Journal of Educational Psychology
2013, Vol. 105, No. 4, 970–987

0022-0663/13/$12.00 DOI: 10.1037/a0032447

A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on K–12 Students' Mathematical Learning

Saiying Steenbergen-Hu and Harris Cooper
Duke University

In this study, we meta-analyzed empirical research of the effectiveness of intelligent tutoring systems (ITS) on K–12 students' mathematical learning. A total of 26 reports containing 34 independent samples met study inclusion criteria. The reports appeared between 1997 and 2010. The majority of included studies compared the effectiveness of ITS with that of regular classroom instruction. A few studies compared ITS with human tutoring or homework practices. Among the major findings are (a) overall, ITS had no negative and perhaps a small positive effect on K–12 students' mathematical learning, as indicated by the average effect sizes ranging from $g = 0.01$ to $g = 0.09$, and (b) on the basis of the few studies that compared ITS with homework or human tutoring, the effectiveness of ITS appeared to be small to modest. Moderator analyses revealed 2 findings of practical importance. First, the effects of ITS appeared to be greater when the interventions lasted for less than a school year than when they lasted for 1 school year or longer. Second, the effectiveness of ITS for helping students drawn from the general population was greater than for helping low achievers. This finding draws attentions to the issue of whether computerized learning might contribute to the achievement gap between students with different achievement levels and aptitudes.

Keywords: intelligent tutoring systems, effectiveness, mathematical learning, meta-analysis, achievement

Intelligent tutoring systems (ITS) are computer-assisted learning environments created using computational models developed in the learning sciences, cognitive sciences, mathematics, computational linguistics, artificial intelligence, and other relevant fields. ITS often are self-paced, learner-led, highly adaptive, and interactive learning environments operated through computers. ITS are adaptive in that they adjust and respond to learners with tasks or steps to suit learners' individual characteristics, needs, or pace of learning (Shute & Zapata-Rivera, 2007).

ITS have been developed for mathematically grounded academic subjects, such as basic mathematics, algebra, geometry, and statistics (Cognitive Tutor: Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997; Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007; AnimalWatch: Beal, Arroyo, Cohen, & Woolf, 2010; ALEKS: Doignon & Falmagne, 1999); physics (Andes, Atlas, and Why/Atlas: VanLehn et al., 2002, 2007); and computer science (dialogue-based intelligent tutoring systems: Lane & VanLehn, 2005; ACT Programming Tutor: Corbett, 2001). Some ITS assist with the learning of reading (READ 180: Haslam, White, & Klinge, 2006; iSTART: McNamara, Levinstein, & Boonthum, 2004), writing (R-WISE writing tutor: Rowley, Carlson, & Miller, 1998), economics (Smithtown: Shute & Glaser, 1990), and research methods (Research Methods Tutor: Arnott, Hastings, & Allbritton, 2008). There are also ITS for specific skills, such as metacognitive skills (see Aleven, McLaren, & Koedinger, 2006; Conati & VanLehn, 2000). The use of ITS as an educational tool has increased considerably in recent years in U.S. schools. Cognitive Tutor by Carnegie Learning, for example, was used in over 2,600 schools in the United States as of 2010 (What Works Clearinghouse, 2010a).

ITS are developed so as to follow the practices of human tutors (Graesser, Conley, & Olney, 2011; Woolf, 2009). They are expected to help students of a range of abilities, interests, and backgrounds. Research suggests that expert human tutors can help students achieve learning gains as large as two sigmas (Bloom, 1984). Although not as effective as what Bloom (1984) found, a recent meta-review by VanLehn (2011) found that human tutoring had a positive impact of $d = 0.79$ on students' learning.

ITS track students' subject domain knowledge, learning skills, learning strategies, emotions, or motivation in a process called *student modeling* at a level of fine-grained detail that human tutors cannot (Graesser et al., 2011). ITS can also be distinguished from computer-based training, computer-assisted instruction (CAI), and e-learning. Specifically, given their enhanced adaptability and power of computerized learning environments, ITS are considered superior to computer-based training and CAI in that ITS allow an infinite number of possible interactions between the systems and the learners (Graesser et al., 2011). VanLehn (2006) described ITS as tutoring systems that have both an outer loop and an inner loop. The outer loop selects learning tasks; it may do so in an adaptive manner (i.e., select different problem sequences for different students), on the basis of the system's assessment of each individual student's strengths and weaknesses with respect to the targeted learning objectives. The inner loop elicits steps within each task (e.g., problem-solving steps) and provides guidance with respect to

This article was published Online First September 9, 2013.

Saiying Steenbergen-Hu and Harris Cooper, Department of Psychology & Neuroscience, Duke University.

Correspondence concerning this article should be addressed to Saiying Steenbergen-Hu, Department of Psychology & Neuroscience, Duke University, 417 Chapel Drive, Box 90086, Durham, NC 27708-0086. E-mail: ss346@duke.edu

these steps, typically in the form of feedback, hints, and error messages. In this regard, as VanLehn (2006) noted, ITS are different from CAI, computer-based training, or web-based homework in that the later lack of an inner loop. ITS are one type of e-learning that can be self-paced or instructor directed, encompassing all forms of teaching and learning that are electronically supported, through the Internet or not, in the form of texts, images, animations, audios, or videos.

The growth of ITS and the accumulation of evaluation research justify a meta-analysis of the effectiveness of ITS on students' mathematical learning for the following three reasons. First, several reviews of the impact of ITS on reading already exist (Becker, 1992; Blok, Oostedam, Otter, & Overmaat, 2002; Kulik, 2003). Most recently, Cheung and Slavin (2012) reviewed the effects of educational technology on K–12 students' reading achievement, relative to traditional instructional methods. They found an average standardized mean difference of 0.16 favoring the educational technology. No similar review regarding ITS with a focus on math has been carried out.

Second, much research on the effectiveness of math ITS has accumulated over the last two decades. Without rigorous summarization, this literature appears confusing in its findings. For example, Koedinger et al. (1997) found that students tutored by Cognitive Tutor showed extremely high learning gains in algebra compared with students who learned algebra through regular classroom instruction. Shneyderman (2001) found that, on average, students who learned algebra through Cognitive Tutor scored 0.22 standard deviations above their peers who learned algebra in traditional classrooms but only scored 0.02 standard deviations better than their comparison peers on the statewide standardized test. However, Campuzano, Dynarski, Agodini, and Rall (2009) found that sixth grade students who were taught math with regular classroom instruction throughout a school year outperformed those who were in regular class 60% of the school year and spent the other 40% of class time learning math with ITS, indicated by an effect size of −0.15. Thus, there is a need to gather, summarize, and integrate the empirical research on math ITS, to quantify their effectiveness and to search for influences on their impact.

Third, there has been increased attention in recent years on the effectiveness of math ITS for students' learning. The What Works Clearinghouse (WWC) has completed several evidence reviews on some math ITS products. For example, the WWC produced four reviews on Carnegie Learning's Cognitive Tutor (i.e., the WWC, 2004, 2007, 2009, 2010a). The WWC also reviewed the evidence on Plato Achieve Now (see WWC, 2010b). The WWC reviews, however, did not include all math ITS. Our literature search identified more than a dozen intelligent tutoring system products designed to help students' mathematical learning. And, important for our effort, the WWC reviews did not examine factors that might influence the direction and magnitude of the ITS effect. In contrast, our effort does not focus on specific intelligent tutoring system programs but on their general effectiveness and on the factors that moderate their impact.

In sum then, a number of questions regarding the effectiveness of ITS can be addressed by a meta-analysis. Most broadly, it can estimate the overall average effectiveness of ITS relative to other types of instruction on students' mathematical learning. But more specific questions also may be answerable. For example, a meta-analysis can explore what kind of settings ITS work best in, as well as for what types of student populations. By using information across as well as within primary studies, a meta-analysis provides a useful quantitative strategy for answering these questions.

Method

Study Inclusion and Exclusion Criteria

For studies to be included in this meta-analysis, the following eight criteria had to be met:

1. Studies had to be empirical investigations of the effects of ITS on learning of mathematical subjects. Secondary data analyses and literature reviews were excluded.

2. Studies had to be published or reported during the period from January 1, 1990, to June 30, 2011, and had to be available in English.

3. Studies had to focus on students in grades K–12, including high achievers, low achievers, and remedial students. However, studies focusing exclusively on students with learning disabilities or social or emotional disorders (e.g., students with attention-deficit/hyperactivity disorder) were excluded.

4. Studies had to measure the effectiveness of ITS on at least one learning outcome. Common measurements included standardized test scores, modified standardized test scores, course grades, or scores on tests developed by researchers.

5. Studies had to have used an independent comparison group. Comparison conditions included regular classroom instruction, human tutoring, or homework. Studies without a comparison group or those with one-group pretest–posttest designs were excluded.

6. Studies had to use randomized experimental or quasi-experimental designs. If a quasi-experimental design was used, evidence had to be provided that the treatment and comparison groups were equivalent at baseline (see WWC, 2008). Studies with a significant difference between the treatment and comparison groups prior to the ITS intervention were excluded, unless information was available for us to calculate effect sizes that would take into account the prior difference.

7. Studies had to have at least eight subjects in treatment and comparison groups, respectively. Studies with sample sizes less than eight in either group were excluded.

8. Studies had to provide the necessary quantitative information for the calculation or estimation of effect sizes.

Study Search

We used the following procedures to locate studies: (a) a search of abstracts in electronic databases including ERIC, PsycINFO, Proquest Dissertation and Theses, Academic Search Premier,

Econlit With Full Text, PsycARTICLES, SocINDEX With Full Text, and Science Reference Center; (b) Web searches using the Google and Google Scholar search engines; (c) a manual examination of reference and bibliography lists of the relevant studies; and (d) personal communications with 18 ITS research experts who had been the first author on two or more ITS studies during the past 20 years.

We used a wide variety of search terms to ensure our searches would identify as many relevant studies as possible. Although some researchers have used the term *intelligent tutoring systems*, many others have used a wide variety of alternative terms, for example, *computer-assisted tutoring, computer-based tutoring, artificial tutoring,* or *intelligent learning environments.* Therefore, we also used the terms *intelligent tutor*, artificial tutor*, computer tutor*, computer-assisted tutor*, computer-based tutor*, intelligent learning environment*, computer coach*, online-tutor*, keyboard tutor*, e-tutor*, electronical tutor*,* and *web-based tutor**. After concluding these searches, we began to focus on math ITS.

