Journal of Educational Psychology

0022-0663/14/$12.00 http://dx.doi.org/10.1037/a0037123

Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis

Wenting Ma
Simon Fraser University

Olusola O. Adesope
Washington State University

John C. Nesbit and Qing Liu
Simon Fraser University

Intelligent Tutoring Systems (ITS) are computer programs that model learners' psychological states to provide individualized instruction. They have been developed for diverse subject areas (e.g., algebra, medicine, law, reading) to help learners acquire domain-specific, cognitive and metacognitive knowledge. A meta-analysis was conducted on research that compared the outcomes from students learning from ITS to those learning from non-ITS learning environments. The meta-analysis examined how effect sizes varied with type of ITS, type of comparison treatment received by learners, type of learning outcome, whether knowledge to be learned was procedural or declarative, and other factors. After a search of major bibliographic databases, 107 effect sizes involving 14,321 participants were extracted and analyzed. The use of ITS was associated with greater achievement in comparison with teacher-led, large-group instruction ($g = .42$), non-ITS computer-based instruction ($g = .57$), and textbooks or workbooks ($g = .35$). There was no significant difference between learning from ITS and learning from individualized human tutoring ($g = -.11$) or small-group instruction ($g = .05$). Significant, positive mean effect sizes were found regardless of whether the ITS was used as the principal means of instruction, a supplement to teacher-led instruction, an integral component of teacher-led instruction, or an aid to homework. Significant, positive effect sizes were found at all levels of education, in almost all subject domains evaluated, and whether or not the ITS provided feedback or modeled student misconceptions. The claim that ITS are relatively effective tools for learning is consistent with our analysis of potential publication bias.

Keywords: Intelligent Tutoring System, student model, effect size, meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/a0037123.supp

In 1970, computer scientist Jaime Carbonell published a report on SCHOLAR, a program he designed to conduct limited, mixed-initiative, instructional dialogues with a student about South American geography (Carbonell, 1970). SCHOLAR used natural language to answer a learner's question or pose a question and give feedback about the correctness of the learner's response. Although the term Intelligent Tutoring System (ITS) was not used in Carbonell's article, SCHOLAR is often regarded as the first ITS (Corbett, Koedinger, & Anderson, 1997). Beyond the mixed-initiative dialogue, what was remarkable about SCHOLAR was the way its architecture represented domain knowledge separately from the natural language interface. The separate, explicit domain representation allowed the program, in theory, to generate a diverse and combinatorally large set of questions and answer a similarly large and diverse set of questions posed by the learner. Framing his work as an extension and application of research in artificial intelligence, Carbonell emphasized the fundamental differences between SCHOLAR and the other types of computer-assisted instruction being designed at the time. In particular, he discussed how a domain representation can serve as the basis for modeling student knowledge.

BIP, another early example of an ITS (Barr, Beard, & Atkinson, 1976) assigned programming tasks to students that matched their individual learning needs and competencies. The BIP researchers constructed a domain representation that mapped goal skills (e.g., printing variables) to the programming tasks that exercised them. Students' performances on a task supported inferences about their acquisition of skills linked to that task. In this early ITS, like many that have been designed since, the student model was an overlay or subset of the domain model. By the time a special issue on Intelligent Tutoring Systems appeared in the *International Journal of Man-Machine Studies* (Sleeman & Brown, 1979) it was clear that a new type of instructional system and a new field of research

Wenting Ma, Faculty of Education, Simon Fraser University; Olusola O. Adesope, Educational Psychology Program, Department of Educational Leadership, Sport Studies, and Educational/Counseling Psychology, College of Education, Washington State University; John C. Nesbit and Qing Liu, Faculty of Education, Simon Fraser University.

Support for this research was provided by a grant from the Social Sciences and Humanities Research Council of Canada (to John C. Nesbit) and Washington State University's College of Education Faculty Funding Award (to Olusola O. Adesope).

