Journal of Educational Psychology
2014, Vol. 106, No. 2, 331–347

© 2013 American Psychological Association
0022-0663/14/$12.00 DOI: 10.1037/a0034752

A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning

Saiying Steenbergen-Hu and Harris Cooper
Duke University

This meta-analysis synthesizes research on the effectiveness of intelligent tutoring systems (ITS) for college students. Thirty-five reports were found containing 39 studies assessing the effectiveness of 22 types of ITS in higher education settings. Most frequently studied were AutoTutor, Assessment and Learning in Knowledge Spaces, eXtended Tutor-Expert System, and Web Interface for Statistics Education. Major findings include (a) Overall, ITS had a moderate positive effect on college students' academic learning ($g = .32$ to $g = .37$); (b) ITS were less effective than human tutoring, but they outperformed all other instruction methods and learning activities, including traditional classroom instruction, reading printed text or computerized materials, computer-assisted instruction, laboratory or homework assignments, and no-treatment control; (c) ITS's effectiveness did not significantly differ by different ITS, subject domain, or the manner or degree of their involvement in instruction and learning; and (d) effectiveness in earlier studies appeared to be significantly greater than that in more recent studies. In addition, there is some evidence suggesting the importance of teachers and pedagogy in ITS-assisted learning.

Keywords: intelligent tutoring systems, effectiveness, college students, academic learning, meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/a0034752.supp

Intelligent tutoring systems (ITS) are computer-assisted learning environments. They are highly adaptive, interactive, and learner-paced learning environments created using computational models developed in the learning sciences, cognitive sciences, mathematics, computational linguistics, artificial intelligence, and other relevant fields (Graesser, Conley, & Olney, 2011). They are designed to follow the practices of expert human tutors (Woolf, 2009). ITS are adaptive in that they adjust and respond to learners with tasks or steps suited to the learners' individual characteristics, needs, or pace of learning (Shute & Zapata-Rivera, 2007). ITS and automated tutoring systems are covered by the National Science Foundation's portfolio in educational technologies (Cherniavsky & Vanderputten, 2003). ITS have been used as educational tools in both K-12 and higher education settings. Cognitive Tutors by Carnegie Learning, for example, were used in over 2,600 schools in the United States as of 2010 (What Works Clearinghouse [WWC], 2010a).

Subject Matter Scope of ITS

ITS have been developed for mathematically grounded academic subjects. To name just a few, there are ITS for algebra (e.g., Cognitive Tutors: Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997; Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007), basic mathematics (e.g., *AnimalWatch*: Beal, Arroyo, Cohen, & Woolf, 2010), statistics (e.g., Web Interface for Statistics Education [WISE]: Aberson, Berger, Healy, & Romero, 2003; Assessment and Learning in Knowledge Spaces [ALEKS]: Doignon & Falmagne, 1999), physics (e.g., Andes, Atlas, and Why/Atlas: VanLehn et al., 2002, 2007), and computer science (e.g., dialogue-based ITS: Lane & VanLehn, 2005; ACT Programming Tutor: Corbett, 2001).

Some ITS assist with learning of other subjects such as reading (e.g., READ 180: Haslam, White, & Klinge, 2006; iSTART: McNamara, Levinstein, & Boonthum, 2004), writing (e.g., R-WISE writing tutor: Rowley, Carlson, & Miller, 1998), economics (e.g., Smithtown: Shute & Glaser, 1990), and research methods (e.g., Research Methods Tutor: Arnott, Hastings, & Allbritton, 2008). There are also ITS for specific skills, such as metacognitive skills (Aleven, McLaren, Roll, & Koedinger, 2006; Conati & VanLehn, 2000).

Theoretical Underpinnings of ITS

ITS are often designed by incorporating pedagogical, psychological, or other cognitive learning theories into computational models (Graesser et al., 2011). For example, Cognitive Tutors are built on a cognitive theory called adaptive control of thought (ACT, or ACT-R in its updated form; Anderson et al., 1995). According to ACT-R, a cognitive skill consists of many units of goal-related domain knowledge (e.g., knowing the side–angle–side theorem) and goal-independent procedural knowledge (e.g., the ability to use the theorem). Cognitive skill acquisition involves converting a large set of declarative knowledge through the formulation of production rules that represent procedural knowledge.

This article was published Online First December 23, 2013.

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

Correspondence concerning this article should be addressed to Saiying Steenbergen-Hu, who is now at the Center for Talent Development, School of Education and Social Policy, Northwestern University, 617 Dartmouth Place, Evanston, IL 60208-4175. E-mail: hu@northwestern.edu

The conversion can be achieved through problem solving. It is a dynamic process in which a learner's behavior and knowledge are constantly assessed so that learning is reconstructed based on what the learner has already mastered and subsequent learning focuses on what the learner has yet to learn.

Constraint-based tutors are based on the constraint-based model (CBM), derived from Ohlsson's theory of learning from performance errors (Ohlsson, 1992, 1996). The core idea of CBM is to deliver the domain knowledge as a set of constraints, which can be used to analyze students' solutions and provide feedback on errors (Mitrovic, 2012).

As another example, AutoTutor is a type of ITS designed to facilitate learning through holding dialogues with students in natural language (Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Tutoring Research Group, University of Memphis, 1999). AutoTutor was built by incorporating explanation-based constructivist theories of learning into the practices of human tutoring, which usually involve collaborative constructive activities (Aleven & Koedinger, 2002; Chi, de Leeuw, Chiu, & LaVancher, 1994).

ITS and Other Forms of Computerized Instruction

ITS are an advanced learning technology that can be distinguished from computer-assisted instruction (CAI), computer-based training (CBT), and e-learning. ITS are considered superior to CAI and CBT 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) based on 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 these steps, typically in the form of feedback, hints, or error messages. In this regard, CAI, CBT, and Web-based homework are different from ITS in that they lack an inner loop (VanLehn, 2006). ITS are one type of e-learning that usually encompasses all forms of teaching and learning that are electronically supported in the forms of texts, images, animations, audios, or videos.

The Need for a Meta-Analysis of the Effectiveness of ITS in Higher Education

A meta-analysis of the effectiveness of ITS on college students' academic learning is needed for four reasons. First, existing research syntheses of the effectiveness of educational technology in higher education usually cover a broad range of technologies, but few have focused on ITS. For example, Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) conducted a second-order meta-analysis of 40 years of research on the impact of technology on learning. They selected 25 meta-analyses that studied a variety of technologies, including CAI, CBI, computer-simulation instruction, word processor, hypermedia, information and communication technology, simulation, and digital media. Sosa, Berger, Saw, and Mary (2011) meta-analyzed the effectiveness of CAI on statistics learning covering three broad categories of computer-based tools: number crunchers (e.g., SPSS, STATA, Minitab), communication-based instructional tools (e.g., e-mail, LISTSERV, Blackboard, or Sakai), and stand-alone tools or tutorials of which some can be identified as ITS. Schenker (2007) and Hsu (2003) reported two similar meta-analyses of the effectiveness of technology on statistics instruction that covered a wide range of technologies. Thus, previous research syntheses either omitted ITS or blended them with other types of technologies. There is a need to distinguish ITS from other types of educational technologies and conduct a separate systematic summarization of their effectiveness.

Second, although much research on the effectiveness of ITS in higher education has accumulated over the past 2 decades, many fundamental questions remain unanswered. Graesser et al.'s (2011) review of ITS posed a number of questions that need answering. For example, do different versions of ITS and different types of ITS have different impacts on student learning; do different computer tutors that handle the same subject have different influences, and if they do, which are the most effective and for what kind of learner populations; and is there an Aptitude × Treatment interaction? Although many of the questions need to be addressed in primary empirical investigations, some are more suited to research syntheses, especially meta-analyses.

Third, researchers have conducted some systematic reviews on ITS's effectiveness on K-12 students' learning, but none exists for higher education. For example, the WWC produced four reviews of Carnegie Learning's Cognitive Tutors (WWC, 2004, 2007, 2009, 2010a) and one review of Plato Achieve Now (WWC, 2010b), all focusing on K-12 students' math learning. More recently, Steenbergen-Hu and Cooper (2013) conducted a meta-analysis of ITS's effectiveness on K-12 students' mathematical learning. VanLehn (2011) reviewed studies of ITS's impact on students' learning of science, technology, engineering, and mathematics (STEM) subjects. This review is an important recent effort of reviewing ITS's effectiveness. However, it included studies of all education levels and did not examine the effectiveness in terms of education contexts or grade levels. Furthermore, it only included studies that compared ITS with human tutoring or no tutoring.

Last, we need not only to know ITS's effectiveness overall but to know it in specific contexts. For example, we need to know whether and how ITS's effectiveness differs when they are compared to different learning interventions, whether ITS affect students' learning differently in different subject domains, and in what circumstances ITS help students most.

The Current Meta-Analysis

The current meta-analysis extends previous systematic reviews of educational technology's impact on learning in higher education by focusing on ITS, a particular type of advanced learning technology. Furthermore, it extends previous ITS research syntheses by focusing on college students' learning. It examines possible differential effectiveness due to the different manners or degrees of ITS's involvement in learning, synthesizes the effectiveness relative to different comparison conditions, and addresses some remaining unanswered questions in literature.

