherspoon, & Gholson (2006) reported results of two follow-up experiments that included a standard interactive tutoring condition and a standard vicarious condition, which were similar to the experimental and control conditions in Craig et al. (2004) Students in the standard interactive condition interacted with the AutoTutor ITS. Control students learned vicariously by viewing the recorded AutoTutor sessions of other students, which Craig et al. (2006) called yoked vicarious sessions. For Experiment 1, the estimated interactivity effect size is 0.65. For Experiment 2, the estimated interactivity effect size is 0.34.

Gholson et al. (2009) evaluated the importance of interactivity in computer literacy and physics instruction for students in Grades 8 through 11. Students in the interactive condition received about 3 hours of instruction via a standard AutoTutor system. Students in a control condition learned vicariously through a monologue presentation of ideal answers to problems. Pre-post gains were nearly identical for the standard AutoTutor group and the monologue group. The estimated interactivity effect size was 0.04.

Lane and VanLehn (2005) compared the performance of college students who used a dialogue-based ITS called ProPl with performance of a control group who read the same content. Participants were college students in an introductory programming course at the University of Pittsburgh. Lane and VanLehn compared performance of the two groups on a timed 2-hour lab assignment and a 75-minute posttest that targeted students' planning and algorithm writing skills. The average effect size for the lab assignments is 0.34. In addition, the ProPl group scored 0.31 standard deviations higher than the control group on a written posttest. The average effect size on the two outcome measures is 0.33.

Moreno, Mayer, Spires, and Lester (2001) carried out a series of five experiments on the use of a tutor representation, called a pedagogical agent, in computer-based teaching. The third experiment examined the importance of student interaction with the pedagogical agent. Participants in the experiment were college students at the University of California at Santa Barbara. These students were asked to design plants suited to specific environments for a lesson on plant ecology. Students in the interactive condition listened to a complete solution to a problem immediately after designing a plant. Students in the control group saw the same problems but listened to explanations without being able to design plants. Outcomes measured in the study were recall of factual information included in the lessons and ability to apply the information to solve new problems. The estimated effect size for interactivity is 0.69.

VanLehn et al. (2007) reported on results from six studies (and seven experiments) on human tutoring, intelligent computer tutoring, and reading textbook material. The first, third, and fifth studies in the series compared the effects of a fully interactive tutoring system with those of a less interactive version of the system. College students in the interactive conditions worked for 2 to 3 hours on physics problems, with help from an intelligent tutor. Students in the comparison groups (called canned-text-remediation groups) entered an essay in response to a physics question, read through the full sets of hints and scaffolding developed for tutorials on the question, and then edited their essays to take into account what they had learned from the text feedback. Individual student essays were not analyzed in the canned-text-remediation condition, and students did not receive adaptive feedback on flaws in their essays. The effect sizes in the three experiments, based on a comparison of tutoring versions with different levels of interactivity, were –0.18 for Study 1, 0.28 for Study 3, and 0.17 for Study 5.

The sixth study in the series, which combined results from the VanLehn et al. Experiments 6 and 7, contained an interactive condition that was similar to the interactive conditions in the other studies. However, unlike Studies 1, 3, and 5, it contained two control conditions: the canned-text-remediation condition described previously and a canned-text-only condition. Students in the canned-text-only condition read the problems along with an ideal answer to each problem, but the students did not write answers of their own. The authors found no significant difference in results from the canned-text-remediation and the canned-text-only conditions. In addition, performance of the students in the interactive condition was similar to performance of students in the control conditions. Effect size for interactivity for Study 6 was 0.11.

Effect sizes varied from moderately positive to moderately negative in the four VanLehn et al. (2007) studies. The authors attributed the variation in results to the amount of challenge in the instructional material used in an experiment. They reported that interactivity effects were positive and significant in studies with challenging lessons and tests (e.g., where students who had not taken college physics studied content written for students who had taken college physics). Findings were insignificant or negative in studies with lessons and tests that were less challenging (e.g., where novices studied material written for novices or students at the intermediate level studied material written for intermediate-level students).

Together, the 11 comparisons in the 6 interactivity evaluations suggest that tutoring sessions that include student-teacher interactions are more effective than systems that substantially reduce or eliminate student-tutor interactions. The average effect size is 0.31 in the 11 comparisons (see Table 8). While interactivity makes a crucial contribution to the effectiveness of ITSs, it does not fully explain their effectiveness. Reducing the interactivity of tutoring systems does not reduce their effectiveness to the level of conventional classroom instruction. It does, however, reduce their effectiveness to the level of older style computer-based instructional systems (Fletcher, 2003; C. L. C. Kulik & Kulik, 1991; J. A. Kulik, 1994).

2. Self-Explanation

Self-explanation prompts encourage students to reflect on their solutions to problems: why they chose a certain approach, why the approach did or did not work, what more general principle the approach represents, and so forth. Researchers have shown that such prompts can enhance learning, especially deeper understanding, in regular classrooms (e.g., Brown & Campione, 1994; Chi, 2000; Palincsar & Brown, 1984; White, Shimoda, & Frederiksen, 1999). We found six studies assessing the effectiveness of self-explanation prompting in ITSs. These studies are summarized in Table 9 and discussed in the remainder of this subsection.