Correspondence concerning this article should be addressed to Olusola O. Adesope, Department of Educational Leadership, Sport Studies, and Educational/Counseling Psychology, College of Education, Washington State University, Cleveland Hall 356, Pullman, WA 99164-2114. E-mail: olusola.adesope@wsu.edu

had emerged. Almost all the articles in the special issue and in the associated book (Sleeman & Brown, 1982) were centrally concerned with student modeling.

As indicated in Figure 1, scholarly interest in ITS has grown significantly since 1980. However, only since about 1997 have a significant number of evaluative studies been published which compared the learning outcomes of students using ITS with those using other instructional methods. Research evaluating the instructional efficacy of ITS has been conducted from elementary to postsecondary levels in a wide range of knowledge domains including algebra (Koedinger, Anderson, Hadley, & Mark, 1997), physics (Albacete & VanLehn, 2000), medical physiology (Woo et al., 2006), law (Pinkwart, Ashley, Lynch, & Aleven, 2009), language learning (Tsiriga & Virvou, 2004), reading comprehension (Mostow et al., 2002), and meta-cognitive skills (Mitrovic, 2003).

The purpose of our research was to review and critically assess research that compared the learning outcomes of ITS with outcomes from other modes of instruction. Expanding on recent meta-analyses of ITS effectiveness, our meta-analysis reviewed evaluative studies published prior to 2013 and covers all knowledge domains and levels of education.

What Is an ITS?

When conducting a meta-analysis it is crucial to articulate a working definition of the subject so that clear and replicable inclusion criteria can be established. The goals we set for our definition of ITS were as follows:

- The definition should broadly conform to usage of the term by theorists and authors of peer-reviewed reports.
- The definition should result in a minimal number of borderline cases whose inclusion status is uncertain.
- Where theory and usage do not offer certain grounds to prefer one definition over another, we accept the more inclusive definition and use moderator variables to mark the less inclusive criteria. For example, although one might assume that ITS exhibit a high degree of interactivity and offer feedback to learners' responses, some systems that otherwise qualify as ITS do not offer feedback and instead provide features such as individualized task selection. Our solution was to include such programs as ITS and use a moderator variable to distinguish between feedback and no-feedback systems.

Shute and Psotka (1996) presented an extended consideration of the definition of ITS that included definitions elicited from leading ITS researchers. In summarizing these expert definitions, they noted that (a) almost all agreed "that the most critical element is real-time cognitive diagnosis (or student modeling)" and (b) "the next most frequently cited feature is adaptive remediation" (p. 14). An emphasis on student modeling as the key to adaptive tutoring remains evident in more recent conceptualizations of ITS (Sottilare, Graesser, Hu, & Holden, 2013). Drawing from these works and our reading of published evaluations of ITS we adopted the following definition.

An ITS is a computer system that for each student:

1. Performs tutoring functions by (a) presenting information to be learned, (b) asking questions or assigning learning tasks, (c) providing feedback or hints, (d) answering questions posed by students, or (e) offering prompts to provoke cognitive, motivational or metacognitive change

2. By computing inferences from student responses constructs either a persistent multidimensional model of the student's psychological states (such as subject matter knowledge, learning strategies, motivations, or emotions) or locates the student's current psychological state in a multidimensional domain model

3. Uses the student modeling functions identified in point 2 to adapt one or more of the tutoring functions identified in point 1

An example of multidimensionality in an ITS is the use of multiple production rules to represent domain knowledge in the Cognitive Tutors developed at Carnegie Mellon University (Anderson, Corbett, Koedinger, & Pelletier, 1995). Multidimensionality of the student or domain model is necessary to distinguish ITS from adaptive systems that model student knowledge as a single ability parameter as do some adaptive instructional systems based on item response theory (Veldkamp, Matteucci, & Eggen, 2011).