Specifically, the current meta-analysis addresses the following six major questions: (a) What is the overall effectiveness of ITS on college students' academic learning? (b) Do ITS impact students' learning differently depending on the types of instructions or learning activities to which they are compared? (c) Do ITS affect learning differently depending on the subject domain? (d) Does the manner in which ITS are used matter, for example, does it matter whether they are used as a principal instruction tool or integrated with regular classroom instruction? (e) Does the length of ITS instruction matter? and (f) Do different ITS impact learning differently? In addition, several other moderators of ITS's effectiveness are also examined in a more exploratory fashion. Taken together, this meta-analysis is intended to add new information regarding the effectiveness of ITS and expand our knowledge about advanced learning technologies.

Method

Distinguish ITS From Other Computer Technologies

When synthesizing the effectiveness of educational technology, an important issue is to address the overlap and diffusion among various types of technologies and clearly define the particular type of technology of interest. Although the term *intelligent tutoring systems* has been used in much research, many studies have used other terms. The great variation of terminology can lead to neglect of relevant studies. It also complicates the interpretation of the results and limits the practical implication of the findings. Therefore, it is not only important but necessary to clearly define ITS and distinguish them from other forms of computer technologies. As defined in the introduction, ITS are highly adaptive, interactive, and learner-paced learning environments operated through computers. Operationally, the current meta-analysis used the following rules to distinguish studies of ITS from those of other computer technologies in the process of selecting primary studies.

First, ITS are domain-related stand-alone computer tutorials. This refers to the fact that each one contains its own unique instructional content and specializes in domain-related knowledge. This distinguishes ITS from many domain-independent computer technologies that are not designed for facilitating conceptual learning. Such technologies include, for example, computational aids or statistical software packages (e.g., SPSS or Minitab) and communication support systems (e.g., Blackboard or Sakai). Sosa et al.'s (2011) meta-analysis of the effectiveness of CAI on statistics learning is helpful to illustrate how the current meta-analysis distinguished ITS from other computer technologies. They categorized CAI into three broad categories: stand-alone tools or tutorials that facilitate learning with little or no input from an instructor, number crunchers (e.g., SPSS or Minitab), and communication tools (i.e., Blackboard or Sakai). Among the 45 studies included in their meta-analysis, 12 of them covered studies of stand-alone tools or tutorials. Only those studies were of interest for the current meta-analysis and were screened for possible inclusion.

In recent years, researchers have developed ITS for teaching specific skills, such as metacognitive skills (Aleven et al., 2006; Conati & VanLehn, 2000), which are seemingly domain independent. However, existing literature suggests that ITS's effectiveness on learning such seemingly domain-independent skills is often assessed within certain well-defined subject areas. For example, Chi and VanLehn (2010) studied college students' use of ITS to learn a metacognitive strategy in probability and its transfer to physics. Therefore, studies of skill-targeting ITS were also an interest of this meta-analysis. However, the current meta-analysis identified no such studies that were qualified for inclusion.

Second, studies of ITS usually provide detailed descriptions to identify whether the technologies studied were ITS or not. Specifically, ITS studies usually have the following features. First, they usually share much of the same literature base. For example, ITS studies often cited Sleeman and Brown (1982) as the original source of ITS terminology, and many ITS studies mentioned artificial intelligence as the precursor of the ITS field. Second, typical ITS studies describe ITS as interactive, self-paced, or learner-controlled; they also provide detailed descriptions of ITS's working processes, which usually consist of delivering learning content to students, tracking and adapting to students' learning pace, assessing learning progress, and providing feedback. Third, they usually discuss the models or theories upon which particular ITS are based, their architectures, and user interfaces. Last, authors usually have associations with certain teams that specialize in particular ITS and have conducted a series of work on them over years. For example, the team of Anderson, Koedinger, and Ritter specializes in Cognitive Tutors; Graesser and his team focus on AutoTutor, a type of intelligent conversational agent; VanLehn and his team focus on dialogue-based ITS (e.g., Andes Physics Tutoring System); and Mitrovic, Ohlsson, and their colleagues have worked on constraint-based tutors (e.g., Structured Query Language-Tutor) for more than a decade. These features were very useful in helping distinguish studies of ITS from those of other computer technologies and ensuring this meta-analysis succeeded in identifying qualified studies. In fact, the important features of typical ITS studies as summarized above are also one of the findings of the current research synthesis.

Study Inclusion and Exclusion Criteria

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

1. Studies had to be empirical primary investigations of the effects of ITS on college students' academic learning. Secondary data analyses and literature reviews were excluded.

2. Studies had to focus on students in general higher education institutions. Studies focusing exclusively on students in professional schools (e.g., medical schools, law schools, or military academies) were excluded; studies focusing exclusively on students with learning disabilities or social or emotional disorders (e.g., students with attention-deficit/hyperactivity disorder) were also excluded.

3. Studies had to have used an independent comparison condition that could have been regular classroom instruction, computerized instruction, human tutoring, self-reliant learning activities, doing homework, or no-treatment control. The following types of studies were excluded: studies without a comparison group, those with one-group pretest–posttest designs, and those comparing one type or version of ITS with another.

4. Studies had to employ randomized 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 (WWC, 2013). Studies with a significant preexisting difference between the treatment and comparison groups were excluded unless information was available for us to calculate effect sizes that would take into account the prior difference.

5. Studies had to measure ITS's effectiveness on at least one learning outcome. Common outcome measurements included course grades or scores on tests developed by researchers.

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

7. Studies had to be published or reported during the period from January 1, 1990, to July 2012 and had to be available in English. The year 1990 was chosen as a cutoff time because many evaluation studies of ITS's effectiveness began to appear in the beginning of the 1990s.

Study Search and Identification

We conducted a four-stage study search and identification procedure. The first stage was an electronic search consisting of two procedures: (a) a search of electronic databases including ERIC, PsycINFO, ProQuest Dissertations and Theses, Academic Search Premier, Econlit with Full Text, PsycARTICLES, SocINDEX with Full Text, and Science Reference Center and (b) Web searches using the Google and Google Scholar search engines. We used a wide variety of search terms, including *intelligent tutor**, *artificial tutor**, *computer tutor**, *computer-assisted tutor, computer-based tutor**, *intelligent learning environment**, *computer coach**, *online-tutor**, *keyboard tutor**, *e-tutor**, *electronic tutor**, and *web-based tutor**.

The second stage was tracking and screening meta-analyses or systematic reviews of educational technology in higher education settings published from the year 2000 to 2012. This search intended to catch ITS studies titled under broad educational technology, especially those under the title of CAI and computer-based instruction. We conducted searches in electronic versions of *Review of Educational Research* and *Review of Research in Education.* We also searched in Google Scholar. We identified and screened 18 recent meta-analyses or systematic reviews of the effects of educational technology in higher education.

The third stage consisted of two parts. One was conducting a search with the names of major ITS that were used in higher education settings. By doing this, we thoroughly examined all the publically available research relevant to each major ITS. The other was tracking the publication records of leading ITS researchers.

Finally, we examined reference or bibliography lists of the relevant studies throughout the search and study coding process. Taken together, we identified 35 reports containing 39 studies qualified for our inclusion.

Study Coding

We designed a detailed coding protocol to guide the coding and information retrieval. The coding protocol covered studies' major characteristics. 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. For cases in which a disagreement was not resolved between the two coders, the second author was consulted. For the coding of study outcomes, the first author conducted the coding and then discussed it with the second author.

Effect-Size Calculation

We used Hedges's g, a standardized mean difference between two groups, as the effect-size index for this meta-analysis. The preference for Hedges's g over other standardized-difference indices, such as Cohen's d and Glass's Δ, is due to the fact that Hedges's 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). WWC (2013) adopted Hedges's g as the default effect-size measure for continuous outcomes in its review. Hedges's g was chosen for this meta-analysis because the samples in many ITS studies are small.

Hedges's g was calculated by subtracting the mean of the comparison condition from that of the ITS learning condition and dividing the difference by the average of the two groups' standard deviations. A positive g indicates that students using ITS learned more than their peers in the comparison condition. In cases for which only inference 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).

Two types of effect sizes were extracted. If a study reported only posttest outcomes, we extracted unadjusted effect sizes that did not take into account other variables that might have had an impact on the outcomes. If a study provided outcomes that either adjusted or controlled for other variables (e.g., pretest scores) or reported information that allowed us to do so, we extracted adjusted effect sizes. In some cases, adjusted effect sizes were based upon means and standard deviations of gain scores (i.e., posttests minus pretests), whereas in other cases, they were based upon covariance-adjusted means and standard deviations. For studies that reported descriptive statistics of both pretests and posttests, as suggested by D. B. Wilson (personal correspondence, 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-size variances.

Data Analysis

We used the Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2006) software for data analysis, using independent samples as the unit of analysis 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 that the average effect size will be an estimate of that value. A random-effects model assumes that there is more than one true effect and that the effect sizes included in a meta-analysis are drawn from a population of effects that can have varying values (Cooper, Hedges, & Valentine, 2009). All effect sizes were weighted by inverse variances that were based on sample sizes. Before the analyses, we conducted Grubbs (1950) tests to examine whether there were statistical outliers among the effect sizes and sample sizes. Testing for moderators was conducted to identify variables that might be associated with the effectiveness of ITS (Cooper et al. 2009).