Aleven and Koedinger (2002) carried out two experiments to determine whether explanation-prompting could improve the effectiveness of Cognitive Tutor. Students in the control group worked on high-school geometry problems for about 7 hours on a standard version of Cognitive Tutor. Students in the treatment group worked on the same problems with a special version of Cognitive Tutor that also prompted students to explain their answers in knowledge construction dialogues (KCDs). The researchers found that students who were required to explain their answers needed fewer problems to reach criterion levels than students who were not required to explain answers. In a first experiment, the amount of time-on-task was not controlled, and the KCD group spent 18% more time working on problems. The self-explanation effect was significant in this

Table 9. Effects of Self-Explanation Prompts in Six Component Evaluations

Study	Treatment	Comparison	Result
Aleven & Koedinger (2002), Experiment 1	Cognitive Tutor, plus KCDs	Cognitive Tutor without supplemental KCDs	T > C (p < .005)
Aleven & Koedinger (2002), Experiment 2	Cognitive Tutor, plus KCDs	Cognitive Tutor without supplemental KCDs	T > C, ns
Conati & VanLehn (1999)	SE-Coach provided intelligent prompting for self-explanation	SE-Coach presented canned text answers	ES = 0.12
Craig et al. (2006), Experiment 1	Vicarious viewing of ideal problem solutions, plus reflective questions posed by AutoTutor.	Vicarious viewing of ideal problem solutions, without reflective questions	ES =0.23
Gholson et al. (2009)	Vicarious viewing of ideal problem solutions, with embedded deep-level questions	Vicarious viewing of ideal problem solutions, without embedded deep-level questions	ES = 0.33
Siler, Rosé, Frost, VanLehn, & Koehler (2002)	Andes2 intelligent tutoring, plus KCDs	Andes2 presented mini-lessons with same content	ES = –0.32

Note for Table 9: ES = effect size; KCD = knowledge construction dialogue; T = treatment group; C = comparison group; ns = not significant; and p = probability

experiment, as measured by the number of problems required to reach criterion levels. In a second experiment, the researchers controlled the amount of time on task for the two groups, and the difference in performance for the two groups was small and marginally significant for success in solving harder problems requiring more student reasoning. We were not able to calculate effect sizes for these experiments.

Conati and VanLehn (1999) studied explanation prompts in a tutoring system, SE-Coach, that was designed primarily to prompt and shape student self-explanations. Their study compared a full version of this system to a stripped-down version from which self-explanation prompts were removed. The full version of SE-Coach guided students through the steps involved in solving a specific physics problem and, at each step, prompted students for explanations. SE-Coach provided menus to make it easier for students to construct their explanations and also offered correctives for unsatisfactory explanations. The stripped-down version of SE-Coach presented the problem-solving steps in each solution as canned text. Students were asked to read the steps but were not asked to explain them. Examination scores of students in the self-explanation group were slightly, but not significantly, higher than scores of students in the control group. The effect size was 0.03 for the researchers' objective-based scoring of the examination problems and 0.21 for their feature-based scoring. The average effect size for the study was 0.12. In two follow-up studies, Conati and VanLehn (2000a, 2000b) looked at effects in subgroups of students. The sample size in most subgroups was too small, however, to yield statistically reliable findings.

The first experiment in Craig et al.’s (2006) report, described in the previous subsection, included vicarious learning conditions with and without deep-level questions. The deep-level questions were designed to stimulate the students to reflect on concepts that were used in solving problems. In the full-questions-vicarious condition, the experimenter added deep-level questions before each step in the ideal answer to each problem. In the monologue-vicarious condition, students listened to ideal answers that did not include embedded questions. Students in both conditions listened to the ideal answers without interacting with the tutor in any other way. The effect size for deep-level questions was 0.23.

In addition to examining interactivity effects, the study by Gholson et al. (2009) examined the effects of deep-level questions in two vicarious learning conditions, which Gholson and his colleagues referred to as dialogue and monologue conditions. In the dialogue condition, the experimenter embedded deep-level questions before each tutorial interaction. In the monologue condition, students listened to recordings without the embedded deep-level questions. The estimated effect size for deep-reasoning questions was 0.33.

Siler, Rosé, Frost, VanLehn, & Koehler (2002) carried out three studies that evaluated the effects of KCDs by comparing two programs delivered by the Andes2 tutoring system. The first program presented students with KCDs, whereas the second program provided mini-lessons on the same content. Students in KCD sessions were asked to explain their answers, were given menus to help them construct explanations, and were also given feedback on their explanations. Students in the control condition read about the same concepts in specially prepared mini-lessons. The mini-lessons severely limited interactions but contained all the conceptual content of the corresponding KCDs. The participants in the experiment were paid college student volunteers who had completed laboratory lessons on physics concepts. In the first and third experiments, the pretest scores of the treatment and comparison groups were significantly and substantially different, so the results of these experiments cannot be used to draw conclusions about the experimental treatment. The second experiment in the series, however, was not flawed by initial differences in comparison groups. The estimated KCD effect size for this study is –0.32.