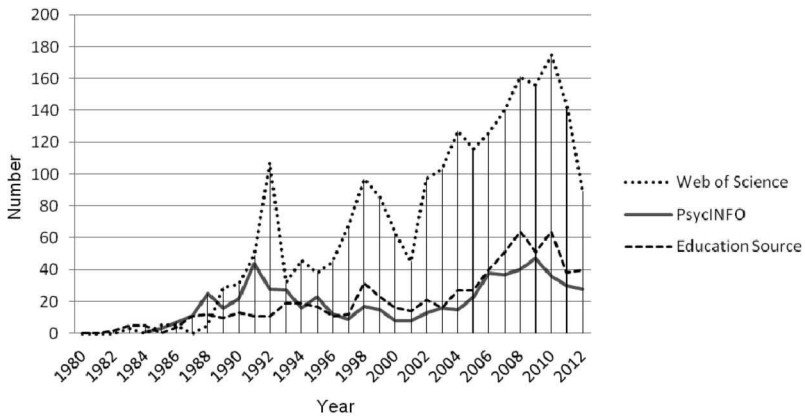

Figure 1. Number of research articles (1980–2012) retrieved with the term "intelligent tutor" from three representative bibliographic databases.

Indeed, the multidimensionality criterion is necessary to distinguish ITS from the many adaptive testing systems that use item response theory to model student knowledge as a single ability parameter and which, by selecting questions matched to student ability, may have incidental or intentional instructional effects (Santopietro, 2011). However, we do categorize as an ITS, and have included in our meta-analysis, an adaptive instructional system that uses item response theory to model student knowledge on multiple dimensions (C. M. Chen, Lee, & Chen, 2005).

There are some sophisticated tutor characteristics that, while likely desirable, we do not regard as essential to the definition of ITS. The modeling of student misconceptions or bugs has been investigated since the 1970s (Brown & Burton, 1978), but because many systems without misconception modeling have been identified as ITS in peer-reviewed research (e.g., Rowe & Schiavo, 1998) we chose to include such systems in our definition. Many ITS collect responses which are inputs the student makes to an interface that supports the stepwise construction of an answer to a problem. VanLehn (2011) assembled evidence that such step-based tutors result in better learning outcomes than tutors that work only with students' final answers. Nevertheless, because peer-reviewed research has identified answer-based systems as ITS (e.g., Tsiriga & Virvou, 2004), we have chosen to also include these in our definition.

Especially in reports appearing in the 1970s and 1980s, ITS were frequently distinguished from other forms of computer-based instruction in terms of their architectural qualities. The "test-and-branch" computer-based instruction systems being deployed and studied during that period required content authors to pre-program sequenced presentations, questions, response feedback and conditional branching to subsequent instruction (Hannum, 1986). Thus, adaptation was usually sensitive to only the student's most recent response. ITS researchers emphasized differences between the design theory of intelligent systems and that of the preceding instructional systems, and they articulated a four-component conceptual structure for ITS that has been remarkably resilient, even as ITS themselves varied significantly in their design (Dede, 1986; Hartley & Sleeman, 1973). The four "generally accepted" (Sottilare et al., 2013, p. ii) conceptual components of ITS are as follows:

1. An interface that communicates with the learner by presenting and receiving information. Often constrained to the subject domain (e.g., algebra), the interface determines the moves the learner can make in solving problems, seeking information or responding to questions.

2. A domain model that represents the knowledge the student is intended to learn. The model is a set of logical propositions, production rules, natural language statements, or any suitable knowledge representation format.

3. A student model that represents relevant aspects of the student's knowledge determined by the student's responses to questions or other interactions with the interface. Although the student model may be a subset or "overlay" of the domain model, in some ITS the student model represents common misconceptions or other "bugs" in the student's knowledge.

4. A tutor model that represents instructional strategies such as offering a hint when the student is unable to generate a correct response or assigning a problem that requires knowledge only slightly beyond the current student model.