We found that some math ITS studies could not be retrieved through the search keywords above and some ITS studies are locatable only through the use of particular ITS names. The reference list of Graesser et al.'s (2011) introduction to ITS, for instance, indicates that large numbers of studies are exclusively connected with particular ITS programs, such as Cognitive Tutor, AutoTutor, or CATO. Dynarski et al. (2007) evaluated the effectiveness of three mathematical educational software programs for sixth graders (i.e., Larson Pre-Algebra, Achieve Now, and iLearn Math) and three software programs for ninth graders (i.e., Cognitive Tutor Algebra, Plato Algebra, and Larson Algebra). All of these educational software programs were actually ITS products. However, we found our previous search only caught studies of Cognitive Tutor, the most widely used and studied ITS, and we missed all studies of the other software. This was also the case for the educational software evaluated by Campuzano et al. (2009). Therefore, we used the names of some major software programs reported in Graesser et al. (2011), Dynarski et al. (2007), and Campuzano et al. (2009) and conducted a third search in ERIC and PsycINFO. No new qualified studies were found. However, by screening all of the studies in the WWC reviews of Cognitive Tutor and Plato Achieve Now (i.e., WWC, 2004, 2007, 2009, 2010a, 2010b), we found five additional studies that qualified for inclusion. In summary, our search concluded with 26 qualified reports evaluating the effectiveness of ITS on K–12 students' mathematical learning.

Study Coding

We designed a detailed coding protocol to guide the study coding and information retrieval. The coding protocol covered studies' major characteristics, which included (a) the basic features of the study reports (e.g., whether the study was published or unpublished and when it was conducted), (b) research design features (e.g., sample sizes; whether the study used a randomized or quasi-experimental design), (c) the contexts of intervention (e.g., subject matter; whether the study compared ITS with regular classroom instruction, human tutors, or other education interventions; the duration of ITS intervention), and (d) the study outcomes (e.g., what and how outcomes were measured; when the assessments took place; the magnitudes and direction of the effect sizes). Two coders independently coded the major features of each study, except the study outcomes, and then met together to check the accuracy of the coding. If there was a disagreement in coding, the two coders discussed and reexamined the studies to settle on the most appropriate coding. If the disagreement could not be resolved, the second author was consulted. The first author coded the study outcomes and then discussed the codes with the second author. The major specific variables coded are described later along with the study results.

Effect Size Calculation

We used Hedges' *g*, a standardized mean difference between two groups, as the effect size index for this meta-analysis. The preference for Hedges' *g* over other standardized-difference indices, such as Cohen's *d* and Glass's Δ, is due to the fact that Hedges' *g* can be corrected to reduce the bias that may arise when the sample size is small (i.e., $n < 40$; Glass, McGaw, & Smith, 1981). Hedges' *g* was chosen for this meta-analysis because the samples in many ITS studies are small.

Hedges' *g* was calculated by subtracting the mean of the comparison condition from that of the ITS tutoring condition and dividing the difference by the average of the two groups' standard deviations. A positive *g* indicates that students tutored by ITS achieved more learning gains than did those in the comparison condition. In cases for which only inference test results were reported but no means and standard deviations were available, *g* was estimated from the inferential statistics, such as *t*, *F*, or *p* values (Wilson & Lipsey, 2001). For studies that did not report specific values of inferential statistics, we assumed a conservative value for effect size calculation. For example, if a study reported a statistically significant difference between the ITS and the comparison condition with $p < .01$, we assumed a *p* value of .01 for effect size calculation.[1]

We calculated unadjusted effect sizes for a study if it only reported the ITS and comparison groups' mean posttest scores, standard deviation, and sample sizes. Unadjusted effect sizes did not take into account other variables that might have had an impact on the outcomes. For some studies, in addition to unadjusted effect sizes, adjusted effect sizes were also extracted. We called them *adjusted effect sizes* because they were calculated after adjusting or controlling for other variables, such as pretest scores. In some cases, adjusted effect sizes were based on means and standard deviations of gain scores (i.e., posttests − pretests), whereas in other cases they were based on covariance-adjusted means and standard deviations. For studies that reported descriptive statistics of both pretests and posttests, as suggested by D. B. Wilson (personal communication, April 18, 2011), adjusted effect sizes were the differences between posttest and pretest effect sizes and their variances were the sum of posttest and pretest effect sizes variances.

[1] This was the case for only one study (i.e., Shneyderman, 2001), in which the effect size for one of the three outcomes was calculated by assumed $p = .01$ when the study reported $p < .01$. Because the effect size representing this study was an average of all three effect sizes from three outcomes, there was a minimal possibility that this would lead to an underestimation of the overall effect sizes in this meta-analysis.

Independent Studies, Samples, and Effect Sizes

To address effect size dependency issues, we used independent samples as the unit of analysis. Each independent sample is not the equivalent of a separate research report. One report could contain two or more independent studies. For example, we coded Beal et al. (2010) as two independent studies, each based on a different sample. The 26 reports contained 34 independent studies based on 34 independent samples. Table 1 presents the major features of all 31 studies in which ITS were compared with regular classroom instruction.

We used a shifting unit of analysis approach (Cooper, 2010) to further address possible dependencies among effect sizes. The benefits of the shifting unit of analysis approach are that it allows us to retain as much data as possible while ensuring a reasonable degree of independence among effect sizes. With this approach, effect sizes were first extracted for each outcome as if they were independent. For example, if a study with one independent sample used both a standardized test and a course grade to measure students' learning, two separate effect sizes were calculated. When estimating the overall average effect of ITS, these two effect sizes were averaged so that the sample contributed only one effect size to the analysis. However, when conducting a moderator analysis to investigate whether the effects of ITS vary as a function of the type of outcome measures, this sample contributed one effect size to the category of standardized test and one to that of course grade.

Data Analysis

We used the Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2006) software for data analysis. Before the analyses, we conducted Grubbs (1950) tests to examine whether there were statistical outliers among the effect sizes and sample sizes. We conducted the meta-analysis using a weighting procedure and with both fixed-effect and random-effects models (Cooper, 2010). A fixed-effect model functions with the assumption that there is one true effect in all of the studies included in a meta-analysis and the average effect size will be an estimate of that value. A fixed-effect model is suited to drawing conditional inferences about the observed studies. However, it is less well suited to making generalizations to the population of studies from which the observed studies are a sample (Konstantopoulos & Hedges, 2009). A random-effects model assumes that there is more than one true effect and the effect sizes included in a meta-analysis are drawn from a population of effects that can differ from each other.

Two approaches were used to assess publication bias. First, a funnel plot was visually inspected. The suggestion of missing studies on the left side of the distribution indicated the possible presence of publication bias. Duval and Tweedie's (2000) trim-and-fill procedure (available as part of the Comprehensive Meta-Analysis software) was then used to further assess and adjust for publication bias. Through this procedure, unmatched observations were removed (trimmed) from the data distribution and additional values were imputed (filled) for projected missing studies. Then, average effect sizes are recalculated.

Moderator Analyses

Testing for moderators was conducted on the groups of effect sizes that had a high degree of heterogeneity (Cooper, Hedges, & Valentine, 2009). The purpose of testing for moderators was to identify variables associated with certain features of the primary studies that might be significantly associated with the effectiveness of ITS.

Results

The literature search located 26 reports that met our study inclusion criteria. The reports appeared between 1997 and 2010. The sample sizes in the reported studies ranged from 18 to 17,164. The 26 reports provided 65 effect sizes. Forty-seven effect sizes were unadjusted, meaning they were calculated from posttest outcome measures and did not control for variables other than the ITS treatment, which might have influenced the outcome measures; 18 were adjusted effect sizes, which were calculated after controlling for other confounding variables, such as pretest scores.

As mentioned in the Method section, to address effect size dependency issues, we used independent studies (i.e., samples) as the unit of analysis. The 26 reports contained 34 independent studies based on 34 independent samples. Of the 34 independent studies, 31compared ITS with regular classroom instruction. These 31 studies provided 61 effect sizes (see Table 1). In general, these 31 independent studies compared learning outcomes of instructions with an ITS component to those without one. Specifically, this comparison refers to four types of comparison situations. First, a large portion of the studies, for example, studies of Cognitive Tutor compared the learning gains of students who learned through instruction in which Cognitive Tutor was a significant part of regular classroom instruction with the learning gains of students who learned through traditional classroom instruction in which no Cognitive Tutor was involved. In such studies, interventions usually lasted for one school year or one semester during which students in the experimental groups generally spent 60% of their time in regular classroom learning and 40% of their time in the computer lab using Cognitive Tutor; students in the control groups spent 100% of their time in regular classrooms. Second, some studies compared students who learned solely through using ITS with those who learned in regular classroom instruction (e.g., Arroyo, Woolf, Royer, Tai, & English, 2010; Beal et al., 2010, Study 2; Walles, 2005). Interventions in these studies usually lasted for just a few days. Third, two studies compared students' learning in conditions in which ITS partially took teacher's responsibilities (e.g., giving students guidance or feedback) to students' learning in conditions in which they received guidance or feedback from teachers (i.e., Hwang, Tseng, & Hwang, 2008; Stankov, Rosic, Zitko, & Grubisic, 2008). Interventions in these studies lasted for several weeks to one semester. Last, one study compared students who used ITS as a supplement in addition to regular classroom instruction with students who learned through regular classroom instruction without using ITS as a supplement (i.e., Biesinger & Crippen, 2008). Intervention in this study lasted for one semester. Because the comparison conditions in all four types of situations above involved either regular classroom instruction or teachers' efforts, we grouped them together as ITS being compared with regular classroom instruction.