Effect-Size Interpretation

Effect sizes are informative only when they are interpreted with appropriate criteria. Different rules exist for effect-size interpretation. For example, Cohen's (1977) guidelines set an effect size of 0.20 as small, 0.50 as medium, and 0.80 as large. More recently, the WWC (2013) described a set of guidelines to determine the rating for an intervention when combining the findings from multiple studies in WWC reviews. For example, "strong evidence of a positive effect with no overriding contrary evidence" is defined as a "positive effect," and "strong evidence of a negative effect with no overriding contrary evidence" is defined as a "negative effect"; "an effect size of 0.25 standard deviations or larger is considered to be substantively important"; and "an indeterminate effect is one for which the single or mean effect is neither statistically significant nor substantively important" (WWC, 2013, pp. 27–28). In sum, the WWC's intervention rating scheme takes into account the statistical significance, the size of the effect, and the number of studies providing the evidence.

The WWC's intervention rating scheme serves as a good reference for the present meta-analysis. However, the purposes of the WWC reviews and the present meta-analysis are different to some extent. A distinguishing feature of the WWC reviews is that they identify and rate studies of specific intervention. The present meta-analysis is to quantitatively synthesize the overall effectiveness of ITS on college students' academic learning on the basis of all existing empirical research that meets the inclusion criteria. Therefore, we did not use the WWC's intervention rating scheme for the present meta-analysis. Instead, we adopted a *meta-analytic thinking* approach for interpreting the results (cf. Cumming & Finch, 2001; Thompson, 2002, 2006). Particularly, we asked two questions when interpreting the effect sizes: (a) Was the effect size noteworthy? (b) Was the effect size consistent with the related prior literature? In addition, we considered practical significance of the results, along with their magnitude, direction, and statistical significance.

Results

Study Features

The literature search identified 35 reports containing 39 independent studies that met our inclusion criteria. Of the 35 reports, 33 each contained one independent study, one report (VanLehn et al., 2007) contained four studies, and another report (Koch & Gobell, 1999) contained two studies. The studies appeared between 1990 and 2011. Study sample sizes ranged from 20 to 1,066. Each independent study was based on one independent sample. Full descriptions of all the included studies, their major features, and their references are available in the online supplemental materials of this article.

Twenty-two types of ITS were investigated. Four ITS were evaluated in at least three different studies. They were AutoTutor (10 studies), ALEKS (five studies), eXtended Tutor-Expert System (xTex-Sys; five studies), and WISE (three studies). Two ITS were evaluated in two studies. They were Transaction Analysis and Recording Tutor (Johnson, Phillips, & Chase; 2009; Phillips & Johnson, 2011) and Design-Statistics Finder (Koch & Gobell, 1999, containing two studies). Sixteen ITS were evaluated once.

There was great variation in the way ITS were used as interventions in the studies. We coded the intervention conditions into five categories, each indicating different manners or levels of ITS's involvement in the intervention. In 18 studies, the main or only treatment was that students used ITS to learn; we called this condition ITS as principal instruction. In eight studies, ITS were integrated into classroom instruction in which they played an important part, and we called it ITS-integrated class instruction. In five studies, ITS were used to supplement classroom instruction for additional learning after regular classes; we called it ITS-supplemented class instruction. In another five studies, students used ITS for laboratory or exercises that usually took place during class time, and we labeled these ITS-assisted activities. In three studies, students used ITS to do after-school homework assignment, and we called it ITS-assisted homework. Such categorizations allowed us to explore whether ITS's effectiveness differed by the way and the degrees of their uses. We termed these five intervention conditions together as ITS-assisted learning.

ITS-assisted learning was compared to seven different comparison conditions. They were (a) traditional classroom instruction (16 studies); (b) reading printed text (eight studies), including reading textbooks or other hard-copy materials; (c) reading computerized materials (six studies), including computerized minilessons or "canned" texts; (d) CAI (six studies), including using nonadaptive or less intelligent computer systems; (e) self-reliant learning activities (five studies), including laboratory, exercise, or doing homework without help; (f) no-treatment control (four studies); and (g) human tutoring (three studies).

Effect Sizes

We calculated adjusted and unadjusted effect sizes based on the information available in the studies. As we noted in the Method section, unadjusted effect sizes did not take into account other variables that might have had an impact on the outcomes, while adjusted effect sizes were obtained after adjusting or controlling for other variables (e.g., pretest scores). Extracting two types of effect sizes allowed us to make the best use of study information and examine whether ITS's effectiveness differed depending on how it was estimated. Of the 39 independent studies, 24 studies provided information for both adjusted and unadjusted effect sizes, 13 studies provided only unadjusted effect sizes, and two studies provided only adjusted ones. Second, we extracted effect sizes within each study for each comparison so we could examine how the magnitude of the effect differed within one study and how ITS's effectiveness varied by comparison conditions across multiple studies. We formed three sets of effect sizes to address the six main research questions.

Effect sizes by comparison conditions. The first data set consists of 48 effect sizes corresponding to the 48 comparisons in the 39 studies. This data set was used to examine whether the estimate of ITS's effectiveness differed depending on what alternative condition it was compared to. Of the 39 studies, 30 had one comparison condition. For example, Graesser et al. (2003) compared ITS-assisted learning to reading printed text. Nine studies had two comparison conditions. For example, Shute and Glaser (1990) studied ITS's effectiveness in comparison to both traditional classroom instruction and no-treatment control. We extracted one effect size for each comparison condition. We chose an adjusted over unadjusted effect size to represent a study if it provided both types of effect sizes. This led to 33 adjusted and 15 unadjusted effect sizes (see Table 1 in the online supplemental materials). The average adjusted and unadjusted effect sizes did not differ significantly under either a fixed-effect model, $Q_b(1) = 2.51$, $p = .113$, or a random-effects model, $Q_b(1) = 1.84$, $p = .175$. This justified our decision to pool the 48 effect sizes together in this data set. Thirty-nine effect sizes were in a positive direction, and nine were in a negative direction. The effect sizes ranged from −.77 to 1.43. Grubbs (1950) tests detected no effect-size outliers.

For studies with two comparison conditions, the two corresponding effect sizes may not be independent of each other because they both were based on a same sample. However, when effect sizes were grouped by each type of comparison condition, all effect sizes within one type of comparison were independent from each other because each sample contributed one effect to the estimation of that comparison. For example, Shute and Glaser (1990) yielded two effect sizes, one for the comparison of traditional classroom instructions and one for that of no-treatment control. When the effect sizes were grouped by the types of comparison condition, this study contributed one effect size to the group of traditional classroom instructions and one for no-treatment control. Therefore, all effect sizes within each group were from different studies, and they were independent from each other.

Adjusted effect sizes. The second data set consisted of 26 adjusted effect sizes from 26 studies. It was formed as follows: (a) the effect sizes of each of the nine studies with two different comparison conditions were averaged within each study, and this reduced the 48 effect sizes in the first data set to 39, one representing each study; (b) for each of the 24 studies that provided both an adjusted and an unadjusted effect size, we chose the adjusted over unadjusted effect size to represent the study; (c) two studies provided only adjusted effect sizes; and (d) we excluded the 13 studies that provided only unadjusted effect sizes. The 26 effect sizes ranged from −.32 to 1.43. Twenty-three were in a positive direction, and three were in a negative direction. Grubbs (1950) tests detected no outliers.

Unadjusted effect sizes. The third data set consisted of 37 unadjusted effect sizes. It was formed as follows: (a) among the 39 effect sizes from the 39 studies, for each of the 24 studies that provided both an adjusted and an unadjusted effect size, we selected the unadjusted effect size to represent the study; (b) 13 studies provided only unadjusted effect sizes; and (c) we excluded two studies that provided only adjusted ones. The effect sizes in this data set ranged from −.05 to 2.12. Thirty-three were in a positive direction, three were in a negative direction, and one was zero. Grubbs (1950) tests detected the effect size of 2.12 from the third qualified study contained in VanLehn et al. (2007) as a significant outlier ($p < .05$). This effect size was reset to its nearest neighbor, 1.29.

In summary, the three data sets were formed to help address the six main research questions. We used the first data set to study ITS's effectiveness by comparison conditions. We used the second and the third data sets to assess ITS's overall effectiveness, how ITS's effectiveness differed by the types of instruction or learning activities to which they were compared, subject domain, the manner in which ITS were used, the length of ITS instruction, and by different ITS, measured with both adjusted and unadjusted effect sizes. We also explored other potential moderators of ITS's effectiveness. Analyses on the second and the third data sets can be considered an alternative sensitivity analysis.

Overall Effectiveness

Measured with adjusted effect sizes. We conducted meta-analyses on the 26 adjusted effect sizes from the 26 studies to examine the overall effectiveness of ITS on college students' learning (see Table 1). Under a fixed-effect model, the average effect size was $g = .37$, 95% confidence interval (CI) [.24, .50], $p = .000$, and was significantly different from zero. Under a random-effects model, the average effect size was also $g = .37$, 95% CI [.21, .53], $p = .000$, and was significantly different from zero. The 26 effect sizes appeared to be homogeneous, $Q_t(25) = 35.47$, $p = .080$, $I^2 = 29.51$. This suggests that the total variance was largely due to within-study rather than between-study variation.