Not all ITS have four distinct architectural components corresponding to these conceptual components. For example, an ITS may have a single knowledge base that serves as a domain and student model and have no explicit representation of teaching strategy. In our view, consistent with the definition we have provided, the student modeling process, and its use to adapt instruction, is the essential feature that distinguishes ITS from other computer-based instructional systems.

Types of ITS

ITS researchers have adopted a variety of student modeling approaches including model-tracing (Roll, Baker, Aleven, & Koedinger, 2004), probabilistic modeling (Conati & VanLehn, 1999; Conati & Zhao, 2004), reconstructive bug modeling (Mitrovic & Djordjevic-Kajan, 1995), and constraint-based modeling (Suraweera & Mitrovic, 2002). Because student modeling is a core element of ITS design and there is value in analyzing the learning outcomes associated with the most prevalent types of student modeling, we categorized the ITS in the studies we analyzed into the following four types.

Expectation and Misconception Tailoring (EMT)

Tutorial dialogues in which the computer and student exchange ideas using natural language have been a challenging goal for ITS research since SCHOLAR (Carbonell, 1970). AutoTutor is an ITS that supports natural language dialogue for instruction that, like much observed human tutoring, involves "imprecise verbal content, a low to medium level of user knowledge about a topic, and earnest literal replies" (Graesser et al., 2004, p. 181). AutoTutor models student knowledge by matching the students' responses to text passages representing expectations (i.e., learning goals) and anticipated misconceptions in the domain. The specified expectations and misconceptions constitute a multidimensional domain model. Matching is performed by a statistical method called latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998) that returns a similarity metric between an aggregation of the student's responses and each expectation and misconception in the domain model. AutoTutor uses the result of the matching process to drive scripted "dialogue moves" such as hints, feedback, prompts, and assertions (Graesser et al., 2004, p. 183).

Model Tracing

The type of student modeling adopted in the cognitive tutors developed at Carnegie Mellon University is rooted in Anderson's ACT-R theory of human learning and cognition (Anderson, 1993; Anderson & Lebière, 1998). In the design of a cognitive tutor, the skill to be learned is modeled by a set of production rules that can be activated to solve automatically the problems in the domain (Anderson et al., 1995). The production rules, which consist of operations and the conditions under which they are triggered, are selected to form a psychologically realistic emulation how humans

solve problems in the domain. The operations in the domain model are exposed in the interface where the student can select them to progress toward a problem solution. As the student selects operations a *model-tracing* process maps them to a series of production rules in the domain model. If an error is detected, the student is given immediate feedback and allowed to choose a different operation. After the student's use of a production rule is identified by model tracing, a Bayesian procedure called *knowledge-tracing* can be used to update an estimate of the probability that it has been correctly learned. Thus, the multidimensional student model in a knowledge-tracing cognitive tutor is constituted as probabilities assigned to production rules in the domain model.

Constraint-Based Modeling (CBM)

Constraint-based modeling is an established technique for student modeling that is fundamentally different from the model tracing and knowledge tracing used in the cognitive tutors (Kodaganallur, Weitz, & Rosenthal, 2005). Based on Ohlsson's theory of learning from performance errors (Ohlsson, 1994), CBM represents domain knowledge as logical constraints by relating each constraint to states that could arise in the solution of a problem. A constraint consists of three components: (a) a relevance condition that indicates when the constraint is applicable, (b) a satisfaction condition that tests the current state of the student's solution, and (c) a feedback message that, when the solution state fails the satisfaction condition, advises the student of the error and reminds them of the principle that was violated by the error (Mitrovic, Martin, & Suraweera, 2007). To explain by analogy, if the relevance condition is "cooking a pot roast," the satisfaction condition might be "oven temperature below 120 degrees Celsius" and the feedback message might be "when cooking a pot roast remember to keep the oven temperature below 120 degrees Celsius." Thus, when a student's behavior violates a constraint, an error is detected and appropriate feedback is provided (Mitrovic et al., 2011). If no constraint is violated, the student is considered to be on the right solution path. For modeling domain knowledge, constraints play a similar role in CBM as production rules play in model tracing, except that constraints cannot be executed to generate a problem solution. Constraints can be used to model student knowledge in a variety of ways. For example, the ITS can represent a student's knowledge as constraints that were found to be relevant, satisfied and violated during a problem solving session.