Two independent studies (i.e., Mendicino, Razzaq, & Heffernan, 2009; Radwan, 1997) provided information on the effects of ITS on mathematical learning relative to that of homework assignments. One independent study (i.e., Beal et al., 2010) compared

Table 1
Studies Comparing Intelligent Tutoring Systems With Regular Classroom Instruction

Study (independent sample)	ITS name	Subject	ITS duration	Sample size[a]	Sample achievement level	Schooling level	Research design	Year of data collection	Counter-balanced testing	Report type	Unadjusted ES[b]	Adjusted ES[c]
Arbuckle (2005)	Cognitive Tutor Algebra I	Algebra	Short term	111	General	High	Quasi-experimental	ng	No	Nonjournal	0.78	0.67
Arroyo et al. (2010)	Math Facts Retrieval Training and Wayang Outpost	Basic math	Short term	250	General	Middle	Quasi-experimental	2006–2010	Yes	Journal	0.39	
Beal et al. (2010) (2)	AnimalWatch	Basic math	Short term	202	General	Middle	Quasi-experimental	ng	ng	Journal	−0.26	
Beal et al. (2007)	Wayang Outpost	Basic math	Short term	28	General	High	Quasi-experimental	ng	Yes	Journal		0.55
Biesinger & Crippen (2008) (1)	Online Remediation software	Basic math	Semester	3,566	Low achievers	High	Quasi-experimental	2003–2005	No	Journal	0.22	0.13
Biesinger & Crippen (2008) (2)	Online Remediation software	Basic math	Semester	17,164	General	High	Quasi-experimental	2003–2005	No	Journal	0.16	
Cabalo & Vu (2007)	Cognitive Tutor Algebra I	Algebra	One semester	364	General	Ng	True experimental	2006–2010	No	Nonjournal	−0.22	0.03
Cabalo, Ma, & Jaciw (2007)	Cognitive Tutor Bridge to Algebra Curriculum	Algebra	One school year	576	General	Ng	True experimental	2006–2010	No	Nonjournal	0.09	0.05
Campuzano et al. (2009) (1)	Larson Pre-Algebra and Achieve Now	Basic math	One school year	659	General	Middle	True experimental	2006–2010	Yes	Nonjournal	−0.23	−0.12
Campuzano et al. (2009) (2)	Cognitive Tutor Algebra I and Larson Algebra I	Algebra	One school year	534	General	High	True experimental	2006–2010	Yes	Nonjournal	0.09	0.09
Carnegie Learning (2001a)	Cognitive Tutor Algebra I	Algebra	One school year	293	General	High	True experimental	Before 2003	No	Nonjournal	0.45	
Carnegie Learning (2001b) (1)	Cognitive Tutor Math	Basic math	One school year	132	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal		0.00
Carnegie Learning (2001b) (2)	Cognitive Tutor Math	Basic math	One school year	174	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal		−0.23
Carnegie Learning (2002) (1)	Cognitive Tutor	Algebra	One school year	58	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal	−0.66	
Carnegie Learning (2002) (2)	Cognitive Tutor	Algebra	One school year	80	General	High	Quasi-experimental	Before 2003	No	Nonjournal	−0.24	
Dynarski et al. (2007) (1)	Larson Pre-Algebra, Achieve Now, and iLearn Math	Basic math	One school year	3,136	General	Middle	True experimental	2003–2005	Yes	Nonjournal	0.05	0.14
Dynarski et al. (2007) (2)	Cognitive Tutor Algebra, Plato Algebra, and Larson Algebra	Algebra	One school year	1,404	General	High	True experimental	2003–2005	Yes	Nonjournal	−0.23	−0.07

Table 1 (*continued*)

Study (independent sample)	ITS name	Subject	ITS duration	Sample size[a]	Sample achievement level	Schooling level	Research design	Year of data collection	Counter-balanced testing	Report type	Unadjusted ES[b]	Adjusted ES[c]
Hwang, Tseng, & Hwang (2008)	Intelligent Tutoring, Evaluation and Diagnosis	Basic math	Semester	76	General	ng	True experimental	ng	No	Journal	0.75	
Koedinger (2002)	Cognitive Tutor Math 6	Basic math	One school year	128	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal	0.42	
Koedinger et al. (1997)	Cognitive Tutor Algebra—PUMP	Algebra	One school year	225	General	High	Quasi-experimental	Before 2003	No	Journal	0.49	
Morgan & Ritter (2002)	Cognitive Tutor Algebra I	Algebra	One school year	384	General	High	True experimental	Before 2003	No	Nonjournal	0.23	
Pane et al. (2010)	Cognitive Tutor Geometry	Geometry	One school year	699	General	High	True experimental	2006–2010	No	Journal		−0.19
Plano, Ramey, & Achilles (2007)	Cognitive Tutor Algebra	Algebra	One school year	779	Low achievers	High	Quasi-experimental	2003–2005	No	Nonjournal	−0.66[d]	−0.48
Ritter et al. (2007)	Cognitive Tutor Algebra I	Algebra	One school year[e]	342	General	High	True experimental	ng	No	Journal	0.29	
Sarkis (2004)	Cognitive Tutor Algebra I	Algebra	One school year	4,649	General	High	Quasi-experimental	2003–2005	No	Nonjournal	0.13	
Shneyderman (2001)	Cognitive Tutor Algebra I	Algebra	One school year	663	General	High	Quasi-experimental	Before 2003	No	Nonjournal	0.14	
Smith (2001)	Cognitive Tutor Algebra	Algebra	More than one school year	445	Low achievers	High	True experimental	Before 2003	No	Nonjournal		−0.07
Stankov et al. (2008) (1)	Tutor-Expert System	Basic math	Short term	18	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.92	1.05
Stankov et al. (2008) (2)	Tutor-Expert System	Basic math	Short term	18	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.08	0.11
Stankov et al. (2008) (3)	Tutor-Expert System	Basic math	Short term	48	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.00	0.31
Walles (2005)	Wayang Outpost	Basic math	Short term	218	General	High	True experimental	2003–2005	No	Nonjournal	−0.25	

Note. ITS = intelligent tutoring system; ES = effect size; ng = not given.
[a] The sample sizes reported in this table are the total sample sizes of each independent study. Grubbs (1950) tests showed that among the total sample sizes, five of them were detected as outliers. They are 17,164, 4,649, 3,566, 3,136, and 1,404, for which the nearest neighbor is 799. [b] These are unadjusted overall effect sizes for each independent sample. [c] These are adjusted overall effect sizes for each independent sample. [d] The original effect size extracted from this study was −1.57. It was detected as an outlier in Grubbs (1950) tests. In the analyses, we reset the value to −0.66, its next nearest neighbor among unadjusted overall effect sizes. [e] The Ritter et al. (2007) study reported an outcome measure after one semester of the intervention, and it also reported two outcome measures after one school year of the intervention. The study sample remained same. So we pooled the outcome measures and extracted one overall effect size for this study. However, we categorized the ITS duration as one school year.

ITS with human tutoring. We narratively reported the results of the studies that compared ITS with human tutors or home work conditions later in this section. We did not include them in the analyses described below so as to have a single clear comparison group. Therefore, the 61 effect sizes of ITS in comparison to regular classroom instruction made up the data for the results that follow.

With the 61 effect sizes, we formed three different data sets. The first data set included unadjusted overall effect sizes. It consisted of 26 effect sizes with each independent sample contributing one overall effect size to the data set. Here, if multiple effect sizes were extracted from the same sample, these effect sizes were averaged to estimate the overall effectiveness of ITS on this independent sample. The second data set included all unadjusted effect sizes. This data set consisted of all 44 unadjusted effect sizes from the 26 independent samples. The third data set consisted of 17 adjusted overall effect sizes from 17 independent samples. Some independent samples provided both an unadjusted and an adjusted overall effect size, whereas some only provided one type of overall effect sizes or the other.

We conducted analyses on adjusted and unadjusted effect sizes separately. One may argue that it would be beneficial to pool the two types of effect sizes so that the analyses would include all of the 31 effect sizes from the 31 studies. However, we think the benefits of conducting the analyses separately outweigh those of analyzing them together. We have three justifications. First, distinguishing adjusted and unadjusted effect sizes would allow us to examine whether estimates of ITS effectiveness differs with or without controlling for confounding factors. Second, the number of studies in the analyses did not increase significantly even if we analyze the effect sizes together. Specifically, if the effect sizes were analyzed together, the total number of studies would be increased from 26 (i.e., the number of unadjusted effect sizes) to 31 (i.e., the total number of independent studies or samples). Finally, analyzing the two types of effect sizes separately helps in interpretation. Differentiating adjusted and unadjusted effect sizes and integrating them separately allows us to provide clearer information regarding what each estimate of effect refers to with regard to the achievement outcome.

We conducted Grubbs (1950) tests to look for statistical outliers before calculating the average effect sizes. The Grubbs tests showed that, among the unadjusted overall effect sizes ($k = 26$), one effect size ($g = -1.57$) appeared to be an outlier (i.e., Plano, Ramey, & Achilles, 2007). We found that the Plano et al. (2007) study provided information for both an adjusted (adjusted by pretest scores, $g = -0.48$) and an unadjusted effect size ($g = -1.57$). Clearly then, the unadjusted effect size was strongly impacted by the preexisting differences between the treatment and comparison groups. We reset the effect size to -0.66, its next nearest neighbor among the unadjusted overall effect sizes. Among all the unadjusted effects sizes ($k = 44$), the effect size ($g = -1.57$) from the Plano et al. (2007) study again appeared to be an outlier. We reset the effect size to -1.03, its next nearest neighbor. The Grubbs tests detected no outliers among the adjusted overall effect sizes. We also conducted Grubbs tests on ITS and comparison group sample sizes. Again, we reset the outlier sample sizes to their nearest neighbors. We conducted analyses after adjusting the outlier sample sizes.[2]

As Table 1 shows, 10 different ITS were studied in the 31 independent studies comparing ITS with regular classroom instruction. Cognitive Tutor by Carnegie Learning was the most frequently studied. Specifically, Cognitive Tutor for algebra learning was evaluated in 16 studies; Cognitive Tutor for math was studied in three studies; in one study, Cognitive Tutor was used for geometry. As mentioned in the introduction, the WWC had completed four reviews on the effectiveness of Cognitive Tutor. Four other ITS (i.e., Larson Pre-Algebra/Algebra I, Achieve Now, iLearn Math, and Plato Algebra) were evaluated in a national-level study (see Campuzano et al., 2009; Dynarski et al., 2007) and were also reviewed by the WWC. The other two ITS that were relatively frequently studied were Wayang Outpost (see Arroyo et al., 2010; Beal et al., 2010; Walles, 2005) and Tutor-Expert System (see the three studies by Stankov, Rosic, Zitko, & Grubisic, 2008). Online Remediation Software appeared in two studies by Biesinger and Crippen (2008). AnimalWatch (Beal, Arroyo, Cohen, &Woolf, 2007) and Intelligent Tutoring, Evaluation and Diagnosis (Hwang et al., 2008) each appeared once.

Because Cognitive Tutor was most frequently studied, we briefly describe its mechanism, scope of use, and the length of its implementation. Cognitive Tutor is built on a cognitive theory called adaptive control of thought (Anderson et al., 1995). Cognitive Tutor presents students with a series of problems and adaptively identifies a student's problem-solving strategy through his or her actions and comparisons with correct solution approaches and misconceptions generated by the program's cognitive model, a process called *model tracing*. Five curricula have been developed with Cognitive Tutor as their software component and have been used by more than 500,000 students in approximately 2,600 schools across the United States as of 2010 (WWC, 2010a). They are Bridge to Algebra, Algebra I, Geometry, Algebra II, and Integrated Math. In these curricula, students generally spend three class periods per week in regular classroom learning and two class periods in computer lab using Cognitive Tutor. In most of the evaluation studies included in this meta-analysis, students used Cognitive Tutor for one school year or one semester.