Measured with unadjusted effect sizes. We also conducted meta-analytical procedures on the 37 unadjusted effect sizes from the 37 studies (see Table 1). Under a fixed-effect model, the average effect size was $g = .32$, 95% CI [.24, .40], $p = .000$, and

Table 1
Overall Effectiveness of Intelligent Tutoring Systems

		Fixed-effect model				Random-effects model				Heterogeneity		
Effect-size type	*k*	*g*	*SE*	95% CI	*p*	*g*	*SE*	95% CI	*p*	Q value	*df* (Q)	*p*
Adjusted	26	.37	.07	[.24, .50]	.000***	.37	.08	[.21, .53]	.000***	35.47	25	.080
Unadjusted	37	.32	.04	[.24, .40]	.000***	.35	.06	[.24, .46]	.000***	56.67	36	.015*

Note. CI = confidence interval.
* $p < .05$. *** $p < .001$.

was significantly different from zero. Under a random-effects model, the average effect size was $g = .35$, 95% CI [.24, .46], $p = .000$, and was also significantly different from zero. The 37 unadjusted effect sizes appeared to be heterogeneous, $Q_t(36) = 56.67$, $p = .015$, $I^2 = 36.50$. This suggests that the total variance could be due to both within- and between-study variation.

Taken together, ITS appeared to have a moderate positive impact on college students' academic learning, with average effect sizes ranging from .32 to .37. The magnitudes of the effectiveness appeared to be quite similar regardless of whether it was measured with adjusted or unadjusted effect sizes. The adjusted effect sizes appeared to be homogeneous across different studies, while unadjusted ones seemed to be heterogeneous, suggesting that adjusted effect sizes could be a better indicator of ITS's effectiveness than unadjusted ones.

Assessing Publication Bias

Among the 26 studies that provided adjusted effect sizes, 18 were journal articles, seven were conference papers, and one was a book chapter. We first produced a funnel plot with each Hedges's g plotted against its standard error. The majority of the studies clustered symmetrically near the mean effect size toward the top of the graph. No study on the left side of the mean was projected as missing. This suggested the absence of publication bias. We further conducted Duval and Tweedie's (2000) trim and fill procedures and found that the adjusted average effect sizes remained the same as the observed ones, which were $g = .37$ under both a fixed-effect and a random-effects model. Therefore, there was no evidence that publication bias had an impact on the effectiveness when measured with adjusted effect sizes.

Among the 37 studies that provided unadjusted effect sizes, 25 were journal articles, nine were conference papers, and three were book chapters. We produced a funnel plot with each Hedges's g plotted against its standard error. The majority of the studies appeared on the right side of the mean effect size and clustered toward the bottom of the graph. Nine studies on the left side of the mean were projected missing. This suggested a presence of publication bias. Statistics from the trim and fill procedures revealed that the overall average effect size after imputing the nine missing values was $g = .22$ under a fixed-effect model and was $g = .23$ under a random-effects model. The observed average effect sizes, as reported previously, were .32 under a fixed-effect model and .35 under a random-effects model. Taken together, the effectiveness might have been slightly overestimated when measured with unadjusted effect sizes.

Grouping the Effectiveness

In addition to the overall effectiveness, we examined whether ITS's effectiveness differed depending on comparison conditions, different ITS, intervention conditions, and subject matter. We drew answers to these issues from results of testing for moderator analyses. Because these issues were relevant to the main research questions of this meta-analysis, we chose to address them with greater narrative details. Table 2 presents the results of grouping the studies by comparison conditions. Tables 3 and 4 show the grouping results by different ITS, intervention condition, and subject matter, measured by adjusted and unadjusted effect sizes, respectively.

Comparison condition. To examine whether ITS's effectiveness differed when compared to different conditions, we grouped the studies by their comparison conditions (see Table 2). Under a fixed-effect model, ITS were more effective than no-treatment control ($g = .90$), self-reliant learning ($g = .82$), reading printed text ($g = .47$), traditional classroom instruction ($g = .37$), computer-assisted learning ($g = .35$), and reading computerized materials ($g = .22$), but not human tutoring ($g = -.25$). The results were essentially unchanged under a random-effects model. With the exception of the cases for computerized materials and

Table 2
Intelligent Tutoring Systems' Effectiveness by Comparison Condition

		Fixed-effect model				Random-effects model			
Comparison condition	k^a	g	SE	95% CI	p	g	SE	95% CI	p
Original comparisons									
Human tutoring	3	−.25	.24	[−.72, .22]	.302	−.25	.24	[−.72, .22]	.302
Reading computerized materials	6	.22	.12	[−.01, .46]	.064	.25	.17	[−.08, .57]	.138
Computer-assisted instruction	6	.35	.12	[.11, .60]	$.004^{**}$	.33	.14	[.06, .60]	$.015^{**}$
Traditional classroom instruction	16	.37	.07	[.24, .50]	$.000^{***}$	.38	.09	[.21, .55]	$.000^{***}$
Reading printed text	8	.47	.10	[.27, .68]	$.000^{***}$	.50	.14	[.22, .78]	$.000^{**}$
Self-reliant learning	5	.82	.19	[.44, 1.20]	$.000^{***}$	.82	.19	[.44, 1.20]	$.000^{***}$
No-treatment control	4	.90	.20	[.52, 1.29]	$.000^{***}$	.90	.20	[.52, 1.29]	$.000^{***}$
Broader comparisons									
Human tutoring	3	−.25	.24	[−.72, .22]	.302	−.25	.24	[−.72, .22]	.302
Instruction learning	36	.37	.05	[.28, .46]	$.000^{***}$	.37	.06	[.25, .49]	$.000^{***}$
Self-reliant learning or no-treatment control	9	.86	.14	[.59, 1.13]	$.000^{***}$	.86	.14	[.59, 1.13]	$.000^{***}$

Note. CI = confidence interval.
[a] The analyses were conducted on the first data set, which included 48 effect sizes from the 39 studies. It is worth noting that, within each group, all effect sizes were independent to each other because each independent study contributed only one effect size to one type of comparison condition. For example, Stankov, Glavinić, and Grubišić (2004) provided one effect size for comparing intelligent tutoring systems with human tutoring and one effect size for comparing intelligent tutoring systems with traditional classroom instructions. Both effect sizes were involved in this process, but they were in two different groups.
** $p < .01$. *** $p < .001$.

Table 3
ITS's Effectiveness by ITS, Intervention, and Subject Measured With Adjusted Effect Sizes

		Fixed-effect model			Random-effects model		
Variable	k^a	g	Q_b	p	g	Q_b	p
ITS			2.65	.618		1.40	.844
Why/AutoTutor	5	.29			.37		
Web Interface for Statistics Education (WISE)	3	.52			.59		
eXtended Tutor-Expert System (xTex-Sys)	4	.20			.20		
Assessment and Learning in Knowledge Spaces (ALEKS)	2	.46			.46		
Other	12	.34			.34		
Intervention condition[b]			8.52	.074		5.95	.203
ITS as principal instruction	14	.26			.30		
ITS-integrated class instruction	4	.29			.29		
ITS-supplemented class instruction	3	.71			.73		
ITS-assisted activities	2	.43			.43		
ITS-assisted homework	3	.21			.21		
Subjects			4.01	.548		2.78	.734
Physics	9	.25			.27		
Statistics	6	.46			.46		
Computer science	5	.31			.31		
Business-related	3	.26			.26		
Mathematical subjects	1	.59			.59		
Other subjects	2	.60			.60		

Note. Q_b denotes the heterogeneity status between all categories of a particular variable. ITS = intelligent tutoring systems.
[a] The analyses were conducted on the second data set. It involved the 26 adjusted effect sizes from 26 studies. [b] Intervention conditions indicate the manners or degrees of ITS's involvement in the interventional instruction or learning activities.

human tutoring, all average effect sizes were statistically significantly different from zero under both models. Taken together, except when compared to human tutoring, ITS had a positive impact on college students' learning relative to all other learning approaches. For all seven comparison conditions, the effects sizes appeared homogeneous within each comparison type.

The average effectiveness of ITS differed by comparison condition under both a fixed-effect model, $Q_b(6) = 21.365$, $p = .002$, and a random-effects model, $Q_b(6) = 19.945$, $p = .003$. We conducted a series of pairwise comparisons between conditions for which the average effect sizes appeared to be relatively similar. Analyses revealed no significant difference between computerized materials and computer-assisted learning. We therefore combined them into computerized learning. Likewise, there was no significant difference between traditional classroom instruction and reading printed text, and we combined them into traditional learning.

Table 4
ITS's Effectiveness by ITS, Intervention, and Subject Measured With Unadjusted Effect Sizes

		Fixed-effect model			Random-effects model		
Variable	k^a	g	Q_b	p	g	Q_b	p
ITS			8.06	.089		6.99	.136
Why/AutoTutor	6	.53			.61		
Web Interface for Statistics Education (WISE)	2	.04			.04		
eXtended Tutor-Expert System (xTex-Sys)	4	.31			.31		
Assessment and Learning in Knowledge Spaces (ALEKS)	5	.30			.30		
Other	20	.34			.36		
Intervention condition			8.25	.083		6.50	.165
ITS as principal instruction	18	.39			.42		
ITS-integrated class instruction	8	.29			.29		
ITS-supplemented class instruction	4	.23			.24		
ITS-assisted activities	4	.62			.64		
ITS-assisted homework	3	.09			.09		
Subjects			8.78	.118		6.73	.242
Computer science	10	.45			.45		
Physics	10	.37			.40		
Statistics	7	.20			.24		
Business-related	4	.16			.16		
Mathematical subjects	3	.65			.65		
Other subjects	3	.29			.32		

Note. Q_b denotes the heterogeneity status between all categories of a particular variable. ITS = intelligent tutoring systems.
[a] The analyses were conducted on the third data set. It involved the 37 unadjusted effect sizes extracted from 37 studies.