Overall, the six self-explanation studies found an average effect size of 0.09 for self-explanation, a null effect. These studies therefore suggest that tutoring systems that prompt students to explain their answers are about as effective as systems that do not. Although explicit self-explanation prompting does not appear to hurt tutoring programs, in aggregate, these studies suggest that it does little to increase their effectiveness.

3. Interface Features

Four component evaluations examined the importance of several interface features of tutoring systems—far too few studies to provide definitive answers about the importance of any specific feature. Nonetheless, these studies suggest that interface features can influence the effectiveness of tutoring systems. Specifically, these interface evaluations suggest that tutorial systems work better when they (1) allow students to explore a knowledge domain flexibly rather than in lockstep fashion, (2) provide information in a game-like rather than purely didactic manner, and (3) provide spoken rather than text-only instruction and feedback. The evaluations, therefore, suggest that interface design may be a fruitful area for further research and development.

a. Flexible vs. Inflexible Exploration

Mark and Greer (1995) compared the effects of different ways of teaching learners to program a video cassette recorder (VCR). Participants in the study were undergraduate students at the University of Saskatchewan and had no prior experience in VCR programming. The students worked with four different computer simulations of a VCR. Two of the simulations differed in the amount of exploration that they allowed students. Specifically, a sequence-based instructional program (Mark-II) required students to follow a rigid sequence of steps in programming the VCR, whereas a device-based version (Mark-III) allowed students to follow a flexible sequence of steps. The students who were given more freedom to explore the simulation outperformed the students who did not have this freedom. The Mark-III students took fewer steps to program a VCR on a post-training lab assignment and also made fewer errors on a post-training test. The effect size for flexibility was 0.35. The authors concluded that flexible programs are more effective because they give students more room to explore domains and build conceptual models of them.

b. Game Playing

Virvou, Katsionis, and Manos (2005) compared the effectiveness of a standard ITS with a conventional user interface to the effectiveness of the same ITS with a VR-ENGAGE interface, which gave the system the look and feel of a virtual reality game. Participants were fourth grade children studying geography in elementary schools in Greece. Effect size for the gaming interface was 0.28. The effect of switching from a standard interface to a game-playing one was small but significant.

c. Tutor Representations

Moreno et al.'s study (2001), described in Section 3.B.1, examined the role that tutor representations can play in computer-based instruction. The researchers carried out five experiments with a tutor representation, or pedagogical agent, named Herman. In the first two experiments in the series, Moreno and her colleagues compared two experimental conditions. In the first condition, Herman appeared on screen and presented spoken directions and feedback to students. In the second condition, directions and feedback were presented as written text, and Herman's image and voice were absent. The effect of the pedagogical agent on student learning was clear. In Experiment 1, the effect size associated with the presence of a pedagogical agent was 0.81. In Experiment 2, the pedagogical agent raised scores by an average of 0.54 standard deviations.

Students in Moreno et al.'s Experiments 1 and 2 were exposed to two aspects of the pedagogical agent: they saw Herman, and they listened to him. Moreno et al. carried out two additional studies to determine which of these attributes was more important for student learning (i.e., Was it more important to see or to hear a pedagogical agent?). Moreno's Experiments 4 and 5 manipulated two aspects of the pedagogical agent: whether or not the agent provided spoken narration and whether or not the pedagogical agent was visually present. The pedagogical agent was an animated character in Experiment 4 and a recorded image of an expressive actor in Experiment 5. Results showed that the sight of the tutor was far less important than his spoken voice. Test scores were nearly identical for conditions with and without a visual image. Test scores from conditions with spoken narration, however, were much higher than test scores for conditions with text only. Spoken narration raised test scores by an average of 0.99 standard deviations.

Moreno et al.'s experiments therefore suggest that spoken narration can be very important in computer-based teaching systems, but further research is needed to determine the degree to which the study's findings generalize to other tutoring systems. In two studies, for example, VanLehn et al. (2007) found very similar results for (1) Why2-Atlas, a tutoring system that provides text-only answers to students, and (2) Why2-AutoTutor, a system that provides spoken guidance and feedback. In the fifth study in VanLehn's report, Why2-Atlas posttest scores were 0.08 standard deviations higher than Why2-AutoTutor scores. In the sixth study, which included results from VanLehn et al.'s Experiments 6 and 7, Why2-Atlas posttest scores were 0.27 standard deviations higher. On average, the posttest difference in results for the two tutoring systems was 0.18 standard deviations. Because feedback modality is not the only way in which Why2-Atlas and Why2-AutoTutor differ, this comparison of the two systems only provides suggestive evidence about the effects of spoken vs. written feedback in tutoring systems.

4. Discussion and Conclusions

A. Scope

In 1983, Richard Clark published a widely cited article with the well wrought and often repeated assertion that “The best current evidence is that media are mere vehicles that deliver instruction but do not influence student achievement any more than the truck that delivers our groceries causes changes in our nutrition“ (page 445). In short and perhaps oversimplified, it is not technology, but what you do with it that matters. All of the system evaluation studies we reviewed compared a technology-based approach (i.e., ITSs) with classroom instruction. It might then be argued that our concern and that of the studies we reviewed was with a medium rather than an instructional method. However, just because a method uses computers does not necessarily condemn it to the media-based dustbin. As our review shows, there are differences in the effectiveness of ITSs, as with any method. However, with the focus on reliably replicating the capabilities of one-on-one human tutoring with computers (thereby increasing the accessibility and affordability of such tutoring), research on ITSs appears to reside securely within the instructional method camp as a concern with a bona fide instructional method.