Bayesian Network Modeling

A Bayesian network is a tool for probabilistic reasoning and representation of uncertain knowledge (Pearl, 1988). In ITS, Bayesian networks are used to represent a multidimensional domain model consisting of multiple variables. Connections between variables are specified to form a network, and inferencing about the value of a variable in the network is accomplished by Bayesian calculations on other variables connected to it (Millán, Loboda, & Pérez-de-la-Cruz, 2010). To take a simple example, a list of binary target variables representing concepts and misconceptions might constitute the domain model and a list of evidence variables representing test items might feed forward with connections to the target variables. Taking the student's performance on the test items as input, the Bayesian network would calculate the probability that the student "has" each concept and misconception. Bayesian networks can be used to create much more complex models, such as dynamic models of student problem solving in tutors that provide hints and coaching (Conati, Gertner, & VanLehn, 2002). Bayesian networking is a flexible method that can be used to implement many different types of student models, including aspects of CBM and knowledge tracing. In practice though, we found that ITS researchers identified their ITS as belonging to at most one of the four types we have described.

Why ITS May Be More Effective Than Other Forms of Instruction

Prior reviews concluded that under some circumstances using ITS results in greater achievement than participating in traditional classroom instruction and studying printed materials (Steenbergen-Hu & Cooper, 2014; VanLehn, 2011). We hypothesize that a portion of the advantage of ITS over traditional classroom instruction and learning activities with printed materials can be attributed to the characteristics ITS share with other forms of computer-based instruction (CBI). A second-order meta-analyses by Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) found an effect size of .31 for CBI as the primary means of instruction compared with traditional classroom teaching. Researchers have explained the CBI advantage as resulting from greater interactivity and adaptation than is available in teacher-led, large-group instruction and presentational modes of instruction. Specifically, they have attributed the effectiveness of CBI to greater immediacy of feedback (Azevedo & Bernard, 1995), feedback that is more response-specific (Sosa, Berger, Saw, & Mary, 2011), greater cognitive engagement (Cohen & Dacanay, 1992), more opportunity for practice and feedback (Martin, Klein, & Sullivan, 2007), increased learner control (Hughes et al., 2013), and individualized task selection (Corbalan, Kester, & Van Merriënboer, 2006).

The prior quantitative reviews also concluded that using ITS is associated with greater achievement than using non-ITS CBI. We hypothesize that multidimensional student modeling enables ITS to outperform non-ITS CBI on each of its advantages cited in the previous paragraph. An ITS that models domain knowledge as production rules can perform task selection by characterizing each task as a set of production rules required to complete it and each student as a set of production rules that most need to be practiced, and then finding the best match. This type of multidimensional matching is likely to be more effective than unidimensionally matching student ability to task difficulty. A system that monitors and models a student's task choices and uses that model to individualize learner-control options is likely to be more effective than a system that provides the same learner-control options to all students. Likewise, feedback adjusted to account for a history of prior interactions is likely to be more effective than feedback that is determined only by the last response.

ITS may also be more effective than non-ITS CBI in the sense that ITS can extend the general advantages of CBI to wider set of learning activities. For example, the ability to score and provide individualized comments on a student's essay would extend the advantage of immediate feedback well beyond what is possible in non-ITS CBI.