Self-reliant learning activities and no-treatment control were combined together for a similar reason. Further analyses revealed no significant difference between computerized learning and traditional learning, and we thus combined them into instruction learning. Therefore, the original seven comparison conditions were transformed into three relatively broader comparisons: human tutoring, instruction learning, and self-reliant learning or no-treatment control.

ITS's effectiveness differed significantly between broader comparisons (also see Table 2). Specifically, the effectiveness appeared to be greater when compared with the condition in which students were engaged in self-reliant learning or no-treatment control than when compared with instruction learning (i.e., the combination of traditional learning and computerized learning), which was higher than when compared with human tutoring. In other words, ITS appeared to be least effective when compared with human tutoring and most effective when compared with students learning through self-reliant activities or no-treatment control.

Different ITS. To examine whether ITS's effectiveness differed among different ITS, we grouped the studies by the five most frequently studied ITS and grouped other miscellaneous ITS (12 studies) together. For adjusted effect sizes (see Table 3), results showed that, under a fixed-effect model, the average effect size was .20 for Why/AutoTutor (five studies), .52 for WISE (three studies), .20 for xTex-Sys (four studies), .46 for ALEKS (two studies), and .34 for other miscellaneous ITS (12 studies). The results remained almost the same under a random-effects model. The average effect sizes of the five groups of ITS did not differ under either a fixed-effect model, $Q_b(4) = 2.65$, $p = .618$, or a random-effects model, $Q_b(4) = 1.40$, $p = .844$.

As Table 4 shows, for unadjusted effect sizes, under a fixed-effect model, the average effect size was .53 for Why/AutoTutor (six studies), .30 for WISE (two studies), .31 for xTex-Sys (four studies), .34 for ALEKS (five studies), and .04 for other miscellaneous ITS (20 studies). The results remained almost the same under a random-effects model (see Table 4). Again, the average effect sizes did not differ under either a fixed-effect model, $Q_b(4) = 8.06$, $p = .089$, or a random-effects model, $Q_b(4) = 6.99$, $p = .136$.

Intervention condition. To examine whether ITS's effectiveness differed depending on how ITS were involved in the interventions, we grouped the studies by intervention conditions. For adjusted effect sizes, under a fixed-effect model, the average effect size was .26 for the category of ITS as principal instruction (14 studies), .29 for ITS-integrated class instruction (four studies), .43 for ITS-assisted activities (two studies), .71 for ITS-supplemented class instruction (three studies), and .21 for ITS-assisted homework (three studies). The results remained almost the same under a random-effects model. The average effect sizes did not differ significantly depending on how ITS were used in the intervention under either a fixed-effect model, $Q_b(4) = 8.52$, $p = .074$, or a random-effects model, $Q_b(4) = 5.95$, $p = .203$.

For unadjusted effect sizes, under a fixed-effect model, the average effect size was .39 for the category of ITS as principle instruction (18 studies), .29 for ITS-integrated class instruction (eight studies), .23 for ITS-supplemented class instruction (four studies), .62 for ITS-assisted other learning (four studies), and .09 for ITS-assisted homework (three studies). The results remained almost the same under a random-effects model. The average effect sizes did not differ significantly depending on how ITS were used in the interventions under either a fixed-effect model, $Q_b(4) = 8.25$, $p = .083$, or a random-effects model, $Q_b(4) = 6.50$, $p = .165$.

Subject matter. To examine whether ITS's effectiveness differed depending on which subjects ITS were used for, we grouped the studies by subject matter. For adjusted effect sizes, under a fixed-effect model, the average effect size was .25 for the learning of physics (nine studies), .46 for statistics (six studies), .31 for computer science (five studies), .26 for business-related subjects (e.g., accounting, economics; three studies), .59 for mathematically related subject (i.e., precalculus; one study), and .60 for other miscellaneous subjects (e.g., psychology, electronic engineering; two studies). The results remained largely unchanged under a random-effects model. ITS's effectiveness did not differ significantly depending on which subjects ITS were used for under either a fixed-effect model, $Q_b(5) = 4.01$, $p = .548$, or a random-effects model, $Q_b(5) = 2.78$, $p = .734$.

For unadjusted effect sizes, under a fixed-effect model, the average effect size was .45 for computer science (10 studies), .37 for physics (10 studies), .20 for statistics (10 studies), .16 for business-related subjects (four studies), .65 for mathematically related subjects (three studies), and .29 for other subjects (three studies). The results remained almost the same under a random-effects model. ITS's effectiveness did not differ significantly depending on which subject ITS were used for under either a fixed-effect model, $Q_b(5) = 8.78$, $p = .118$, or a random-effects model, $Q_b(5) = 6.73$, $p = .242$.

Other Moderators

In addition to the four variables reported above, we conducted tests for 17 other variables that could have possibly significantly impacted the effectiveness. Tables 5 and 6 present the results of four significant moderators in at least one data set (i.e., adjusted or unadjusted) or under one analysis model (i.e., fixed-effect or random-effects model). These led to four noteworthy findings. First, ITS's effectiveness in earlier studies was significantly greater than that in more recent studies. This finding was robust because it was consistent across unadjusted or adjusted effect sizes, under both analysis models. Second, there was some evidence that the situation of teacher involvement (e.g., whether the intervention and comparison groups had the same teachers, different teachers, no teachers, or no information was provided) affected ITS's effectiveness. Third, there was some evidence that ITS's effectiveness was greater in studies that used embedded assessments (e.g., midterms or final exams) with which the content taught during the intervention was measured along with other content taught in a certain period of time (e.g., a semester) than that in studies that used specific assessments (i.e., specifically developed for measuring the instructional content) with which only the content taught during the intervention was measured. Last, when measured with unadjusted effect sizes, significant difference was found between studies that reported effect sizes and those that did not do so, under a fixed-effect model but not under a random-effects model.

None of the 13 remaining variables tested appeared to be statistically significant moderators of the relationship between ITS-

Table 5
Results of Testing for Moderators on Adjusted Effect Sizes

Variable	*k*	Fixed-effect model			Random-effects model		
		g	Q_b	*p*	*g*	Q_b	*p*
Study time			6.78	.034*		5.07	.079
2006–2011	11	.19			.19		
2000–2005	12	.53			.53		
1990s	3	.50			.50		
Study time (further analysis)			6.77	.009**		5.21	.023*
2006–2011	11	.19			.19		
1990–2005	15	.52			.52		
Teacher involvement[a]			9.35	.025*		9.08	.028*
Same teachers	9	.59			.59		
Different teachers	4	.03			.03		
No teachers	5	.21			.21		
Not given	8	.32			.37		
Teacher involvement (further analysis 1)			7.30	.007**		7.30	.007**
Same teachers	9	.59			.59		
Different teachers	4	.03			.03		
Teacher involvement (further analysis 2)			0.58	.446		0.58	.446
Different teachers	4	.03			.03		
No teachers	5	.21			.21		
Teacher involvement (further analysis 3)			4.09	.043*		4.09	.043*
Same teachers	9	.59			.59		
No teachers	5	.21			.21		
Teacher involvement (further analysis 4)			8.49	.004**		8.49	.004**
Same teachers	9	.59			.59		
Different teachers or no teachers	9	.13			.13		
Assessment type[b]			3.90	.048*		2.57	.109
Specific	21	.29			.32		
Embedded	5	.57			.57		
Effect-size reporting[c]			0.94	.332		0.86	.355
Yes	6	.53			.57		
No	20	.34			.33		

Note. Q_b denotes the heterogeneity status between all categories of a particular variable.
[a] Teacher involvement refers to the circumstances in which both the intervention and comparison groups had the same teachers, different teachers, or no teachers involved, or no teacher information was given. [b] Assessment type refers to the test types of the posttests. Specifically, embedded assessment denotes the type of assessment in which the target test contents were incorporated into a broader assessment, such as midterm or final exam; specific assessment denotes assessment tools entirely composed of target test contents specifically developed for measuring the instructional content. [c] Effect-size reporting refers to whether the study reported effect sizes of interest to this meta-analysis.
* $p < .05$. ** $p < .01$.

assisted learning and student learning gains. A list of the 13 variables, their definitions, specific categories, and the testing results are available in Tables 2 and 3 in the online supplemental materials. Although testing of these variables did not lead to any significant findings, it did show what this meta-analysis had explored in searching for moderators of the effectiveness. Therefore, it provided a foundation and inspiration for future research.

Summary

Taken together, results of this meta-analysis revealed answers to the six main research questions. Overall, ITS had a moderate positive effect on college students' learning, with average effect sizes ranging from $g = .32$ to $.37$. When the effectiveness was measured after controlling for the influence of other variables (e.g., pretest scores), the average adjusted effect size was $g = .32$ under a fixed-effect model and was $g = .35$ under a random-effects model; both effects were significantly different from zero, favoring ITS-assisted learning over its comparisons. When measured without taking into account the influence of other factors, the average unadjusted effect size was $g = .37$ and was significantly different from zero, also favoring ITS-assisted learning over its comparisons under both analysis models. ITS's effectiveness appeared to be largely similar regardless of whether it was measured with adjusted or unadjusted effect sizes and regardless of the analysis model. However, the adjusted effect sizes appeared to be homogeneous across different studies, while unadjusted ones seemed to be heterogeneous.