B. Findings

This review shows that ITSs can be effective instructional tools. Students who received intelligent tutoring outperformed students from conventional classes in 41 (91%) of the 45 system evaluations that we examined, and the improvement in performance was great enough to be considered of substantive importance in 36 (80%) of the 45 studies. The median effect size in the 45 studies was 0.63, 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 72nd percentile.

This effect size may underestimate the true effectiveness of ITSs, however, because the posttests used in some of the system evaluations were poorly aligned with the teaching objectives of the tutoring systems. Specifically, nine evaluations of Carnegie Learning’s Cognitive Tutor employed standardized, multiple-choice tests that were not closely aligned with the instructional objectives of the Cognitive Tutor programs. In addition, three other system evaluations included posttests that were poorly aligned with the instructional objectives emphasized in the tutoring programs. Most of the trivial effects that we found came from these poorly aligned tests. When results from poorly aligned tests were eliminated from our analysis, median effect size for intelligent tutoring was 0.75 in 39 evaluations, and effect sizes were large enough to be considered of substantive importance in 38 (97%) of the 39 studies.

Another factor that can influence the results of an intelligent tutoring program is the fidelity in implementing the program. Very few system evaluations measured fidelity of implementation, but the few that did suggested that intelligent tutoring effects are much stronger when teachers implement intelligent tutoring programs carefully and completely and are weaker when teachers do not implement intelligent tutoring programs properly or when technical problems affect the implementations. One large study found a difference in effect sizes of one-half standard deviation between strong and weak implementations of an ITS (Pane et al., 2010). Another study found a difference in effect size well in excess of 1.00 when the learning by students with a teacher experienced in using the Cognitive Tutor was compared to learning by students whose teachers lacked this advantage (Koedinger & Anderson, 1993). These findings suggest the need to better understand and systematically develop (human) teacher skills and strategies for working effectively with ITSs.

Whether we consider findings from all studies or only from studies with well-aligned posttests and strong implementations, it is clear that ITSs surpass older forms of computer-based instruction in effectiveness. For example, a 1994 review, which aggregated results from 12 separate meta-analyses carried out at 8 different universities, found an average effect size of 0.35 for these older approaches (J. A. Kulik, 1994). The largest of the meta-analyses cited in the review covered 254 reports. The average effect size in the 254 studies was 0.3, roughly equivalent to an increase from the 50th to the 62nd percentile (C. L. C Kulik & Kulik, 1991). Thus, older forms of computer-based instruction, on average, raised posttest scores about one-third standard deviation over scores from conventional classrooms. The gains from ITSs are twice as high.

The instructional gains from intelligent tutoring are also greater than the gains most often found with human tutoring. The five meta-analyses we reviewed earlier found that tutored students outperformed students who learned in conventional classrooms (P. A. Cohen et al., 1982; Hartley, 1977; Mathes & Fuchs, 1994; G. W. Ritter et al., 2009; VanLehn, 2011). The median effect size in the five meta-analyses was 0.4. For a long time, the goal of developers of ITSs has been to match the success of human tutoring. Our results suggest that ITSs have already met this goal.

Interestingly, the average effect size in the system evaluations we reviewed is very close to the average effect that VanLehn (2011) found for ITSs. VanLehn found 27 studies that compared posttest scores of students taught with and without intelligent tutoring. He found average effect sizes of 0.40 for substep-based forms of intelligent tutoring and 0.76 for step-based forms. The overall average effect size was 0.58. The similarity of VanLehn's overall findings to ours is remarkable, given that the two reviews differed substantially in search procedures, inclusion criteria, and effect-size calculation. For example, VanLehn examined studies found in computer science journals and conferences, whereas we cast a wider net for studies. When a study reported results on several learning measures, VanLehn used the strongest effect size to represent the study results, whereas we averaged effects over all measures to derive an overall effect size. VanLehn also included data from several kinds of control groups in his analyses, whereas we restricted our analyses to comparisons with conventionally instructed controls. Despite these differences, VanLehn's review and our review found overall effects of intelligent tutoring to be similar.

VanLehn's review also examined 15 evaluations that directly compared the effects of ITSs and human tutoring. Ten of the evaluations examined step-based intelligent tutoring, and five examined substep-based tutoring. The average difference in posttest performance from intelligent tutoring and human tutoring was only 0.10 standard deviations. Test scores of students who were tutored by humans were 0.10 standard deviations higher than the test scores of students who received intelligent tutoring via computer. The difference is not large enough to be considered of substantive importance. VanLehn's finding is also consistent with our conclusions about human tutoring.