Overall Effectiveness of ITS on Students' Mathematical Learning

We conducted meta-analyses on the data sets of unadjusted and adjusted overall effect sizes to examine the overall effectiveness of ITS on students' mathematical learning, compared with that of regular classroom instruction. All effect sizes were weighted by inverse variances. Of the 26 unadjusted overall effect sizes, 17 were in a positive direction, eight were in a negative direction, and one was exactly 0. Under a fixed-effect model, the average effect size was 0.05, 95% CI [.02, .09], $p = .005$, and was significantly different from 0. Under a random-effects model, the average effect size was 0.09, 95% CI [−.03, .20], $p = .136$, and was not significantly from 0. There was a high degree of heterogeneity

[2] It is worth mentioning that we also calculated the average effect sizes and ran moderator analyses on the effect sizes without adjusting the outlier sample sizes. We found very minor differences in the analysis results between those with and without adjusted outlier sample sizes. These differences were not sufficient to lead to any major changes in conclusions. Therefore, we choose to only report the analysis results with the sample size outliers adjusted.

among the 26 unadjusted overall effect sizes, $Q_t(25) = 180.80$, $p = .000$. This indicates that it was unlikely that sampling error alone was responsible for the variance among the effect sizes; instead, some other factors likely played a role in creating variability as well.

Of the 17 adjusted overall effect sizes, 10 were in a positive direction and seven were in a negative direction. Under a fixed-effect model, the average effect size was 0.01, 95% CI [−.04, .06], $p = .792$, and was not significantly different from 0. Under a random-effects model, the average effect size also was 0.01, 95% CI [−.10, .12], $p = .829$, and was also not statistically significantly different from 0. There was a high degree of heterogeneity among the 17 adjusted overall effect sizes, $Q_t(16) = 54.01$, $p = .000$. Again, this suggested that it was unlikely that sampling error alone was responsible for the total variance among the effect sizes.

Examining Publication Bias

We conducted Duval and Tweedie's (2000) trim and fill procedure to assess the possible effects of publication bias. For unadjusted overall effect sizes, there was evidence that three studies on the left side of the distribution might have been missing under both a fixed-effect model and a random-effects model. The overall average effect size after imputing the three additional values was 0.04 under a fixed-effect model and was 0.03 under a random-effects model. The average effect size for the observed effect sizes, as reported previously, was 0.05 under a fixed-effect model and 0.09 under a random-effects model. This implies that the average effect of ITS might have been slightly overestimated.

For adjusted overall effect sizes, three studies on the left side of the distribution were projected as missing under a fixed-effect model, and six effect sizes on the left side of the mean were projected as missing under a random-effects model. The overall average effect sizes after imputing the three additional values ranged from −0.04 to −0.01 using a fixed-effect model; the overall average effect size after imputing the six additional values was −0.09 using a random-effects model. The average of the observed effect sizes, as reported previously, were 0.01 under both a fixed-effect model and random-effects model. Therefore, there was little evidence that publication bias had much impact on the average effect size in this case.

Testing for Moderators on the Unadjusted and Adjusted Overall Effect Sizes

We conducted moderator analyses exploring nine variables that could possibly have an impact on ITS's effects. We chose these nine variables for two reasons. First, they represented important features of ITS intervention or research methodology. Second, there were at least two effect sizes associated with each of the category of the variable, in the data sets of both unadjusted and adjusted overall effect sizes, to allow meaningful analyses.[3]

Tables 2 and 3 present the results of testing for moderators on the unadjusted and adjusted overall effect sizes, respectively. In each data set, the number of effect sizes involved might be different. For variables with more than two categories, we first conducted comparisons using all of the categories and then regrouped them to create a two-group comparison. We included the results of further analyses on the two-group comparison in the tables as well.

Subject matter. Testing results showed that the effectiveness of ITS did not differ for different subject matters under a fixed-effect model, $Q_b(1) = .12$, $p = .726$, nor did it differ under a random-effects model, $Q_b(1) = .62$, $p = .431$, for unadjusted effect sizes. The advantage of using ITS, compared with regular classroom instruction, was significant only for basic math under the fixed-effect model, indicated by the fact that the confidence interval of the effect size ($g = .06$) did not contain 0.

For adjusted effect sizes, results showed the effectiveness of ITS on students' learning of basic math appeared to be greater than that of learning algebra under a fixed-effect model, $Q_b(1) = 9.10$, $p = .003$, but not under a random-effects model, $Q_b(1) = 1.62$, $p = .204$.[4] Specifically, under a fixed-effect model, the average effectiveness of ITS was $g = 0.11$, 95% CI [.04, .19], on helping students learn basic math and $g = -0.05$, 95% CI [−.13, .02], on learning algebra.

ITS duration. For unadjusted effect sizes, the effectiveness of ITS differed depending on the length of instruction under both a fixed-effect model, $Q_b(2) = 16.28$, $p = .000$, and a random-effects model, $Q_b(2) = 6.42$, $p = .04$. Further analyses revealed no difference between the short-term and one-semester ITS interventions. We therefore compared the combination of short-term and one-semester interventions with interventions of one school year or longer. We found that under a fixed-effect model, the average effectiveness of ITS was greater when the interventions lasted for less than one school year, $g = 0.23$, 95% CI [.13, .32], than when they lasted for one school year or longer, $g = .02$, 95% CI [−.02, .06]; under a random-effects model, the average effectiveness of ITS was also greater when the interventions lasted for less than one school year, $g = 0.26$, 95% CI [.08, .44], than that of when they lasted for one school year or longer, $g = -.01$, 95% CI [−.15, .14].

For adjusted effect sizes, results showed that ITS effectiveness also differed depending on the duration of intervention under both a fixed-effect model, $Q_b(2) = 13.88$, $p = .001$, and a random-effects model, $Q_b(2) = 14.71$, $p = .001$. Further analyses revealed that the effects differed depending on whether the ITS intervention lasted for one school (or longer) or less than one school year under both a fixed-effect model, $Q_b(1) = 6.48$, $p = .011$, and a random-effects model, $Q_b(1) = 7.40$, $p = .007$.

Sample achievement level. We categorized study samples in terms of the academic achievement level of the subjects, using the way they were reported in the primary studies to categorize samples. Two types of student samples appeared. One consists of general students, a population that includes students of all achievement levels. Another consists of low achievers. There were three studies that reported results for low achievers (i.e., Biesinger & Crippen, 2008; Plano et al., 2007; Smith, 2001). For unadjusted effect sizes, under a fixed-effect model, the effectiveness of ITS on

[3] The second reason led us to drop a number of variables we initially intended to study. For example, we hoped to compare whether there was a difference in the effects of ITS when they were used to substitute for regular classroom instruction and when they were used only as a supplement to regular classroom instruction. We were unable to do so because, for the 17 adjusted effect sizes, only one effect size was associated with ITS used as substitute, versus 16 effect sizes associated with ITS as a supplement.

[4] For adjusted overall effect sizes, we dropped one effect size associated with geometry, $g = -.19$, 95% CI [−.34, −.04]

Table 2
Testing for Moderators of the Unadjusted Effect Sizes

Variable	*k*	Fixed				Random			
		g	95% CI	Q_b	p_b	*g*	95% CI	Q_b	p_b
Subject				.12	.726			0.62	.431
Basic math	12	.06	[.01, .09]			.14	[−.03, .32]		
Algebra	14	.05	[−.01, .14]			.05	[−.11, .21]		
ITS duration				16.28***	.000			6.42*	.040
One school year or longer	15	.02	[−.02, .06]			−.01	[−.15, .14]		
One semester	4	.24	[.13, .36]			.28	[.10, .45]		
Short term	7	.20	[.04, .36]			.23	[−.13, .58]		
ITS duration (further analysis)				16.07***	.000			5.10*	.024
One school year or longer	15	.02	[−.02, .06]			−.01	[−.15, .14]		
Less than one school year	11	.23	[.13, .32]			.26	[.08, .44]		
Sample achievement level				46.13***	.000			.60	.438
General students	24	.09	[.05, .12]			.12	[.02, .21]		
Low achievers	2	−.42	[−.55,−.28]			−.23	[−1.08, .63]		
Schooling level				2.11	.349			.58	.749
Elementary school	3	.21	[−.22, .62]			.25	[−.28, .77]		
Middle school	6	−.001	[−.09, .09]			.02	[−.24, .29]		
High school	14	.06	[.02, .11]			.09	[−.07, .25]		
Schooling level (further analysis)				.51	.473			.37	.545
Elementary school	3	.21	[−.22, .62]			.25	[−.28, .77]		
Secondary school	23	.05	[.01, .09]			.08	[−.04, .20]		
Sample size				14.28**	.001			.70	.704
Less than 200	9	.27	[.09, .45]			.21	[−.14, .56]		
Over 200 but less than 1,000	12	−.02	[−.08, .03]			.05	[−.14, .25]		
Over than 1,000	5	.09	[.04, .13]			.06	[−.09, .20]		
Sample size (further analysis)				5.94*	.015			.74	.389
Less than 200	9	.27	[.09, .45]			.21	[−.14, .56]		
Over 200	17	.04	[.01, .08]			.05	[−.08, .17]		
Research design				7.97**	.005			.37	.544
Quasi-experimental	15	.09	[.05, .14]			.12	[−.07, .31]		
True experimental	11	−.01	[−.07, .05]			.05	[−.09, .19]		
Year of data collection				10.51**	.005			3.01	.222
Before 2003	7	.20	[.09, .30]			.18	[−.02, .39]		
Between 2003–2005	7	.02	[−.02, .07]			−.08	[−.30, .14]		
Between 2006–2010	8	−.01	[−.09, .07]			.05	[−.14, .24]		
Year of data collection (further analysis)				1.69	.193			.02	.883
Before 2006	14	.05	[.01, .09]			.03	[−.13, .19]		
2006 and after	8	−.01	[−.09, .07]			.05	[−.14, .24]		
Counterbalanced testing				10.43**	.001			1.09	.297
No	20	.09	[.05, .13]			.13	[−.02, .27]		
Yes	5	−.05	[−.12, .02]			−.002	[−.20, .19]		
Report type				24.45***	.000			10.03**	.002
Peer-reviewed journal	10	.28	[.18, .37]			.30	[.17, .43]		
Nonjournal	16	.02	[−.02, .05]			−.01	[−.15, .13]		
Measurement timing[a]				18.81***	.000			5.71	.058
End of school year	18	.01	[−.03, .04]			.01	[−.14, .15]		
End of semester	3	.29	[.15, .43]			.34	[.11, .57]		
Immediately after intervention	6	.19	[.02, .35]			.16	[−.29, .60]		

Table 2 (*continued*)

Variable	*k*	Fixed				Random			
		g	95% CI	Q_b	p_b	*g*	95% CI	Q_b	p_b
Measurement timing (further analysis)				17.84***	.000			3.02	.082
End of school year	18	.01	[−.03, .04]			.01	[−.14, .15]		
Sooner than end of school year	9	.25	[.14, .35]			.26	[.01, .51]		
Outcome type[b]				16.57**	.002			5.36	.252
Course grades	3	.29	[.15, .43]			.33	[.10, .56]		
Course passing rates	2	.14	[−.01, .16]			.22	[−.12, .55]		
Specifically designed tests	11	.13	[−.02, .27]			.10	[−.24, .45]		
Modified standardized tests	6	.05	[−.05, .16]			.04	[−.21, .28]		
Standardized tests	14	.02	[−.02, .05]			.03	[−.13, .19]		
Outcome type (further analysis)				13.19***	.000			2.09	.148
Course-related outcome measures	16	.19	[.11, .27]			.20	[.01, .39]		
Standardized test measures	20	.02	[−.02, .06]			.03	[−.11, .16]		

Note. CI = confidence interval; Q_b denotes the heterogeneity status between all subcategories of a particular variable under testing, with degrees of freedom equal to moderator levels minus one; ITS = intelligent tutoring system.