There was evidence that ITS's effectiveness did not differ significantly by the different ITS, by subject matter, by the length of ITS instruction, or by how they were used and the level of their involvement in instructions or learning activities. However, significant differences appeared when comparison conditions differed. Specifically, ITS had a significant positive impact on college students' learning relative to both instruction learning (i.e., both traditional and computerized instruction learning; $g = .37$, $p = .000$) and self-reliant learning activities or no-treatment control ($g = .86$, $p = .000$). However, when compared to human tutoring, ITS appeared to have a nonsignificant negative impact ($g = -.25$, $p = .302$) on students' learning. Further analyses revealed a ranking as follows: human tutoring > ITS-assisted learn-

Table 6
Results of Testing for Moderators on Unadjusted Effect Sizes

		Fixed-effect model			Random-effects model		
Variable	*k*	*g*	Q_b	*p*	*g*	Q_b	*p*
Study time			13.76	.001**		11.39	.003**
2006–2011	16	.26			.27		
2000–2005	14	.25			.29		
1990s	7	.69			.69		
Study time (further analysis 1)			0.01	.919		0.02	.889
2006–2011	16	.26			.27		
2000–2005	14	.25			.29		
Study time (further analysis 2)			13.75	.000***		11.61	.001**
2000–2011	30	.25			.27		
1990s	7	.69			.69		
Teacher involvement			2.71	.438		2.01	.570
Same teachers	11	.26			.28		
Different teachers	5	.29			.29		
No teachers	10	.33			.35		
Not given	11	.45			.51		
Assessment type			0.04	.843		0.05	.819
Specific	26	.31			.35		
Embedded	11	.33			.33		
Effect-size reporting			4.37	.037*		2.68	.102
Yes	9	.49			.49		
No	28	.27			.31		

Note. Q_b denotes the heterogeneity status between all categories of a particular variable.
* $p < .05$. ** $p < .01$. *** $p < .001$.

ing > instruction learning > self-reliant learning activities or no-treatment control.

Discussion

Human Tutoring Outperforms ITS: The Story Continues

The finding that ITS was least effective relative to human tutoring appears to be consistent with findings in many previous studies. We should be quick to point out that the results were not statistically significant (see Table 2) and that there were only three studies that compared ITS-assisted learning with human tutoring (i.e., Reif & Scott, 1999; Stankov, Glavinić, & Grubišić, 2004; VanLehn et al., 2007, Study 1) in the current meta-analysis. Bloom (1984) reported that expert human tutors can help students achieve learning gains as large as two sigmas. 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 $d = .79$ on students' learning. Many ITS are designed to follow the practices of human tutors (Graesser et al., 2011; Woolf, 2009). It is a common practice to compare ITS with human tutoring, considered the highest performance standard in fostering learning. VanLehn (2011) reviewed randomized experiments that compared the effectiveness of human tutoring, computer tutoring, and no tutoring. This review is particularly relevant to the current meta-analysis. Comparing these two reviews reveals many nuances in the issue of ITS's effectiveness.

Specifically, VanLehn's (2011) meta-review covered studies that compared at least two types of five different instructions: human tutoring, two types of ITS, (i.e., substep-based and step-based tutoring systems), other less intelligent computer tutoring systems (e.g., CAI or CBI), and conventional instruction without tutoring (i.e., typically a combination of text reading and problem solving without feedback). He generated two integrated effect sizes that are particularly relevant to the findings of the current meta-analysis: (a) an effect size of $d = .79$, indicating that human tutoring outperformed conventional instruction, and (b) an effect size of $d = .71$, indicating that ITS also outperformed conventional instruction. He concluded that (a) human tutoring was not as effective as what previous research had found and (b) ITS were almost as effective as human tutoring. The current meta-analysis found a mean effect size of $g = .37$ when ITS-assisted learning was compared to instruction learning (i.e., the combination of traditional and computerized instruction) and uncovered a ranking of human tutoring > ITS-assisted learning > instruction learning > self-reliant learning activities or no-treatment control.

VanLehn (2011) found that the comparisons between human tutoring and step-based tutoring yielded a mean effect size of $d = .21$ and that the comparisons between human tutoring and substep-based tutoring yielded $d = -.12$. If reversing the direction of the comparison by designating ITS as the experimental condition and human tutoring as a control condition, as was the case in the current meta-analysis, the effect sizes became $-.21$ and .12, respectively. The current meta-analysis aggregated all types of ITS and did not distinguish step-based from substep-based ITS. To tentatively compare the results of the current meta-analysis with the above results from VanLehn, we averaged the above two effect sizes to obtained a $d = -.05$, indicating that human tutoring outperformed ITS. As reported earlier, the current meta-analysis found an effect size of $g = -.25$. Taken together, one could tentatively conclude that the findings of the current meta-analysis are largely consistent with the findings of VanLehn's (2011) meta-review for ITS compared with human tutoring.

It is relevant to note that these two systematic reviews are different in three major ways. First, the two reviews differ in subject domains and grade levels. VanLehn's (2011) review included studies of STEM subjects, with no restriction of grade levels. As a result, it included a large portion of studies of ITS's use in K-12 and professional students' learning. Second, the two reviews had different methodological standards and applied different study inclusion criteria. VanLehn 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. The above two differences might have led to the result that only six overlapping studies are found in these two reviews. Last, VanLehn's review 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 are worth noting for an appropriate interpretation of the findings from these two reviews. Nonetheless, both reviews suggest there is still a lot for ITS developers to learn from human tutors. However, given the revealed effect, ITS will continue to be one important alternative resource for fostering learning.

It is worth noting that the effectiveness of ITS is an issue of significant importance because the development, research, and use of ITS involve multiple disciplines, including computer science, artificial intelligence, psychology, and education and consume substantial financial, intellectual, and educational resources. For example, Woolf and Cunningham (1987) reported that each hour of ITS instruction costs 200 hours of work to build. One may still argue that human tutoring is likely to be far more expensive than ITS. The complicated nature of this issue precludes a cost-effectiveness argument of ITS versus human tutoring. at least for now. However, the meaning of ITS is more than simply their past or current performance in helping students' learning and their comparison with human tutoring or other instruction approaches. Rather, it seems reasonable to expect that ITS can be instrumental in advancing the learning sciences. For example, ITS can be used to test specific tutoring and learning hypotheses that cannot be examined consistently with human instructors. As Lane (2006) pointed out, ITS's strength is particularly evident in the study of feedback in that ITS allow researchers to experimentally adjust the content, form, and frequency of feedback and therefore test for learning differences. ITS can provide adaptive support for a number of educational activities, such as problem solving, studying examples, exploring interactive simulations, and playing educational games (Conati, 2009). Research communities, such as the Pittsburg Science of Learning Center, have tapped ITS to address a wide range of learning issues.

Do ITS Affect College and K-12 Students Differently?

The current meta-analysis and Steenbergen-Hu and Cooper's (2013) meta-analysis of ITS's effectiveness on K-12 students' mathematical learning both found a positive impact of ITS on students' learning, although the magnitude of the effectiveness appear to be differential. The overall average effect sizes uncovered in the current meta-analysis appeared to be larger than those revealed in the latter. The current meta-analysis found that the overall effectiveness of ITS ranged from .32 to .37, whereas the latter found effect sizes ranging from .01 to .09. There are three possible explanations for this difference. First, some differences in the characteristics of interventions and study methodological features might be responsible. In the current meta-analysis, ITS were used for a short time in the majority of studies studying ITS that were relatively less known and less widely used in real educational settings, these studies were more often based on small sample sizes, they used less rigorous research methods, and they often used specifically designed or nonstandardized outcome measures. In contrast, in Steenbergen-Hu and Cooper's (2013) meta-analysis, ITS were used for one semester, one school year, or longer in the majority of the studies. These studies, such as those of Cognitive Tutors, were more often based on large national samples, and they employed more rigorous study methods, such as random assignment, and used more distal outcome measures, such as standardized achievement tests (e.g., Campuzano, Dynarski, Agodini, & Rall, 2009; Dynarski et al., 2007). There is evidence that the differences mentioned above have an impact on the magnitude of effect sizes (Cheung & Slavin, 2012; Sosa et al., 2011).

Second, the degree of intervention implementation may contribute to the difference. In the current meta-analysis, a large number of ITS were used in laboratories for experimental purposes rather than in real learning environments. For example, 10 of the 26 adjusted effect sizes were associated with laboratories, 15 were associated with real environments, and one was from both. In contrast, in the latter, the majority of the ITS were widely used in real educational settings. For example, 20 of the 34 included studies assessed the effects of Cognitive Tutors, which were used in over 2,600 schools in the United States as of 2010 (WWC, 2010a). It is possible that educational interventions in laboratory environments usually produce larger effects than real environments where researchers usually have no or very little involvement in the implementation of the interventions.

Last, it is possible that ITS's effectiveness differs as the users' age or educational levels differ. The current meta-analysis focused on studies of ITS's impact on college students' learning, while the latter focused on ITS's influence on K-12 students' mathematical learning. It is likely that ITS may function better for more mature students who have sufficient prior knowledge, self-regulation skills, learning motivation, and experiences with computers than for younger students who may still need to develop the above characteristics and need more human inputs to learn. This hypothesis needs to be tested in future research.