Student modeling also enables ITS to interact with students at a finer level of granularity than test-and-branch CBI systems. VanLehn (2011) argued that while CBI systems typically pose a question and offer feedback about students' answers, ITS are able to interact with students at the level of the system interface that students operate on as they construct an answer. He hypothesized that whereas such interface-level interaction may scaffold students to successfully complete multi-step problems, the answer-level hints and try-again loops of typical CBI systems "offer such weak scaffolding and feedback that students are usually allowed to [quit] after several failed attempts" (p. 212).

Prior Quantitative Reviews of ITS and the Need for a Comprehensive Synthesis

VanLehn (2011) conducted a quantitative review to investigate whether computerized or human tutors that deal with more finely grained student problem-solving responses are associated with greater achievement in STEM subjects than tutoring systems that work at the level of problem solutions. He found no difference in outcomes among human tutoring, step-based systems, and substep-based systems but a significant difference favoring these finer-grained approaches over answer-based interventions that interact with learners at the level of problem solutions. Another notable contribution of VanLehn's review was that he reported an effect size of $d = 0.79$ for human tutoring compared with no-treatment controls, a value much smaller than the frequently cited, two-sigma effect sizes reported by Bloom (1984). VanLehn attributed the larger effects reported by Bloom as the result of combining individual tutoring with mastery learning, a strategy that other tutoring research did not use.

Steenbergen-Hu and Cooper (2014) conducted a meta-analysis of 39 studies evaluating the use of ITS for college students' academic learning The studies appeared between 1990 and 2011and reported on 22 different types of ITS While most types of ITS in the analysis were evaluated by only a single study, those evaluated in multiple studies included AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005), ALEKS (Hagerty & Smith, 2005), and xTEx-Sys (Grubišić, Stankov, & Hrepic, 2008). The researchers found an overall, moderate, positive effect ($g = .35$) favoring the use of ITS. When compared specifically to alternatives that were either "self-reliant learning activities" or no-treatment conditions, the use of ITS appeared to offer a large advantage ($g = .86$). However, the three studies that compared ITS to human tutoring resulted in a negative mean effect size ($g = –.25$) that was not statistically significant.

The same authors (Steenbergen-Hu & Cooper, 2013) conducted a meta-analysis of 34 studies from evaluations of ITS in K–12 mathematical learning. The studies were published between 1997 and 2010 and mostly compared the used of ITS to regular classroom instruction. Using a random effects model, they obtained an overall effect size that was not significantly different from zero. Notably, there was a statistically significant mean effect size favoring ITS when learning was measured by course-specific tests but not when measured by standardized tests.

The previous meta-analyses were limited to subsets of the ITS evaluation literature defined by subject or level of schooling. There are two major reasons why a comprehensive meta-analysis is preferable to a collection of smaller analyses collectively covering the same studies. First, the greater number of studies included in a more comprehensive review increases statistical power for detecting whether a mean effect size is significantly greater than zero or significantly different from an effect size at a different level of a moderator variable. Second, to perform a statistical comparison of the effect sizes of two or more categories of studies it is necessary to establish them as two or more levels of a moderator variable in the same meta-analysis. In their discussion, Steenbergen-Hu and Cooper (2014) noted the disparity between the overall effect sizes of their two meta-analyses and questioned whether ITS might affect college students and K–12 students differently. Only by including the different levels of schooling as different levels of a moderator variable in the same ITS meta-analysis can we determine whether the difference between these levels is statistically significant. In addition to these reasons, a comprehensive meta-analysis of ITS is currently needed to incorporate and expose to comparative review the studies that evaluated the effects of ITS in K–12 subjects other than mathematics.

Research Questions

This review synthesizes research on the relative effectiveness of Intelligent Tutoring Systems and addresses the following research questions:

1. Do students using ITS have different learning outcomes from students using other modes of instruction?
2. Do the effects associated with ITS vary with characteristics of the ITS?
3. Do the effects associated with ITS vary with characteristics of the students, outcome assessments, and research setting?
4. Do the effects associated with ITS vary with the methodological features of the research?