[a] Testing for moderators was conducted on the all unadjusted effect sizes so that total number of *k* exceeded 26. [b] Testing for moderators was conducted on the all unadjusted effect sizes so that total number of *k* exceeded 26.

* $p < .05$. ** $p < .01$. *** $p < .001$.

helping general students learn mathematical subjects, $g = .09$, 95% CI [.05, .12], was greater than on helping low achievers, $g = -.42$, 95% CI [−.55, −.28], $Q_b(1) = 46.13$, $p = .000$. The difference was not significant under a random-effects model, $Q_b(1) = 0.60$, $p = .438$. Overall, ITS appeared to have a positive impact on general students. For low achievers, the average effect was negative under both fixed-effect and random-effects models.

For adjusted effect sizes, the effects of ITS on helping general students learn mathematical subjects, $g = .04$, 95% CI [−.02, .09], were greater than on helping low achievers, $g = -.18$, 95% CI [−.32, −.05], $Q_b(1) = 9.24$, $p = .002$, under a fixed-effect model. The difference was not significant under a random-effects model, $Q_b(1) = 1.31$, $p = .253$. In this analysis, the only average effect size that was significantly different from 0 was the negative effect indicating that regular classroom instruction compared favorably with ITS under a fixed-effect model.

Schooling level. The unadjusted overall effect sizes were associated with samples of three schooling levels: (a) elementary school, which included K–5 grade levels; (b) middle school, which included Grades 6–8; and (c) high school, which included Grades 9–12.[5] The relative effectiveness of ITS on students' mathematical learning did not vary significantly in terms of schooling level under either a fixed-effect model, $Q_b(2) = 2.11$, $p = .349$, or a random-effects model, $Q_b(2) = 0.58$, $p = .749$. We regrouped the effect sizes into elementary school and secondary school levels. Again, results showed that the difference was not significant under either a fixed-effect model, $Q_b(1) = 0.51$, $p = .473$, or a random-effects model, $Q_b(1) = 0.37$, $p = .545$.

For the adjusted overall effect sizes, the effectiveness of ITS varied significantly in terms of schooling level under a fixed-effect model, $Q_b(2) = 14.29$, $p = .001$, but not under a random-effects model, $Q_b(2) = 3.07$, $p = .215$. The average effect sizes suggested that the effects of ITS might be most pronounced for students in elementary school, $g = .41$, 95% CI [−.01, .84], compared with those in middle school, $g = .09$, 95% CI [.01, .17] and in high school, $g = -.09$, 95% CI [−.17, −.02]. However, when the effect sizes were regrouped into elementary and secondary school levels, no statistically significant difference was found between them under either a fixed-effect model, $Q_b(1) = 3.61$, $p = .057$, or a random-effects model, $Q_b(1) = 3.19$, $p = .074$.

Sample size. Among the 26 unadjusted overall effect sizes, nine were associated with sample sizes less than 200, 12 were associated with sample sizes over 200 but less than 1,000, and five were associated with sample sizes over 1,000. The unadjusted effectiveness of ITS corresponding to each of these three sample size categories varied significantly under a fixed-effect model, $Q_b(2) = 14.28$, $p = .001$, but not under a random-effects model, $Q_b(2) = 0.70$, $p = .704$. Further analyses revealed that the effects were greater when the study sample sizes were less than 200 than when the sample sizes were over 200 under a fixed-effect model, $Q_b(2) = 5.94$, $p = .015$, but not under a random-effects model, $Q_b(2) = 0.74$, $p = .389$.

[5] We did not include three unadjusted and two adjusted effect sizes associated with studies in which samples were across both middle school and high school. It is also worthy to note that there were only three studies involved elementary school students and all of them were conducted by a same research team.

Table 3
Testing for Moderators of the Adjusted Overall Effect Sizes

Variable	*k*	Fixed				Random			
		g	95% CI	Q_b	p_b	*g*	95% CI	Q_b	p_b
Subject				9.10**	.003			1.62	.204
Basic math	9	.11	[.04, .19]			.11	[−.05, .28]		
Algebra	7	−.05	[−.13, .02]			−.03	[−.19, .12]		
ITS duration				13.88**	.001			14.71**	.001
One school year or longer[a]	10	−.02	[−.07, .04]			−.08	[−.20, .05]		
One semester	2	.06	[−.13, .24]			.06	[−.13, .24]		
Short term	5	.52	[.24, .79]			.52	[.24, .79]		
ITS duration (further analysis)				6.48*	.011			7.40**	.007
One school year of longer	10	−.02	[−.07, .04]			−.08	[−.20, .05]		
Less than one school year	7	.19	[.04, .34]			.29	[.06, .53]		
Sample achievement level				9.24**	.002			1.31	.253
General students	14	.04	[−.02, .09]			.05	[−.06, .16]		
Low achievers	3	−.18	[−.32,−.05]			−.16	[−.49, .18]		
Schooling level				14.29**	.001			3.07	.215
Elementary school	3	.41	[−.01, .84]			.42	[−.04, .89]		
Middle school	4	.09	[.01, .17]			−.004	[−.20, .19]		
High school	8	−.09	[−.17,−.02]			−.01	[−.19, .16]		
Schooling level (further analysis)				3.61	.057			3.19	.074
Elementary school	3	.41	[−.01, .84]			.42	[−.04, .89]		
Secondary school	14	.001	[−.05, .05]			−.01	[−.13, .10]		
Sample size				12.49**	.002			2.77	.251
Less than 200	6	.18	[−.06, .43]			.24	[−.11, .59]		
Over 200 but less than 1000	8	−.08	[−.16, .01]			−.06	[−.20, .09]		
Over than 1,000	3	.09	[.01, .16]			.06	[−.10, .22]		
Sample size (further analysis)				2.03	.154			1.96	.162
Less than 200	6	.18	[−.06, .43]			.24	[−.11, .59]		
Over 200	11	−.001	[−.05, .05]			−.02	[−.14, .09]		
Research design				.92	.338			1.09	.296
Quasi-experimental	9	−.06	[−.20, .08]			.17	[−.16, .50]		
True experimental	8	.02	[−.04, .07]			−.01	[−.11, .08]		
Year of data collection				2.31	.315			.715	.699
Before 2003	3	−.08	[−.25, .08]			−.08	[−.25, .08]		
Between 2003–2005	4	.03	[−.04, .10]			−.07	[−.33, .19]		
Between 2006–2010	8	−.03	[−.12, .05]			.00	[−.13, .14]		
Year of data collection (further analysis)				.79	.375			.53	.468
Before 2006	7	.01	[−.05, .08]			−.08	[−.26, .10]		
2006 and after	8	−.03	[−.12, .05]			.004	[−.13, .14]		
Counterbalanced testing				8.89**	.003			.44	.507
No	12	−.08	[−.16,−.004]			−.01	[−.17, .14]		
Yes	5	.08	[.01, .14]			.06	[−.09, .21]		
Report type				.69	.407			1.98	.160
Peer-reviewed journal	6	−.04	[−.17, .08]			.23	[−.11, .57]		
Nonjournal	11	.02	[−.04, .07]			−.03	[−.15, .09]		

Note. CI = confidence interval; Q_b denotes the heterogeneity status between all subcategories of a particular variable under testing; ITS = intelligent tutoring system.
[a] This subcategory included one study in which the ITS intervention lasted more than one school year.
* $p < .05$. ** $p < .01$.

The adjusted effectiveness of ITS associated with each of these three sample size categories varied significantly under a fixed-effect model, $Q_b(2) = 12.49$, $p = .002$, but not under a random-effects model, $Q_b(2) = 2.77$, $p = .251$. Further analyses showed that the effects did not differ significantly between studies with sample sizes of less than 200 and those with sample sizes over 200 under a fixed-effect model, $Q_b(1) = 2.03$, $p = .154$, nor did it under a random-effects model, $Q_b(1) = 1.96$, $p = .162$.

Research design. For unadjusted overall effect sizes, 15 were from quasi-experimental studies and 11 were from true experiments. The average of the effect sizes from the quasi-experiments, $g = .09$, 95% CI [.05, .14], was larger than that of from true experiments, $g = -.01$, 95% CI [−.07, .05], under a fixed-effect model, $Q_b(1) = 7.97$, $p = .005$. Only the average effect for studies using quasi-experimental designs was significantly different from 0. The difference was not significant under a random-effects model, $Q_b(1) = 0.37$, $p = .544$.

The average of adjusted overall effect sizes from quasi-experiments and true experiments did not differ under either a fixed-effect model, $Q_b(1) = 0.92$, $p = .338$, or under a random-effects model, $Q_b(1) = 1.09$, $p = .296$. None of the average effects were significantly different from 0.

Year of data collection. For unadjusted overall effect sizes, the effects varied depending on the year in which the data were

collected under a fixed-effect model, $Q_b(2) = 10.51$, $p = .005$, but not under a random-effects model, $Q_b(2) = 3.01$, $p = .222$. Only the average effect for studies conducted before 2003 (showing a positive ITS effect) appeared significantly different from 0 under a fixed-effect model. For adjusted overall effect sizes, the effects did not differ significantly in terms of data collection time under either a fixed-effect model, $Q_b(2) = 2.31$, $p = .315$, or a random-effects model, $Q_b(2) = 0.715$, $p = .699$.

Counterbalanced testing. For unadjusted overall effect sizes, the impact of ITS on students' mathematical learning appeared to be lower in studies with counterbalanced testing, $g = -.05$, 95% CI [−.12, .02], than in studies without counterbalanced testing, $g = .09$, 95% CI [.05, .13], under a fixed-effect model, $Q_b(1) = 10.43$, $p = .001$. The average effect size from counterbalanced studies did not differ from 0, whereas the average effect from studies not using counterbalancing did. The difference was not significant under a random-effects model, $Q_b(1) = 1.09$, $p = .297$.