Two additional differences between these two meta-analyses are noteworthy. First, the two meta-analyses covered studies in which ITS were compared to different types of instruction or learning activities. The overall effect sizes in the current meta-analysis were drawn from three broad comparison conditions: human tutoring, instruction learning, and self-reliant activities or no-treatment control. Steenbergen-Hu and Cooper (2013) generated the overall effect sizes from 31 studies that compared ITS with regular classroom instruction. Second, there was a difference in the types of subject matters. The current meta-analysis placed no restriction on subject matters and covered subjects from physics, statistics, computer science, accounting, economics, psychology, and electronic engineering. Steenbergen-Hu and Cooper focused on ITS's effectiveness on K-12 students' mathematical learning (i.e., basic math, algebra, and geometry). The average effect sizes of statistics or mathematical subjects in the current meta-analysis ranged from .20

to .65 and appear to be greater than the overall effect size in the latter.

How Do ITS Perform Relative to Educational Technologies in General?

The conclusion that ITS had a moderate positive effect on college students' academic learning is largely congruent with the findings of several recent systematic syntheses of the impact of educational technology in higher education settings. In particular, it is consistent with those of Sosa et al.'s (2011) meta-analysis of the effectiveness of CAI on statistics in at least three ways. First, both meta-analyses found that technology-enabled learning outperformed traditional classroom instruction, and the average effect sizes were quite similar. Specifically, Sosa et al. found an average performance advantage of CAI over lectured-based traditional statistics instruction (d = .33), and the current meta-analysis found ITS outperformed traditional classroom instruction with a similar effect (g = .37). They found that stand-alone tools or tutorials, of which some can be considered to be ITS, showed a performance advantage over traditional statistics instruction, as indicated by an average effect size of d = .43, which is still quite close to what the current meta-analysis found. Second, the current meta-analysis confirmed Sosa et al.'s finding that, although not consistently across different types of effect sizes or analysis models, studies using embedded assessments yielded a larger average effect size than those using specific assessments. Last, zooming in on ITS's effectiveness on statistics, which was the target subject of Sosa et al., the current meta-analysis found similar results (i.e., effect sizes ranging from .20 to .46) as in Sosa et al. (d = .43 for stand-alone tools or tutorials). In summary, the current meta-analysis supports Sosa et al.'s (2011) conclusion that computer-based tools (including stand-alone tutorials) outperformed traditional instruction in helping students' learning of statistics by making the case that very similar findings were obtained when the particular type of computer-based tools was ITS.

Likewise, findings of this meta-analysis are consistent with what Tamim, Bernard, et al. (2011) found in a second-order meta-analysis to summarize 40 years of research on the effects of technology-enhanced instruction on students' achievement relative to more traditional types of classrooms without technology. They identified 25 meta-analyses encompassing 1,055 primary studies. They found that the weighted mean effect size was .35 and was significantly different from zero under a random-effects model. Furthermore, they found that the 11 meta-analyses that focused on postsecondary levels generated a mean effect size of .29, which was significantly different from zero. Although the types of technologies and comparisons in the two reviews were different to some degrees, once again, the current meta-analysis makes a case that ITS's effectiveness on college students' learning falls within the general range of educational technologies' impact on students' learning in higher education.

Comparing this meta-analysis with two other meta-analyses of the effectiveness of CAI on college students' statistics learning leads to quite similar conclusions as above. For example, based on 47 studies of the effectiveness of technology use in statistics instruction in higher education, Schenker's (2007) meta-analysis found the mean achievement effect size was d = .44. Likewise, integrating 25 primary studies that compared CAI with traditional methods in teaching statistics for undergraduate and graduate students, Hsu (2003) found an overall effect size of .43.

In summary, it is not surprising to find that ITS had a moderate positive effect on college students' academic learning. This finding fits into the big picture concerning the impact of educational technology in higher education that many previous meta-analyses have collectively created.

What Moderates ITS's Impact on Learning?

One noteworthy finding from testing for moderators is that ITS's effectiveness in earlier studies was significantly greater than in more recent studies. Previous research syntheses that explored the relationship of the time of the study and the magnitude of intervention effects produced mixed results. For example, Steenbergen-Hu and Cooper (2013) found that, when measured by unadjusted effect sizes, under a fixed-effect model, ITS's effectiveness on K-12 students' mathematical learning was significantly greater in studies in which data were collected before the year 2003 than in studies between 2003 and 2010. However, this was not the case under a random-effects model or when the effectiveness was measured by adjusted effect sizes. Cheung and Slavin's (2012) meta-analysis of how features of educational technology affected student reading achievement grouped the studies by the year of publication into four decade intervals: 1980s, 1990s, 2000s, and 2010s. They found no trend toward more or less positive results in recent years. It could be argued that categorically breaking the time of study or data collection is often subjective or arbitrary. It is possible that different groupings of the time (e.g., Cheung & Slavin, 2012, could have grouped 1980s and 1990s together and done the same with the 2000s and 2010s) could lead to different results. Therefore, it is best to think of such efforts as exploratory and consider their results as tentative, as with many cases when testing for moderators. Nevertheless, it is worth mentioning that the finding in the current meta-analysis was robust because it was consistent across both types of effect sizes, under both analysis models.

The findings also underline the importance of teachers and pedagogy and point to their possible relationship with the effectiveness of ITS. Researchers ought to continue this line of inquiry and study how teachers play a role in ITS-assisted learning and how they impact the end results. An example of such studies is Tamim, Lowerison, Schmid, Bernard, and Abrami's (2011) multiyear investigation of the relationship between pedagogy, computer use, and postsecondary students' perceptions about course effectiveness over time. They found that course structure (e.g., whether or how the course was learner centered), active learning, and computer use were predictive of students' perceived course effectiveness, with course structure being the most predictive over time. Like the current meta-analysis, this study also points to the importance of teachers and pedagogy in technology-enabled learning.

Future ITS Research and Development

This meta-analysis leads to three points concerning the future of ITS research and development. First, expanding ITS to open-ended or ill-defined domains deserves particular attention. Along with previous research syntheses, the current meta-analysis revealed that extensive work has been done on well-defined domains, such as mathematics, algebra, physics, statistics, and computer programming, where the boundaries between right and wrong are clear and straightforward. However, relatively little has been done on open-ended or ill-defined domains, such as history. Moving beyond well-defined domains requires ITS being able to model domains, student behaviors, and their mental states that are usually not as structured or well defined as those involved in earlier ITS's functions in educational activities, such as problem solving. However, if ITS are able to do so, they could serve many more educational purposes.

Second, research is needed to examine how teachers and pedagogy play a role in ITS-assisted learning and how they impact learning along with ITS. This meta-analysis found that teacher involvement might have significantly influenced ITS's effectiveness, although this finding was not quite robust. Furthermore, although no significant difference was found in the effectiveness depending on how ITS were used and the levels of their involvement in learning, this meta-analysis underlines the role of pedagogy in ITS-assisted learning. These findings support Cherniavsky and Vanderputten (2003)'s argument that further research is needed on the roles teachers play in ITS-enabled learning environments and on the classroom organizations in which ITS were used. This line of inquiry could be fruitful.

Last, ITS hold great potential in enhancing self-regulated learning. Using ITS as metacognitive tools to enhance learning has attracted increased attention in recent years. For example, Chi and VanLehn (2010) studied metacognitive strategy instruction with ITS (i.e., target variable strategy, a domain-independent problem-solving strategy) for college students. They found that strategy instruction closed the learning gap between the learners who were less sensitive to learning environments and those who were more sensitive. Many questions surrounding self-regulation and ITS-assisted learning are worth exploring. For instance, what role does self-regulation play when students learn with ITS? What is the dynamic nature of self-regulated learning when students learn with ITS? How can ITS determine if a learner is a high or low self-regulating learner, and what effects will this determination have on learners' subsequent self-regulation? Furthermore, what types and levels of scaffolding strategies should ITS provide for high or low self-regulating learners? The last two questions are particularly important because research has found that high and low self-regulating learners usually exhibit different learning characteristics (Pintrich, 2000, 2004). Meanwhile, using ITS to enhance self-regulated learning poses great challenges in future ITS design. One of the challenges is how ITS detect, trace, and model the multiple phases and processes of self-regulated learning and foster self-regulatory skills.

Conclusion

With what this meta-analysis has found, answers to several questions concerning ITS's effectiveness on learning have become clearer, although one ought to bear in mind that this meta-analysis is based on a limited number of studies. As an advanced learning technology, ITS have demonstrated their ability to outperform many instructional methods or learning activities in facilitating college students' learning of a wide range of subjects, although they are not yet as effective as human tutors. ITS appear to have a more pronounced effect on college-level learners than on K-12 students. There is no evidence yet suggesting that any ITS are significantly better than others or that they work better for one subject domain than for another. It is also not clear what, if any, substantive factors significantly moderate the effectiveness of ITS-assisted learning. How well ITS can do might depend on the teachers and pedagogical strategies used in a particular learning environment. However, having shown a consistent positive effect across various circumstances, ITS can be used either as a stand-alone principal instruction, to be part of regular classroom instruction, or as a supplemental learning tool. With what they have accomplished so far, ITS certainly add one more good choice to the array of educational technologies available to educators and students.

References

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

*Aberson, C. L., Berger, D. E., Healy, M. R., Kyle, D. J., & Romero, V. L. (2000). Evaluation of an interactive tutorial for teaching the central limit theorem. *Teaching of Psychology, 27,* 289–291. doi:10.1207/S15328023TOP2704_08

*Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2002). An interactive tutorial for teaching statistical power. *Journal of Statistics Education, 10*(3). Retrieved from www.amstat.org/publications/jse/v10n3/aberson.html

*Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2003). Evaluation of an interactive tutorial for teaching hypothesis testing concepts. *Teaching of Psychology, 30,* 75–78. doi:10.1207/S15328023TOP3001_12

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

Aleven, V., McLaren, B., 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.