Method

Selection Criteria

To capture evidence relevant to the research questions, studies were considered eligible for inclusion in the meta-analysis if they:

(a) reported original data;

(b) assessed learning outcomes after interaction with software that matched the definition of ITS presented in the introductory section of this review;

(c) compared learning outcomes from the ITS with outcomes from a non-ITS mode of instruction;[1]

(d) were publicly available, online or in library archives;

[1] Studies that compared a group learning from ITS with a control group that received no instructional treatment were retained but were meta-analyzed separately to provide interpretive context for the results bearing more directly on the research questions.

(e) reported sufficient data to calculate effect size;

(f) reported measurable cognitive outcomes such as recall, transfer, or a mix of both.

Search, Retrieval, and Selection of Studies

We conducted a comprehensive search for relevant research in four major bibliographic databases: ERIC, PsycINFO, Springer Link, and Web of Science. The search, which returned 26,613 titles, applied the following key terms: intellige* tutor*, intellige* agent*, cognit* tutor*, adapt* tutor*, cognit* virtual companion*, and intellige* coaching system*. Also, the reference sections of review articles on Intelligent Tutoring Systems were manually searched to add studies to the selection pool (Arnott, Hastings, & Allbritton, 2008; Conati, 2009; Steenbergen-Hu & Cooper, 2013, 2014; VanLehn, 2011; N. Wang et al., 2008).

In the initial screening phase, the abstracts of the articles were compared with criteria a, b, c and d to exclude irrelevant studies. The 362 articles that passed the initial screening were retrieved, and full-text copies were further evaluated against all six inclusion criteria. Finally, the 107 studies (involving 14,321 participants) that met the inclusion criteria were coded using a pre-defined coding form and coding instructions developed for this meta-analysis. All effect sizes were calculated with Hedges's correction for bias due to small sample sizes (Lipsey & Wilson, 2001).

Coding Study Characteristics and Effect Sizes Extraction

The coding form included 44 fixed-choice items and 37 comment items, not all reported here, that elicited detailed information about the studies such as author, year published, source of the study, research questions, type of ITS, control treatment, grade level of participants, research settings, duration of the study, reliability reporting and statistics needed for computing the effect size of each study.

During the coding process, we observed that some studies evaluated the learning outcomes of more than two groups. For instance, in addition to a control group there might be a group that interacted with an ITS via text and another group that interacted with the same ITS via an animated agent. If each ITS and control contrast were entered in the coding spreadsheet, the control group would be counted twice, the overall weight attributed to the study would be inflated, and the two contrasts would be statistically dependent. A coding strategy was followed to avoid statistical dependence when there were more than two groups in a study (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001). Specifically, when there was more than one control group in a study, any control group that received no instructional treatment was dropped from the main meta-analysis according to selection criterion c and other control groups that did receive instructional treatment were combined by calculating their weighted mean. Similarly, when there was more than one group in a study that learned from an ITS, we combined them by calculating their weighted mean.

Data Analysis and Interpretation

We followed standard guidelines for conducting a meta-analysis (Adesope & Nesbit, 2012; Adesope, Lavin, Thompson, & Ungerleider, 2010; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Nesbit & Adesope, 2006). After coding had been completed, the spreadsheet was imported to IBM® SPSS® Statistics software (Version 21) and later to Comprehensive Meta-analysis 2.2.048 for further analysis (Borenstein et al., 2009). The Comprehensive Meta-Analysis software was used to generate the unbiased mean effect size (Hedges's g), the standard error of Hedges's unbiased estimate of the mean effect size, 95% lower and upper confidence interval around each mean, and values for the test of heterogeneity including Q, p and I-squared.

We interpreted the confidence intervals spanning a range above zero as indicating a statistically significant result favoring learning from Intelligent Tutoring Systems over learning from other instructional treatments. Moreover, the upper and lower 95% confidence intervals were used to detect between-levels differences among different categories of analyses. Specifically, when the confidence intervals of categories were not overlapping, their effect sizes were judged to be significantly different.