For adjusted overall effect sizes, the impact of ITS appeared to be significantly larger in studies with counterbalanced testing, $g = .08$, 95% CI [.01, .14], than that of studies without counterbalanced testing, $g = -.08$, 95% CI [−.16, −.004] under a fixed-effect model, $Q_b(1) = 8.89$, $p = .003$. The difference was not statistically significant under a random-effects model, $Q_b(1) = 0.44$, $p = .507$.

Report type. We grouped the reports into two categories: peer-reviewed journal reports and nonjournal reports. Nonjournal reports include government reports, conference papers, private reports, master's theses, and doctoral dissertations. For unadjusted overall effect sizes, the average effect size in peer-reviewed journals, $g = .28$, 95% CI [.18, .37], was higher than that of nonjournal reports, $g = .02$, 95% CI [–.02, .05], under a fixed-effect model, $Q_b(1) = 24.45$, $p = .000$. The average effect of ITS was positive in studies in peer-reviewed journals. Under a random-effects model, the average effect size in peer-reviewed journals, $g = .30$, 95% CI [.17, .43], was also statistically significantly higher than that of nonjournal reports, $g = -.01$, 95% CI [−.15, .13], $Q_b(1) = 10.03$, $p = .002$; again, the average effect of ITS was positive in studies in peer-reviewed journals. For adjusted overall effect sizes, the average effect size in peer-reviewed journals was not different from that of nonjournal reports under a fixed-effect model, $Q_b(1) = 0.69$, $p = .407$, nor was the case under a random-effect model, $Q_b(1) = 1.98$, $p = .160$.

Testing for Moderators on All Unadjusted Effect Sizes

The data set of all unadjusted effect sizes consisted of all of the 44 unadjusted effect sizes, not averaged within independent samples. This data set allowed us to study two moderators: the measurement timing and the types of outcomes. The analysis results are also included in Table 2.

Measurement timing. Within a single independent sample, we averaged the effects sizes that related to the same measurement timing. This reduced the 44 unadjusted effect sizes to 27 unadjusted effect sizes that were either associated with outcomes measured at the end of the school year or measured sooner than the end of the school year. The effectiveness of ITS when measured at the end of school year, $g = .01$, 95% CI [−.03, .04], was lower than that when measured sooner than that, $g = .25$, 95% CI [.14, .35], under a fixed-effect model, $Q_b(1) = 17.84$, $p = .000$, but not under a random-effects model, $Q_b(1) = 3.02$, $p = .082$.

Outcome type. As above, we averaged the effect sizes corresponding to the same outcome type within each independent sample. This resulted in 36 effect sizes in our analyses. Five different types of outcomes appeared in the studies. They are (a) course grades, (b) course passing rates, (c) scores from tests developed by teachers or researchers to specifically measure students' learning on the knowledge content that was covered by interventions, (d) scores from modified standardized tests that were either substrands of standardized tests or tests made up of some of the released standardized test questions, and (e) scores from standardized tests. Our preliminary analyses showed that there was no statistically significant difference between the average effect size associated with course grades or course passing rates and that of specifically designed tests, nor was a statistically significant difference between the average effect size associated with modified standardized tests and that of standardized tests. Thus, we grouped the first three outcome types into course-related measures and the last two outcome types into measures from standardized tests. Results show that under a fixed-effect model, the average effect size for course-related measures was $g = 0.19$, 95% CI [.11, .27], and $g = 0.02$, 95% CI [−.02, .06] for measures from standardized tests, $Q_b(1) = 13.19$, $p = .000$. The difference was not significant under a random-effects model, $Q_b(1) = 2.09$, $p = .148$. Course-related measures showed a larger and positive ITS effect and were significantly different from 0 under both models.

Effectiveness of ITS in Comparison to Other Treatment Conditions

In comparison to homework. Mendicino et al. (2009) compared 28 fifth-grade students learning math in two different homework conditions over a period of 1 week. In one condition, students completed paper-and-pencil homework. In another condition, students completed Web-based homework using the ASSISTment system. ASSISTment is a Web-based homework system that facilitates students' learning by providing scaffolds and hints. A number sense problem set and a mixed-problem test were used to measure students' learning after each homework condition. To reduce the possibility that other factors might impact learning, Mendicino et al. implemented counterbalanced procedures so that all students participated in both paper-and-pencil and Web-based conditions. They were tested both before and after the intervention. Mendicino et al. reported an adjusted effect size of 0.61 favoring ITS and concluded that students learned significantly more with the help of the Web-based system than by working on paper-and-pencil homework.

Radwan (1997) compared the math performance of 52 fourth graders. Half were tutored using the Intelligent Tutoring System Model and the other half received no tutoring but worked on completing homework, both during the fifth period of school days. The experiment lasted for a total of 15 hr 40 min every day for 4 weeks. Students' learning was measured through a pretest and posttest of the Computerized Achievement Tests. Radwan's *t* tests on the test scores concluded that the experimental group performed significantly better than the control group did. On the basis of the overall score on the Computerized Achievement Tests, we found that this study yielded an unadjusted $g = .40$, $SE = .28$, and an adjusted $g = .60$, $SE = .28$.

In comparison to human tutoring. Beal et al. (2010) studied the effectiveness of AnimalWatch, an intelligent tutoring system designed to help students learn basic computation and fraction skills to enhance problem-solving performance. The participants were sixth graders enrolled in a summer academic skills class in Los Angeles, California. Once per week for 4 weeks, 13 sixth graders spent 1 hr with math tutors and then 45 min with AnimalWatch, and 12 sixth graders learned math with their tutors (each tutor helped four to six students) using small group activities including blackboard lessons and worksheet practice. The mean proportion of correct scores was used to measure students' performance. Beal et al. concluded that students who spent half of their time using ITS and half of time with human tutors improved as much as did those who spent the entire time learning with a human tutor. We found that this study yielded an unadjusted $g = .20$, $SE = .39$.

Discussion

Summary of the Evidence

Findings of this meta-analysis suggest that, overall, ITS had no negative and perhaps a very small positive effect on K–12 students' mathematical learning relative to regular classroom instruction. When the effectiveness was measured by posttest outcomes and without taking into account the potential influence of other factors, the average unadjusted effect size was .05 under a fixed-effect model and .09 under a random-effects model favoring ITS over regular classroom instruction. After controlling for the influence of other variables (e.g., pretest scores), the average adjusted effect size was .01 under both a fixed-effect model and random-effects model also favoring the ITS condition. However, the average relative effectiveness of ITS did not appear to be significantly different from 0 except when effect sizes were unadjusted and a fixed-effect analysis model was used. Also, whether controlling for other factors or not, there was a high degree of heterogeneity among the effect sizes.

Very few studies compared ITS with homework or human tutoring. The few existing studies showed that when compared with homework or human tutoring, the relative effectiveness of ITS appeared to be small to modest, with effect sizes ranging from .20 to .60.

Testing for moderators yielded some informative findings. Table 4 presents a summary of the findings from moderator analyses using two different estimates of effect (i.e., unadjusted and adjusted effect sizes) and two analysis models (i.e., fixed-effect and random-effects models). Two findings were relatively robust. First, the effects appeared to be greater when the ITS intervention lasted for less than a school year than when it lasted for one school year or longer. This effect appeared regardless of whether the moderator analyses were conducted on unadjusted or adjusted effect sizes with a fixed-effect or random-effects model. Second, the effects of ITS appeared to be greater when the study samples were general students than when the samples were low achievers. And under a fixed-effect model, this difference was statistically significant regardless of whether the analyses were conducted on unadjusted or adjusted effect sizes.

Also, there was some evidence for the following three findings related to the methodology of the study: (a) The effectiveness of ITS appeared to be largest when the learning outcomes were measured before the end of the school year, (b) the effects of ITS appeared to be greater when measured by course-related outcomes than when measured by standardized tests, and (c) the average effect size of studies with smaller sample sizes appeared to be bigger than that of larger sample sizes. In general, these results are consistent with the findings related to methodological characteristics of primary studies in numerous meta-analyses.

Overall Effectiveness of ITS

The conclusion that ITS had no negative and perhaps a very small positive effect on K–12 students' mathematical learning relative to regular classroom instruction is largely congruent with the WWC's conclusions regarding the effects of math educational software programs (WWC, 2004, 2010a, 2010b). Specifically, the WWC (2010a) concluded that Carnegie Learning Curricula and Cognitive Tutor software had no discernible effects on mathematics achievement for high school students but Cognitive Tutor®

Table 4
Findings From Testing for Moderators Across Two Types of Effect Sizes and Two Analysis Models

Variable	ITS favored for	Unadjusted		Adjusted	
		Fixed	Random	Fixed	Random
Subject	Basic math	Yes	Yes	Yes+	Yes
ITS duration	Less than one school year	Yes+	Yes+	Yes+	Yes+
Sample achievement level	General students	Yes+	Yes	Yes+	Yes
Schooling level	Elementary school	Yes	Yes	Yes	Yes
Sample size	Sample size less than 200	Yes+	Yes	Yes	Yes
Research design	Quasi-experiments	Yes+	Yes	No	Yes
Year of data collection	Before 2006	Yes	No	Yes	No
Counterbalanced testing	No	Yes+	Yes	No+	No
Report type	Peer-reviewed journal	Yes+	Yes+	No	Yes
Measurement timing	Sooner than end of school year	Yes+	Yes		
Outcome type	Course-related outcome measures	Yes+	Yes		

Note. *Yes* denotes that the subcategory, for example, basic math, appears to be favored over the other subcategory or subcategories of that variable (i.e., subject). A + denotes that the effects of the intelligent tutoring system (ITS) on the favored feature (i.e., subcategory), for example, basic math, was statistically significantly greater than those on the other feature (i.e., subcategory), such as algebra.

Algebra I had potentially positive effects on ninth graders' math achievement. The WWC (2004) found that students who used Cognitive Tutor earned significantly higher scores on the Educational Testing Service Algebra I test and on their end-of-semester grades than their counterparts who were taught with traditional instruction. Furthermore, the WWC (2010b) concluded that PLATO Achieve Now had no discernible effects for six graders' math achievement but the WWC considered the extent of evidence to be small.

It is relevant to mention that the WWC conclusions were based on a very limited number of studies that met their evidence standards or met their standards with reservation. The present meta-analysis included all seven reports that had been identified as meeting the WWC's evidence standards or meeting their standards with reservation. This meta-analysis covered studies of many other ITS programs in addition to the two reviewed by the WWC. Thus, despite the differences in review scopes and methodology, the finding that ITS appeared to have no negative and perhaps a small positive effect on students' mathematical achievement is largely consistent with the conclusion from the WWC reviews.