On the other hand, our conclusions about Carnegie Learning's Cognitive Tutor are different from those drawn by Slavin et al. (2009) and the What Works Clearinghouse (August 2010). We found that Carnegie Tutor effects differed substantially on locally developed, problem-solving tests and standardized multiple-choice tests. The average effect size was 0.72 on local tests designed to measure the higher order objectives stressed in the Carnegie Tutor curriculum, but the average effect size was 0.10 on multiple-choice tests that did not directly measure problem-solving skills. Slavin found an average effect size of 0.12 in seven studies of Cognitive Tutor Algebra, and the What Works Clearinghouse found a near-zero average effect in four studies. We concluded that Cognitive Tutor effects were large enough to be of substantive importance. Slavin and his colleagues concluded that there was "limited evidence of effectiveness" of Cognitive Tutor, based on their finding that at least one study in their analysis had an effect size of at least 0.10. The What Works Clearinghouse researchers concluded that evidence was moderate to strong that Cognitive Tutor had no discernible effect on student achievement.

Reasons for the different conclusions about Cognitive Tutor seem evident. The conclusions drawn by Slavin and his colleagues and by the What Works Clearinghouse were based on results from standardized tests that were poorly aligned with the instructional objectives of the Cognitive Tutor curriculum. Results from tests of problem-solving that fit the Cognitive Tutor curriculum were not included in their analyses. If we had based our conclusions about Cognitive Tutor solely on such tests, we might have reached similar conclusions. We believe, however, that it is a mistake to draw conclusions solely on tests that seem poorly aligned to the higher order instructional objectives stressed in most ITSs.

Studies that examined components of ITSs provided some additional insights into their workings. For example, the component evaluations reinforce earlier findings that frequent and meaningful interactions between students and the system contribute substantially to learning effectiveness. However, this interactivity alone does not explain the effectiveness of the systems. Eliminating interactivity from ITSs reduces their effectiveness by around 0.3 standard deviations in a typical study. It thus reduces the effectiveness of these computer systems to the level of older forms of computer-based instruction but not to the level of conventional teaching.

The component evaluations also suggest that interface improvements might increase the effectiveness of ITSs. Interface studies suggest that tutorial systems work better when they (1) require interactive rather than passive participation, (2) provide spoken rather than text-only instruction and feedback, (3) provide information in a game-like, rather than purely didactic, manner, and (4) allow students to explore a knowledge domain flexibly rather than in lockstep fashion. A caveat, however, is that each of these findings is based on only one or two laboratory studies. For greater confidence in these conclusions, we need to see these findings replicated in other settings, including real school settings, with different learners, and with a wider selection of instructional materials.

C. Final Word

The results of this review suggest that ITSs make substantial improvements over those of other instructional approaches by accelerating learning, expanding learners' problem-solving competencies, and developing the deep conceptual understanding that is needed for retention and transfer—doing so at scales that would be unaffordable if based on human tutoring or classroom instruction. Overall, the findings of this review are sufficiently promising to recommend continued research, development, and application of ITSs.

Illustrations

References

Aleven, V. A., & Koedinger, K. R. (2002). “An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor.” *Cognitive Science 26*(2), 147–179. doi: 10.1207/s15516709cog2602_1

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). “Cognitive modeling and intelligent tutoring.” *Artificial Intelligence 42*(1), 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*(2), 167–207. Retrieved from http://act-r.psy.cmu.edu/papers/129/CogTut_Lessons.pdf.

Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Columbus, OH: Allyn & Bacon.

Arbuckle, W. J. (2005). *Conceptual understanding in a computer-assisted Algebra 1 classroom* (Doctoral dissertation). The University of Oklahoma, Norman, Oklahoma. Retrieved from Dissertations & Theses: Full Text database (Publication No. AAT 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*(3), 694–698. doi: 10.3758/BRM.40.3.694

Atkinson, R. C., & Paulson, J. A. (1972). “An approach to the psychology of instruction.” *Psychological Bulletin 78*(1), 49–61. doi: 10.1037/h0033080

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

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. Retrieved from http://www.jstor.org/stable/1175554?seq=2.

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Burns, L. M. (1993). *MEADOW: An integrated system for intelligent tutoring of subtraction concepts and procedures* (Doctoral dissertation). Columbia University, New York, NY. Retrieved from *Dissertation Abstracts International 54*(07), 2510A.

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, Inc.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts.* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Carbonell, J. R. (1970). "AI in CAI: An artificial intelligence approach to computer-assisted instruction." *IEEE Transactions on Man-Machine Systems 11*(4), 190–202. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04081977.

Carlson, P. A., & Miller, T. M. (1996). *Beyond word processing: Using an interactive learning environment to teach writing* (Technical Report AL/HR-TR-1996-0090). Brooks AFB, TX: Human Resources Directorate, Technical Training Research Division. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA319034.

Chi, M. T. H. (2000). Self-explaining: The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates.

Clark, R. E. (1983). "Reconsidering research on learning from media." *Review of Educational Research, 53*(4), 445–459. doi: 10.3102/00346543053004445

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*(2), 237–248. doi: 10.3102/00028312019002237

Conati, C., & VanLehn, K. (1999). Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In S. P. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 297–304). Amsterdam: IOS Press. Retrieved from http://www.public.asu.edu/~kvanlehn/Stringent/PDF/99AIED_CC_KVL.pdf.

Conati, C., & VanLehn, K. (2000a). Further results from the evaluation of an intelligent computer tutor to coach self-explanation. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems: 5th International Conference* (Vol. 1839, pp. 304–313). Berlin: Springer-Verlag.