*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

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.

*Bliwise, N. G. (2005). Web-based tutorials for teaching introductory statistics. *Journal of Educational Computing Research, 33,* 309–325. doi:10.2190/0D1J-1CE1-5UXY-3V34

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. doi:10.3102/0013189X013006004

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

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts* (ED-01CO0039/0007). Retrieved from http://ies.ed.gov/ncee/pubs/20094041/pdf/20094042.pdf

*Chang, K.-E., Sung, Y.-T., Wang, K.-Y., & Dai, C.-Y. (2003). Web_Soc: A Socratic-dialect-based collaborative tutoring system on the World Wide Web. *IIEE Transactions on Education, 46,* 69–78.

Cherniavsky, J. C., & Vanderputten, E. (2003, April). *The controversy of technology in education*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Cheung, A., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7,* 198–215. doi:10.1016/j.edurev.2012.05.002

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanation improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M., & VanLehn, K. (2010). Meta-cognitive strategy instruction in intelligent tutoring systems: How, when, and why. *Educational Technology & Society, 13,* 25–39.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

Conati, C. (2009, July). *Intelligent tutoring systems: New challenges and directions*. Paper presented at the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), Pasadena, CA.

*Conati, C., Muldner, K., & Carenini, G. (2006). From example studying to problem solving via tailored computer-based meta-cognitive scaffolding: Hypotheses and design. *Technology, Instruction, Cognition and Learning, 4,* 1–52.

*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.), *Intelligent tutoring systems: 5th International Conference* (pp. 304–313). Montreal, Quebec, Canada. doi:10.1007/3-540-45108-0_34

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.), *User modeling: Proceedings of the Eighth International Conference, UM 2001* (pp. 137–147). Berlin, Germany: Springer-Verlag.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

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* (ED-01-CO-0039/0007). Retrieved from 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.

*Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., . . . Tutoring Research Group. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (pp. 1–6). Mahwah, NJ: Erlbaum.

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

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & Tutoring Research Group, University of Memphis. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 35–51. doi:10.1016/S1389-0417(99)00005-4

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21*(1), 27–58.

*Grubišić, A., Stankov, S., & Hrepic, Z. (2008, November). *Comparing the effectiveness of learning content management systems to intelligent tutoring systems*. Paper presented at the IASK International Conference, Madrid, Spain.

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

*Grubišić, A., Stankov, S., & Žitko, B. (2006, September). *An approach to automatic evaluation of educational influence*. Paper presented at the 6th WSEAS International Conference on Distance Learning and Web Engineering, Lisbon, Portugal.

*Hagerty, G., & Smith, S. (2005). Using the Web-based interactive software ALEKS to enhance college algebra. *Mathematics and Computer Education, 39,* 183–194.

*Hampikian, J., Guarino, J., Chyung, S. Y., Gardner, J., Moll, A., Pyke, P., & Schrader, C. (2007). *Benefits of a tutorial mathematics program for engineering students enrolled in precalculus: A template for assessment*. Retrieved from http://www.icee.usm.edu/ICEE/conferences/asee2007/papers/1998_BENEFITS_OF_A_TUTORIAL_MATHEMATICS_PROGR.pdf

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 Associates.

*Heyden, D. C. (1990). A *DSM-III-R* computer tutorial for abnormal psychology. *Teaching of Psychology, 13,* 203–206. doi:10.1207/s15328023top1703_21

Hsu, Y. C. (2003). *The effectiveness of computer-assisted instruction in statistics education: A meta-analysis* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3089963)

*Hu, X., Luellen, J. K., Okwumabua, T. M., Xu, Y., & Mo, L. (2007, April). *Observational findings from a Web-based intelligent tutoring system: Elimination of racial disparities in an undergraduate behavioral statistics course*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, IL.

*Johnson, B. G., Phillips, F., & Chase, L. G. (2009). An intelligent tutoring system for the accounting cycle: Enhancing textbook homework with artificial intelligence. *Journal of Accounting Education, 27,* 30–39. doi:10.1016/j.jaccedu.2009.05.001

*Koch, C., & Gobell, J. (1999). A hypertext-based tutorial with links to the Web for teaching statistics and research methods. *Behavior Research Methods, Instruments & Computers, 31,* 7–13. doi:10.3758/BF03207686

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 Intelligent in Education, 8,* 30–43.

Lane, H. C. (2006, August). *Intelligent tutoring systems: Prospects for guided practice and efficient learning*. Paper presented at the Army's Science of Learning Workshop, Hampton, VA.

*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

*Livergood, N. D. (1994). A study of the effectiveness of a multimedia intelligent tutoring system. *Journal of Educational Technology Systems, 22,* 337–344.

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

Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction, 22,* 39–72. doi:10.1007/s11257-011-9105-9

*Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database language. *International Journal of Artificial Intelligence in Education, 10,* 238–256.

*Morris, E. (2001). The design and evaluation of Link: A computer-based learning system for correlation. *British Journal of Educational Technology, 32,* 39–52. doi:10.1111/1467-8535.00175

Ohlsson, S. (1992). Constraint-based student modeling. *International Journal of Artificial Intelligence in Education, 3,* 429–447.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103,* 241–262. doi:10.1037/0033-295X.103.2.241

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

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). New York, NY: Academic Press. doi:10.1016/B978-012109890-2/50043-3

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16,* 385–407. doi:10.1007/s10648-004-0006-x

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

Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. *Supporting Learning Flow Through Integrative Technologies, 162*(1), 13–20.

*Rosé, C. P., Bhembe, D., Siler, S., Srivasteva, R., & VanLehn, K. (2003). Exploring the effectiveness of knowledge construction dialogues. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies* (pp. 497–499). Amsterdam, the Netherlands: IOS Press.

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

Schenker, J. D. (2007). *The effectiveness of technology use in statistics instruction in higher education: A meta-analysis using hierarchical linear modeling* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3286857)

*Shute, V. J., & Glasser, 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* (ETS Research Report RR-07–05). Princeton, NJ: Educational Testing Service.

Sleeman, D., & Brown, J. S. (Eds.). (1982). *Intelligent tutoring systems*. Orlando, FL: Academic Press.

Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research, 81,* 97–127. doi:10.3102/0034654310378174

*Stankov, S., Glavinić, V., & Grubišić, A. (2004, April). *What is our effect size: Evaluating the educational influence of a Web-based intelligent authoring shell?* Paper presented at the IEEE International Conference on Intelligent Engineering Systems 2004–INES 2004, Cluj-Napoca, Romania.

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

*Stylianou, D. A., & Shapiro, L. (2002). Revitalizing algebra: The effect of the use of a cognitive tutor in a remedial course. *Journal of Educational Media, 27,* 147–171. doi:10.1080/1358165022000081404

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

Tamim, R. M., Lowerison, G., Schmid, R. F., Bernard, R. M., & Abrami, P. C. (2011). A multi-year investigation of the relationship between pedagogy, computer use and course effectiveness in postsecondary education. *Journal of Computing in Higher Education, 23,* 1–14. doi:10.1007/s12528-010-9041-4

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–31. doi:10.3102/0013189X031003025

Thompson, B. (2006). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (pp. 583–603). Mahwah, NJ: Erlbaum.

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., & Carolyn, P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3–62. doi:10.1080/03640210709336984

VanLehn, K., Jordan, P., Rosé, C. P., Behmbe, C., Böttner, D., Gaydos, M., . . . Srivastava, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent tutoring systems: 6th International Conference* (pp. 158–167). Berlin, Germany: Springer.

*VanLehn, K., Van de Sande, B., Shelby, R., & Gershman, S. (2010). The Andes Physics Tutoring System: An experiment in freedom. *Studies in Computational Intelligence, 308,* 421–443. doi:10.1007/978-3-642-14363-2_21

*Wang, H.-C., Li, T.-Y., & Chang, C.-Y. (2006). A Web-based tutoring system with styles-matching strategy for spatial geometric transformation. *Interacting With Computers, 18,* 331–355. doi:10.1016/j.intcom.2005.11.002

What Works Clearinghouse. (2004, December). *What Works Clearinghouse topic report: Curriculum-based interventions for increasing K-12 math achievement-middle school*. Retrieved from 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 http://www.aea9.k12.ia.us/documents/filelibrary/pdf/cognitive_tutor/WWC_Cognitive_Tutor_052907_3B8688D14AA44.pdf

What Works Clearinghouse. (2009, July). *WWC intervention report middle school math: Cognitive Tutor Algebra I.* Retrieved from 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 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 http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=378

What Works Clearinghouse. (2013). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf

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: Morgan Kaufman Publishers.

Woolf, B. P., & Cunningham, P. A. (1987). Multiple knowledge sources in intelligent teaching systems. *IEEE Expert, 2*(2), 41–54. doi:10.1109/MEX.1987.4307063

*Xu, Y. J., Meryer, K. A., & Morgan, D. D. (2009). A mixed-methods assessment of using an online commercial tutoring system to teach introductory statistics. *Journal of Statistics Education, 17,* 1–17. Retrieved from http://www.amstat.org/publications/jse/v17n2/xu.pdf

Received January 11, 2013
Revision received April 4, 2013
Accepted July 29, 2013 ■