Comparing the present meta-analysis with a recent meta-review by VanLehn (2011) illuminates an interplay of many issues pertaining to the effectiveness of ITS. VanLehn (2011) reviewed randomized experiments that compared the effectiveness of human tutoring, computer tutoring, and no tutoring. He found that the effect size of ITS was 0.76, which was nearly as effective as human tutoring, $d = 0.79$. It appears that VanLehn (2011) found a larger effect than what the present meta-analysis revealed. However, we found that these two systematic reviews are different in at least three ways. First, the two reviews differ in subject domains and grade levels of students. The VanLehn (2011) review included studies of science, technology, engineering, and mathematics, with no restriction of grade levels. As a result, it included a large portion of studies on the use of ITS in college students' learning. The present meta-analysis focuses on the effectiveness of ITS on K–12 students' mathematical learning. Second, the two reviews had different methodological standards and applied different study inclusion criteria. VanLehn (2011) covered experiments that manipulated ITS interventions while controlling for other influences and excluded studies in which the experimental and comparison groups received different learning content. For example, it excluded all studies of Carnegie Learning's Cognitive Tutor because students in the experimental groups used a different textbook and classroom activities than did those in the comparison groups. In contrast, in the present meta-analysis, we placed no such restrictions. In fact, our meta-analysis includes studies that compared two ecologically valid conditions in which ITS may or may not be the only difference between the conditions. As a result, 20 out of the 31 independent studies included in this meta-analysis are studies of Cognitive Tutor. Last, VanLehn (2011) selected the outcome with the largest effect size in each primary study. The present meta-analysis extracted effect sizes for all the outcomes possible in each study and averaged them. Taken together, the differences mentioned above may help explain the seemly discrepant findings from these two reviews. An overarching message from this is that when addressing the effectiveness of ITS, as is the case with many other educational interventions, one ought to ask a few questions: for whom, compared with what, and in what circumstances?

We compared the findings of the current meta-analysis with those of four recent reviews that focused on the effectiveness of computer technology or educational software on Pre-K to 12th graders' mathematical achievement. Methodologically, these reviews are also largely comparable with the current meta-analysis. In general, compared with the findings of some similar meta-analyses or systematic reviews of the effectiveness of educational technology, the effects of ITS appear to be relatively small.

Kulik (2003) reviewed 36 controlled evaluation studies to examine the effects of using instructional technology on mathematics and science learning in elementary and secondary schools. He found that the median effect of integrated learning systems was to increase mathematics test scores by 0.38 standard deviations, or from the 50th to the 65th percentiles. He also found that the median effect of computer tutorials was to raise student achievement scores by 0.59 standard deviations, or from the 50th to the 72th percentiles.

Murphy et al. (2002) reviewed 13 studies of the efficacy of discrete educational software on Pre-K to 12th grade students' math achievement. They found that the overall weighted mean effect size for discrete educational software applications in math instruction was 0.45, and the median effect size was 0.27. On the basis of the distribution of confidence intervals, they concluded that $d = 0.30$ or greater appeared to be a reasonable estimate for the effectiveness of discrete educational software on mathematics achievement.

Slavin and Lake (2008) reviewed 38 studies to investigate the effects of CAI on elementary mathematics achievement. They found that the median effect size was 0.19. In their review of middle and high school math programs, Slavin, Lake, and Groff (2009) found that the weighted mean effect size was 0.10 for the effectiveness of CAI.

We should be quick to point out that conclusions based on the comparisons of the findings from different reviews ought to be tentative because there were variations among the reviews regarding the types of educational technology. As the use of educational technology became such a common practice in teaching and learning, it is increasingly difficult to picture a matrix of existing and ever-changing educational technology. As described in the introduction and Method section, we defined ITS as self-paced, learner-led, highly adaptive, and interactive learning environments operated through computers. In the studies included in this meta-analysis, ITS delivered learning content to students, tracked and adapted to students' learning paces, assessed learning progress, and gave students feedback. We believe these features distinguish ITS from other educational technologies in previous reviews.

One possible explanation for the small effects revealed in this meta-analysis is related to the degree of technology implementation and the purposes of technology use in classrooms. Evidence suggests that computer technology appears to have stronger effects when being used as supplemental tools than when used as the only or main instructions. For example, Schmid et al. (2009) found that in terms of degree of technology use, low ($g = 0.33$) and medium use of technology ($g = 0.29$) produced significantly higher effects than did high use ($g = 0.14$). They also found that in terms of type of technology use, when used as cognitive support (e.g., simulations), educational technology produced better results ($g = 0.40$) than when it was used as a presentational tool ($g = 0.10$) or for multiple uses ($g = 0.29$). Also, Tamim, Bernard, Borokhovski,

Abrami, and Schmid (2011) found that computer technology produced a slightly but significantly higher average effect size when used as supporting instruction than when it was used for direct instruction. Taken together, these findings imply that computer technology's major strengths may lie in supporting teaching and learning rather than substituting or replacing main instructional approaches or acting as a stand-alone tool for delivering content. However, further research is needed to reach a firm conclusion. Schmid et al. (2009) argued that future research ought to move away from the "yes-or-no" question and move to other issues, such as how much technology is desirable for improving student learning and how to best use technology to promote educational outcomes.

In addition, much research has supported the view that educational technology can improve student motivation and therefore positively influence student academic performance (Beeland, 2002; Roblyer & Doering, 2010; U.S. Department of Education, 1995). We speculate that as educational technology has become such a common part of learning environments in today's educational settings, student motivation and novelty effects related to the access of educational technology might have decreased. As a result, the relative effectiveness of educational technology may be declining.

Last, findings of this meta-analysis need to be interpreted with caution. As we described earlier, the results were largely based on the 31 independent studies that compared learning outcomes of instructions with an ITS component with those without one. This broad comparison covered four types of categorized situations that were different from one another to varying degrees. For studies in which ITS were the only difference between the treatment and comparison conditions, it is reasonable to conclude that the detected effectiveness difference can be attributed to ITS. However, when ITS were not the only difference between the conditions, differences in outcome measures cannot be attributed solely to ITS. For example, for studies of Cognitive Tutors, the treatment and comparison conditions could differ not only in the use of Cognitive Tutors but also in teachers or school environments. In such cases, it cannot be ruled out that the effectiveness of ITS is masked by the relative ineffectiveness of the other intervention components, such as teachers or school environments. It also could be the case that the effectiveness of the other intervention components is masked by the relative ineffectiveness of ITS. Taken together, this meta-analysis provides information regarding whether and how students' learning outcomes might differ depending on the involvement or absence of ITS from the instructions. However, one ought to be aware that the effectiveness differences may or may not be attributed solely to ITS.

Findings From Testing for Moderators

As summarized earlier, two robust findings stand out from this meta-analysis. The first finding was that the effects of ITS appeared to be greater when the intervention lasted for less than a school year than when it lasted for one school year or longer. We offer three possible explanations for this finding. First, it might be that the novelty factor wears off and students' motivation declines. This explanation would suggest that, just as is the case for many other interventions, too much of a good thing is not a good thing. Again, this brings us back to the important issue regarding how much technology is desirable and how to best use technology to improving student learning. Second, when ITS were in regular or long-term use in schools, researchers usually had no or very little involvement in the actual use of ITS during the study. In other words, the degree of implementation might have impacted the effectiveness of ITS.

Third, some differences in the durations of interventions might be responsible for the differential effectiveness of ITS. Specifically, we found that a number of major characteristics of the studies in which ITS lasted for one school year or longer might account for the small effect sizes yielded in the studies. These studies, such as the studies of Cognitive Tutor (e.g., Campuzano et al., 2009; Dynarski et al., 2007), were more often based on big national samples; used more rigorous study methods, such as random assignment; and used more distal outcome measures, such as standardized achievement tests. In contrast, studies in which ITS lasted for short time or one semester generally produced bigger effect sizes for a number of reasons. For example, they studied relatively less known ITS, they were more often based on small sample sizes, they used less rigorous study methods, and they often used specifically designed or nonstandardized outcome measures. Many previous meta-analyses have concluded that the study differences mentioned above have an impact on the magnitude of effect sizes. Moderator analyses of this meta-analysis confirm this conclusion. We need to use caution in applying this finding to practices before further research is conducted and a firmer conclusion is reached.

The second finding was that ITS helped general students learn mathematical subjects more than it helped low achievers. One possible explanation is that ITS may function best when students have sufficient prior knowledge, self-regulation skills, learning motivation, and familiarity with computers. It is possible that general students have more of the characteristics needed to navigate ITS than low achievers do. Therefore, they benefited more from using ITS. For low achievers, classroom teachers, rather than ITS, might be better leaders, motivators, and regulators to help them learn. Research has found that there are differences in the ways that high achievers and low achievers used ITS and other computer-based instruction tools (Hativa, 1988; Hativa & Shorer, 1989; Wertheimer, 1990). For example, Hativa (1988) found that low achievers, more than high achievers, were prone to make software- and hardware-related errors when working with a CAI system. He further concluded that it was possible that high achievers were much more able than low achievers to adjust to a CAI learning environment so that they were able to benefit more from it.

This finding draws new attention to the debate regarding whether the use of computer technology actually widens the achievement gap between high achievers and low achievers, students of high and low learning aptitudes, students with advantaged and disadvantaged backgrounds, or White and minority students. The results from some longitudinal studies of CAI have provided support for the notion that computerized learning contributes to the increasing achievement gaps between students with different socioeconomic statuses, achievement levels, and aptitudes (Hativa, 1994; Hativa & Becker, 1994; Hativa & Shorer, 1989). Ceci and Papiero (2005) noted that nontargeted technology intervention that is used differently by advantaged and disadvantaged groups of students leads to achievement gap widening. This meta-analysis

adds further support to the above conclusions with the evidence that ITS might have contributed to the achievement gap between higher and lower achieving students. It is worth noting that only three studies provided results for low achievers.

This issue merits considerable attention. As mentioned earlier, the motivation of ITS development is to help students achieve learning gains as they do with the help of expert human tutors. There has been the expectation that ITS, as a form of advanced learning technology, ought to be able to provide optimal conditions needed to teach all children, given their interactivity, adaptability, and ability to provide immediate feedback and reinforcement. Developers of ITS may also want to consider way to adapt ITS for students with a variety of aptitudes and design culturally relevant technology learning environments. Further research with more nuanced approaches for ITS evaluation is needed to provide more information for this issue.