Conati, C., & VanLehn, K. (2000b). "Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation." *International Journal of Artificial Intelligence in Education 11*(4), 389–415. Retrieved from http://iaied.org/pub/943/file/943_paper.pdf.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation Publications.

Corbett, A. T. (2001a). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewica, & J. Vassileva (Eds.), *Lecture Notes in Computer Science: Proceedings of the 8th International Conference on User Modeling* 2001 (Vol. 2109, pp. 137–147). London: Springer-Verlag.

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

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

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, M. Beaudouin-Lafon, & R. Jacob (Eds.), *Proceedings of the ACM CHI 2001 Conference on Human Factors in Computing Systems* (pp. 245–252). New York: ACM Press. doi: 10.1145/365024.365111

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*(2), 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*(4), 565–591. Retrieved from http://www.tandfonline.com/doi/pdf/10.1207/s1532690xci2404_4.

Fletcher, J. D. (1989). "The effectiveness and cost of interactive videodisc instruction." *Machine-Mediated Learning 3*(4), 361–385.

Fletcher, J. D. (1992). Individualized systems of instruction. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp. 613–620). New York, NY: Macmillan.

Fletcher, J. D. (2003). Evidence for learning from technology-assisted instruction. In H. F. O'Neil Jr. & R. Perez (Eds.) *Technology applications in education: A learning view* (pp. 79–99). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fletcher, J. D. (2009). "Education and training technology in the military." *Science 323*(5910), 72–75. doi: 10.1126/science.1167778

Fletcher, J. D. (2011). *DARPA education dominance program: April 2010 and November 2010 digital tutor assessments* (Report IDA-NS-D-4260). Alexandria, VA: Institute for Defense Analyses.

Fletcher, J. D., & Atkinson, R. C. (1973). An evaluation of the Stanford CAI program in initial reading (grades K through 3). *Journal of Educational Psychology 63*(6), 597–602. doi: 10.1037/h0034065

Gabriel, T., & Richtel, M. (2011, October 9). Inflating the software report card: School technology companies ignore some results." *New York Times,* pp. A1, A22.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., Sullins, J., & 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*(5), 487–493. doi: 10.1007/s11251-008-9069-2

Glass, G. V. (1976). "Primary, secondary, and meta-analysis of research." *Educational Researcher 5*(10), 3–8. doi: 10.3102/0013189X005010003

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

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 Publications.

Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., … the Tutoring Research Group. (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 25th Annual Conference of the Cognitive Science Society* (pp. 1–5). Boston, MA: Cognitive Science Society. 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*(2), 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.), *Proceedings of Artificial Intelligence in Education* (pp. 47–54). Amsterdam: IOS Press. Retrieved from http://www.cs.pitt.edu/~chopin/references/tig/AIED-graesser-2003.pdf.

Grubišić, A., Stankov, S., Rosić, M., & Žitko, B. (2009). "Controlled experiment replication in evaluation of e-learning system's educational influence." *Computers and Education 53*(3), 591–602. doi:10.1016/j.compedu.2009.03.014

Grubišić, 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 (WSEAS).

Hartley, S. S. (1977). *Meta-analysis of the effects of individually paced instruction in mathematics* (Doctoral dissertation). University of Colorado, Boulder, CO. Retrieved from *Dissertation Abstracts International 38*(7-A), 4003.

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 1* (Vol. 6094, pp. 204–213). Berlin: Springer-Verlag. Retrieved from http://reed.cs.depaul.edu/peterh/papers/Hastingsits2010.pdf.

Hategekimana, C. P. (2008). *Cognition and technology: Effectiveness of intelligent tutoring systems for software training* (Doctoral dissertation).Iowa State University, Ames, IA. Retrieved from http://lib.dr.iastate.edu/etd/11409/.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). "Eliciting knowledge from experts: A methodological analysis." *Organizational Behavior and Human Decision Processes 62*(2), 129–158. Retrieved from http://eprints.soton.ac.uk/252301/1/El%20Know%20from%20Exp%20OBHD.pdf.

Jeremic, Z., Jovanovic, J., & Gasevic, D. (2009). "Evaluating an intelligent tutoring system for design patterns: The DEPTHS experience." *Educational Technology and Society 12*(2), 111–130. Retrieved from http://www.ifets.info/journals/12_2/9.pdf.

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*(3), 199–214.

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. doi: 10.1007/978-3-540-30139-4_16

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://pact.cs.cmu.edu/koedinger/pubs/Koedinger%20&%20Anderson%2093.pdf.

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*(1), 30–43. Retrieved from http://ctat.pact.cs.cmu.edu/pubs/Koedinger-Anderson.pdf.

Kulik, C. L. C., & Kulik, J. A. (1991). "Effectiveness of computer-based instruction: An updated analysis." *Computers in Human Behavior 7*(1–2), 75–94. Retrieved from http://deepblue.lib.umich.edu/bitstream/2027.42/29534/1/0000622.pdf.

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*(3), 183–

201. Retrieved from http://www.tandfonline.com/doi/pdf/10.1080/08993400500224286.

Le, N. T., Menzel, W., & Pinkwart, N. (2009). Evaluation of a constraint-based homework assistance system for logic programming. In S. C. Kong, H. Ogata, H. C. Arnseth, C. K. K. Chan, T. Hirashima, F. Klett, J. H. M. Lee, … S. J. H. Yang (Eds.), *Proceedings of the 17th International Conference on Computers in Education* [CDROM]. Hong Kong: Asia-Pacific Society for Computers in Education. Retrieved from http://www.icce2009.ied.edu.hk/pdf/C1/proceedings051-058.pdf.

Mark, M. A., & Greer, J. E. (1995). “The VCR tutor: Effective instruction for device operation.” *Journal of the Learning Sciences 4*(2), 209–246. Retrieved from http://www.jstor.org/stable/1466691?seq=2.

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*(1), 59–80.

Mendicino, M., & Heffernan, N. (2007). “Comparing the learning from intelligent tutoring systems, non-intelligent computer-based versions, and traditional classroom instruction.” Morgantown, WV: West Virginia University, Educational Psychology Program.

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*(3), 331–358. Retrieved from http://teacherwiki.assistment.org/wiki/images/1/13/Jrte_layout.pdf.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). “The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?” *Cognition and Instruction 19*(2), 177–213. Retrieved from http://www.jstor.org/stable/3233816.

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*(1), 109–114. Retrieved from http://www.aensiweb.com/jasr/jasr/2009/109-114.pdf.

Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them* (Technical Report IHMC Cmap Tools 2006-01 Rev 01-2008). Pensacola, FL: Florida Institute for Human and Machine Cognition. Retrieved from http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf.

Palincsar, A., & Brown, A. (1984). “Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities.” *Cognition and Instruction 1*(2). 117–175. Retrieved from http://people.ucsc.edu/~gwells/Files/Courses_Folder/ED%20261%20Papers/Palincsar%20Reciprocal%20Teaching.pdf.

Pane, J. F., McCaffrey, D. F., Steele, J. L., Ikemoto, G. S., Slaughter, M. E. (2010). "An experiment to evaluate the efficacy of cognitive tutor geometry." *Journal of Research on Educational Effectiveness 3*(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*(3), 183–189. Retrieved from http://lvstem.cse.lehigh.edu/documents/papers/ccsce06_parvez.pdf.

Pek, P.–K., & Poh, K.–L. (2005). "Making decisions in an intelligent tutoring system." *International Journal of Information Technology and Decision Making 4*(2), 207–233.

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: IOS Press.

Reif, F., & Scott, L. A. (1999). "Teaching scientific thinking skills: Students and computers coaching each other." *American Journal of Physics 67*(9), 819–831.

Reiser, B. J., Anderson, J. R., & Farrell, R. G. (1985). Dynamic student modeling in an intelligent tutor for LISP programming. In A. K. Joshi (Ed.), *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (Vol. 1, pp. 8–14). San Francisco, CA: Morgan Kaufmann.

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*(1), 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.), *Proceedings of the 2007 Conference on Supporting Learning Flow through Integrative Technologies* (pp. 13–20). Amsterdam: IOS Press.

Shneyderman, A. (2001). *Evaluation of the cognitive tutor Algebra I program.* Miami, FL: Miami–Dade County Public Schools, Office of Evaluation and Research. Retrieved from http://oer.dadeschools.net/algebra.pdf.

Siler, S., Rosé, C. P., Frost, T., VanLehn, K., & Koehler, P. (2002). Evaluating knowledge construction dialogs (KCDs) versus minilessons within Andes2 and alone. In *ITS2002 Workshop on Empirical Methods for Tutorial Dialogue Systems* (pp. 9–15). Spain: San Sebastian. Retrieved from http://www.public.asu.edu/~kvanlehn/Not%20Stringent/PDF/02ITSW_SS_CR_TF_KVL_PK.pdf.

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*(2), 839–911. doi: 10.3102/0034654308330968

Smith, J. E. (2001). *The effect of the Carnegie Algebra Tutor on student achievement and attitude in introductory high school algebra.* (Doctoral dissertation). Virginia Polytechnic Institute and State University: Blacksburg, VA. Retrieved from *Dissertation Abstracts International 63*(09), 3078A.

Stankov, S., Glavinić, V., & Grubišić, 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.), *IEEE International Conference on Intelligent Engineering Systems* (pp. 545–550). Romania: Cluj-Napoca.

Stankov, S., Rosić, M., Žitko, B., & Grubišić, A. (2008). "TEx-Sys model for building intelligent tutoring systems." *Computers and Education 51*(3), 1017–1036. Retrieved from http://bib.irb.hr/datoteka/343294.CE_November_2008.pdf.

Steuck, K., & Miller, T. M. (1997). "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.

Suppes, P., Fletcher, J. D., & Zanotti, M. (1976). "Models of individual trajectories in computer-assisted instruction for deaf students." *Journal of Educational Psychology 68*(2), 117–127. doi: 10.1037/0022-0663.68.2.117

Suppes, P. & Morningstar, M. (1972). *Computer-assisted instruction at Stanford 1966–68: Data, models, and evaluation of the arithmetic programs.* New York: Academic Press.

Suraweera, P., & Mitrovic, A. (2002). KERMIT: A constraint-based tutor for database modeling. In *Procedures of the 6th International Conference on Intelligent Tutoring Systems, ITS 2002, Lecture Notes in Computer Science 2363*, 377–387. Berlin: Springer-Verlag.