Conclusion

This meta-analysis synthesized studies of the relative effectiveness of ITS compared with regular class instruction on K–12 students' mathematical learning. Findings suggest that overall, ITS appeared to have no negative and perhaps a small positive impact on K–12 students' mathematical learning. The main contributions of this meta-analysis lie on three fronts. First, it provided further evidence for the conclusions that educational technology might be best used to support teaching and learning. Second, this meta-analysis revealed that ITS appeared to have a greater positive impact on general students than on low achievers. This finding will likely draw considerable attention in policy debates on the issue of whether computerized learning might contribute to the achievement gap between students with different achievement levels or prior backgrounds. Meanwhile, this finding implies that ITS research might be helpful in gaining a better understanding of how better learners learn through ITS, especially in terms of cognitive and metacognitive factors. Third, findings of this meta-analysis confirm several conclusions from many previous meta-analyses concerning the association between methodological features (e.g., sample size, research design, and outcome measure) of primary research and the effectiveness of the intervention studied. On the basis of the findings of this meta-analysis and similar reviews of educational technology, it seems best to think of ITS as one option in the array of education resources that educators and students can use to support teaching and learning. For students who are motivated and can self-regulate learning, ITS might be effective supplements to regular class instruction. However, ITS may not be efficient tools to boost low achievers' or at-risk students' achievement.

References

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

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16,* 101–128.

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

*Arbuckle, W. J. (2005). *Conceptual understanding in a computer assisted Algebra I classroom* (Doctoral dissertation). Retrieved from ProQuest Information and Learning Company. (UMI No. 3203318)

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

*Arroyo, I., Woolf, B. P., Royer, J. M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Lecture Notes in Computer Science: Vol. 6094. Intelligent tutoring systems* (pp. 423–432). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-642-13388-6_46

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

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

Becker, H. J. (1992). Computer-based integrated learning systems in the elementary and middle grades: A critical review and synthesis of evaluation reports. *Journal of Educational Computing Research, 8,* 1–41. doi:10.2190/23BC-ME1W-V37U-5TMJ

Beeland, W. D., Jr. (2002). *Student engagement, visual learning and technology: Can interactive whiteboards help?* Retrieved from Valdosta State University website: http://chiron.valdosta.edu/are/Artmanscrpt/vol1no1/beeland_am.pdf

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

Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research, 72,* 101–130. doi:10.3102/00346543072001101

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.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2006). Comprehensive Meta-Analysis (Version 2.2.027) [Computer software]. Englewood, NJ: Biostat.

*Cabalo, J. V., Ma, B., & Jaciw, A. (2007). *Comparative effectiveness of Carnegie Learning's "Cognitive Tutor Bridge to Algebra" curriculum: A report of a randomized experiment in the Maui School District*. Palo Alto, CA: Empirical Education.

*Cabalo, J. V., & Vu, M.-T. (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.

*Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts—Executive summary* (NCEE 2009-4042). Retrieved from U.S. Department of Education, Institute of Education Sciences, National Center for Education and Regional Assistance website: http://ies.ed.gov/ncee/pubs/20094041/pdf/20094042.pdf

*Carnegie Learning. (2001a). *Cognitive Tutor research results: Freshman Academy, Canton City Schools, Canton, OH* (Cognitive Tutor Research Report OH-01–91). Pittsburgh, PA: Author.

*Carnegie Learning. (2001b). *Cognitive Tutor results report: Freshman Academy, Canton City Schools, Canton, OH, 2001*. Retrieved from http://www.carnegielearning.com/static/web_docs/OH-01-01.pdf

*Carnegie Learning. (2002). *Cognitive Tutor results report*. Pittsburgh, PA: Author.

Ceci, S. J., & Papiero, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *American Psychologist, 60,* 149–160. doi:10.1037/0003-066X.60.2.149

Cheung, A., & Slavin, R. E. (2012). How features of educational technology programs affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215. doi:10.1016/j.bbr.2011.03.031

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

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

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

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 2109. User modeling 2001* (pp. 137–147). Berlin, Germany: Springer-Verlag.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin, Germany: Springer. doi:10.1007/978-3-642-58625-5

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. doi:10.1111/j.0006-341X.2000.00455.x

*Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., . . . Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort—Report to Congress* (NCEE 2007-4005). Retrieved from U.S. Department of Education, Institute of Education Sciences website: http://ies.ed.gov/ncee/pdf/20074005.pdf

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

Graesser, A. C., Conley, M., & Olney, A. (2011). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 3. Applications to learning and teaching* (pp. 451–473). Washington, DC: American Psychological Association.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21,* 27–58.

Haslam, M. B., White, R. N., & Klinge, A. (2006). *Improving student literacy: READ 180 in the Austin Independent School District, 2004–05*. Washington, DC: Policy Studies.

Hativa, N. (1988). Computer-based drill and practice in arithmetic: Widening the gap between high- and low-achieving students. *American Educational Research Journal, 25,* 366–397. doi:10.3102/00028312025003366

Hativa, N. (1994). What you design is not what you get (WYDINWYG): Cognitive, affective, and social impacts of learning with ILS—An integration of findings from six-years of qualitative and quantitative studies. *International Journal of Educational Research, 21,* 81–111. doi:10.1016/0883-0355(94)90025-6

Hativa, N., & Becker, H. J. (1994). Integrated learning systems: Problems and potential benefits. *International Journal of Educational Research, 21,* 113–119. doi:10.1016/0883-0355(94)90026-4

Hativa, N., & Shorer, D. (1989). Socioeconomic status, aptitude, and gender differences in CAI gains of arithmetic. *Journal of Educational Research, 83,* 11–21.

*Hwang, G.-J., Tseng, J. C. R., & Hwang, G.-H. (2008). Diagnosing student learning problems based on historical assessment records. *Innovations in Education and Teaching International, 45,* 77–89. doi:10.1080/14703290701757476

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

*Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30–43.

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279–294). New York, NY: Russell Sage Foundation.

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say* (SRI Project No. P10446.001). Arlington, VA: SRI International.

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

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers, 36,* 222–233. doi:10.3758/BF03195567

*Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education, 41,* 331–358.

*Morgan, P., & Ritter, S. (2002). *An experimental study of the effect of Cognitive Tutor Algebra I on student knowledge and attitude*. Retrieved from the Carnegie Learning website: http://carnegielearning.com/web_docs/morgan_ritter_2002.pdf

Murphy, R. F., Penuel, W. R., Means, B., Korbak, C., Whaley, A., & Allen, J. E. (2002). *E-DESK: A review of recent evidence on the effectiveness of discrete educational software*. Menlo Park, CA: SRI International.

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

*Plano, G. S., Ramey, M., & Achilles, C. M. (2007, January). *Implications for student learning using a technology-based algebra program in a ninth-grade algebra course*. Paper presented at the 13th Annual Office of Superintendent of Public Instruction January Conference and High School Summit, Seattle, WA.

*Radwan, Z. R. (1997). *Evaluation of the effectiveness of a computer assisted intelligent tutor system model developed to improve specific learning skills of special needs student* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 9729551)

*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, U. Hoppe, & S. S.-C. Young (Eds.), *Frontiers in Artificial Intelligence and Applications: Vol. 162: Supporting learning flow through integrative technologies* (pp. 13–20). Amsterdam, the Netherlands: IOS Press.

Roblyer, M., & Doering, A. (2010). *Integrating educational technology into teaching* (5th ed.). Boston, MA: Allyn & Bacon.

Rowley, K., Carlson, P., & Miller, T. (1998). A cognitive technology to teach composition skills: Four studies with the R-Wise writing tutor. *Journal of Educational Computing Research, 18,* 259–296. doi:10.2190/KW4V-FJKD-L7J1-EFK0

*Sarkis, H. (2004). *Cognitive Tutor Algebra 1 program evaluation: Miami–Dade County Public Schools*. Lighthouse Point, FL: Reliability Group.

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R., Abrami, P. C., Wade, C. A., . . . Lowerison, G. (2009). Technology's effect on achievement in higher education: A Stage I meta-analysis of classroom applications. *Journal of Computing in Higher Education, 21,* 95–109. doi:10.1007/s12528-009-9021-8

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

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1,* 51–77. doi:10.1080/1049482900010104

Shute, V. J., & Zapata-Rivera, D. (2007). *Adaptive technologies* (Research Report RR-07-05). Princeton, NJ: Educational Testing Service.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary math: A best-evidence synthesis. *Review of Educational Research, 78,* 427–515. doi:10.3102/0034654308317473

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

*Smith, J. E. (2001). *The effect of the Carnegie Algebra Tutor on student achievement and attitude in introductory high school algebra* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.

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

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

U.S. Department of Education. (1995). Effects on students. In *Technology and education reform: Technical research report*. Retrieved from http://www.ed.gov/pubs/SER/Technology/ch9.html

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16,* 227–265.

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

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

VanLehn, K., Jordan, P. W., Rosé, C. P., Bhembe, D., Böttner, M., Gaydos, A., . . . Srivastava, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Lecture Notes in Computer Science: Vol. 2363. Intelligent Tutoring Systems: 6th International Conference* (pp. 158–167). Berlin, Germany: Springer.

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

Wertheimer, R. (1990). The geometry proof tutor: An "intelligent" computer-based tutor in the classroom. *Mathematics Teacher, 83,* 308–317.

What Works Clearinghouse. (2004, December). *What Works Clearinghouse topic report: Curriculum-based interventions for increasing K-12 math achievement–middle school*. Retrieved from Department of Education, Institute of Education Sciences, website: http://www.eric.ed.gov/PDFS/ED485395.pdf

What Works Clearinghouse. (2007, May). *WWC intervention report middle school math: Cognitive Tutor Algebra I*. Retrieved from Department of Education, Institute of Education Sciences, website: http://www.aea9.k12.ia.us/documents/filelibrary/pdf/cognitive_tutor/WWC_Cognitive_Tutor_052907_3B8688D14AA44.pdf

What Works Clearinghouse. (2008). *What Works Clearinghouse: Procedures and standards handbook* (Version 2.0). Retrieved from Department of Education, Institute of Education Sciences, website: http://ies.ed.gov/ncee/wWc/pdf/reference_resources/wwc_procedures_v2_standards_handbook.pdf

What Works Clearinghouse. (2009, July). *WWC intervention report middle school math: Cognitive Tutor Algebra I*. Retrieved from Department of Education, Institute of Education Sciences, website: http://www.aea9.k12.ia.us/documents/filelibrary/pdf/cognitive_tutor/WWC_CogTutor_Report_July2009_B2A3C279D0481.pdf

What Works Clearinghouse. (2010a, August). *WWC intervention report high school math: Carnegie Learning curricula and cognitive tutor software*. Retrieved from Department of Education, Institute of Education Sciences, website: http://ies.ed.gov/ncee/wWc/pdf/intervention_reports/wwc_cogtutor_083110.pdf

What Works Clearinghouse. (2010b, March). *WWC intervention report middle school math: Plato Achieve Now*. Retrieved from Department of Education, Institute of Education Sciences, website: http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=378

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6,* 413–429. doi:10.1037/1082-989X.6.4.413

Woolf, B. P. (2009). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Kaufman.

Received November 10, 2011
Revision received February 18, 2013
Accepted March 4, 2013 ■