Timms, M. J. (2007). Using item response theory (IRT) to select hints in an ITS. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proceedings of the Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (pp. 213–221). Amsterdam, Netherlands: IOS Press.

Van Campen, J. (1981). A computer-assisted course in Russian. In P. Suppes (Ed.), *University-level computer-assisted instruction at Stanford: 1968–1980.* Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences.

VanLehn, K. (2011). "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems." *Educational Psychologist 46*(4), 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*(1), 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*(3), 147–204. Retrieved from http://www.andestutor.org/Pages/AndesLessonsLearnedForWeb.pdf.

Virvou, M., Katsionis, G., & Manos, K. (2005). “Combining software games with education: Evaluation of its educational effectiveness.” *Educational Technology & Society 8*(2), 54–65. Retrieved from http://www.ifets.info/journals/8_2/5.pdf.

What Works Clearinghouse. (2010). *Carnegie learning curricula and cognitive tutor software*. WWC Intervention Report: High School Math. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_cogtutor_083110.pdf.

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*(2), 243–254. doi: 10.1016/S0747-5632(99)00021-7

White, B. Y., Shimoda, T. A., & Frederiksen, J. R. (1999). “Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development.” *International Journal of Artificial Intelligence in Education 10*(2), 151–182. Retrieved from http://thorndike.tc.columbia.edu/~david/MTSU4083/Readings/Inquiry-based%20ID/White-EnablingStudentstoConstruct.pdf.

Woolf, B., & McDonald, D. (1984). “Design issues in building a computer tutor.” *IEEE Computer 17*(9), 61–73. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01659246.

Abbreviations

AFB	Air Force Base
APT	ACT Programming Tutor
CID	Center for Information Dominance
DEPTHS	Design Patterns Teaching Helping System
DTex-Sys	Distributed Tutor Expert System
DTIC	Defense Technical Information Center
ERIC	Educational Resources Information Clearinghouse
FOSS	Full Option Science System
GREATERP	Goal-Restricted Environment for Tutoring and Educational Research on Programming
ICAI	intelligent computer-assisted instruction
ISIS	Instruction in Scientific Inquiry Skills
ITS	intelligent tutoring system
KCD	knowledge construction dialogue
LSAT	Law School Admissions Test
NTIS	National Technical Information Service
PAL	Personal Assistant for Learning
PAT	Practical Algebra Tutor
PUMP	Pittsburgh Urban Math Project
RMT	Research Methods Tutor
R-WISE	Reading and Writing in a Supportive Environment
VCR	video cassette recorder
WPS	Word Problem Solving
xTex-Sys	eXtended Tutor-Expert System

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE June 2012	2. REPORT TYPE Final	3. DATES COVERED *(From–To)* October 2011 – May 2012

4. TITLE AND SUBTITLE

Effectiveness of Intelligent Tutoring Systems

5a. CONTRACT NUMBER
W91WAW-11-C-0003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

James A. Kulik
J.D. Fletcher

5d. PROJECT NUMBER

5e. TASK NUMBER
AI-2-3370

5f. WORK UNIT NUMBER

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

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

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-4664
Log: H12-000988

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

Assistant Secretary of Defense for
Research and Engineering
4800 Mark Center Drive
Alexandria, VA 22350-3600

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (4 March 2013).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This review examined findings from 66 studies of intelligent tutoring systems: 45 system evaluations and 21 component evaluations. The system evaluations compared learning gains from intelligent tutoring systems to learning gains from conventional classroom instruction. The component evaluations assessed the effectiveness of specific features of intelligent tutoring systems. A typical component evaluation compared learning from two different versions of an intelligent tutoring system, one with and one without a certain feature.

The average effect of intelligent tutoring in the system evaluations was to raise student test scores by approximately 0.60 standard deviations over the test scores of conventionally taught students, roughly equivalent to an improvement from the 50th to the 72nd percentile. Although tutoring effects were much greater in some evaluations, they were near zero in other studies. All but one of the near-zero effects came from tests that were poorly aligned with the higher-order teaching objectives emphasized in the tutoring systems. When results from these poorly aligned tests were eliminated from the analysis, median effect size for intelligent tutoring was 0.75 in 39 evaluations. Another factor that influenced study results was the implementation fidelity of the tutoring program – the care and attention with which the program was implemented in a classroom. Programs with careful attention to implementation were significantly more effective than those without it.

The component evaluations concerned the effectiveness of specific factors in intelligent tutoring programs. Interactivity was found to be a key factor in tutoring effectiveness. Reducing or eliminating student-tutor interactivity severely reduced the effectiveness of computer tutoring programs. Component evaluations also found that improvements in user-interfaces will improve tutoring results. Specifically, tutoring systems produced higher post-instruction scores when they (a) offered spoken rather than text-only instruction and feedback; (b) provided information in a game-like rather than purely didactic manner; and (c) allowed students to explore a domain flexibly rather than in a pre-specified fashion.

15. SUBJECT TERMS

Intelligent Tutoring Systems, Computer Assisted Instruction, Computer Based Instruction, Education, Training, Meta-analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Dr. Pat Mason
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	60	19b. TELEPHONE NUMBER *(include area code)* (703) 588-7420