An important step in meta-analysis is testing whether the observed effect sizes of individual studies that are averaged into a mean effect size all estimate the same population effect size. This assumption of homogeneity of effects is tested by the Q statistic. When all findings are drawn from the same population, Q has an approximate chi-square distribution with k-1 degrees of freedom, where k is the number of studies that constitute a particular subset of analysis. When Q exceeds the critical value of the chi-square distribution, (i.e., $p < .05$), the mean effect size is said to be significantly heterogeneous, indicating that individual effect sizes do not estimate a common population mean (Borenstein et al., 2009; Lipsey & Wilson, 2001).

Two primary effect sizes produced extreme standardized scores ($-3.3 \geq Z \geq 3.3$; $p < .001$) and were thereby identified as outliers. One of these studies produced an effect size $g = 2.25$, whereas the other produced an effect size $g = -1.10$. Further examination of these two studies did not reveal any methodological flaws. Comprehensive Meta-Analysis was used to determine whether removing the two outliers would yield a homogeneous distribution (Hedges & Olkin, 1985). First, we examined the forest plot of all 107 effect sizes and then removed the two potential outliers one at a time. The recalculated results showed that the removal of potential outliers did not improve the fit of the remaining effect sizes to a simple model of homogeneity. However, as recommended by Tabachnick and Fidell (2013), we adjusted each effect size toward the nearest other effect size in the distribution. The adjusted effect sizes were $g = 1.5$ and -0.5.

To maximize interpretation of results, we used both the fixed-effect and random-effects model in all data analyses. A fixed-effect model operates under the assumption that all the studies included in the meta-analysis (107 studies in the present article) share one true effect size, which is estimated to be the average effect size. Conversely, a random-effects model operates under the assumption that there is more than one true effect and that the effect sizes could vary from one study to the other (Borenstein et al., 2009; Lipsey & Wilson, 2001). Considering the great diversity in the ways that research interventions are implemented (e.g., the many different ways that ITS are designed and used) and the variability that could potentially exist from aggregating such a multiplicity of conditions, a random-effects model is usually regarded as a more accurate model than a fixed-effect model (Bo-

renstein et al., 2009; Denson, 2009; Hedges & Vevea, 1998; National Research Council, 1992). Therefore, we reported detailed results for the random-effects model and added summary results for the fixed-effect model to allow comparison with the fixed-effect results reported by prior ITS meta-analyses. Because the fixed-effect results may give additional indication to researchers about the areas in which further research might yield significant effect sizes we also reported which moderator variables and levels showed significant differences under a fixed-effect model. Any mean effect size reported without specifying the type of model was generated by a random-effects model.

Results

We found nine effect sizes (785 participants) derived from six publications in which the control group did not receive instructional treatment but which otherwise met the inclusion criteria (Arroyo, Royer, & Wolf, 2011; Beal, Arroyo, Cohen, & Wolf, 2010; L. H. Chen, 2011; Halpern, Millis, Graesser, & Butler, 2012; Shute, Hansen, & Almond, 2007; H. C. Wang, Rosé, & Chang, 2011). As these effect sizes, shown in Table S1 in the online supplementary materials, did not bear directly on our research questions, they were analyzed separately to provide interpretive context for the main results. Under a random-effects model, they were found to have a statistically significant, weighted mean effect size of $g = 1.23$.

Figure 2 shows the distribution of effect sizes for the main meta-analysis after adjustment of the two outliers. The effect sizes are mainly clustered between –.25 and.75 standard deviations, indicating that in most studies the Intelligent Tutoring Systems groups outperformed their respective control groups. Throughout the results section, a positive effect size indicates that students who used ITS outperformed those who experienced other modes of instruction. Conversely, a negative effect size indicates that students who used other modes of instruction performed better than those who used ITS.

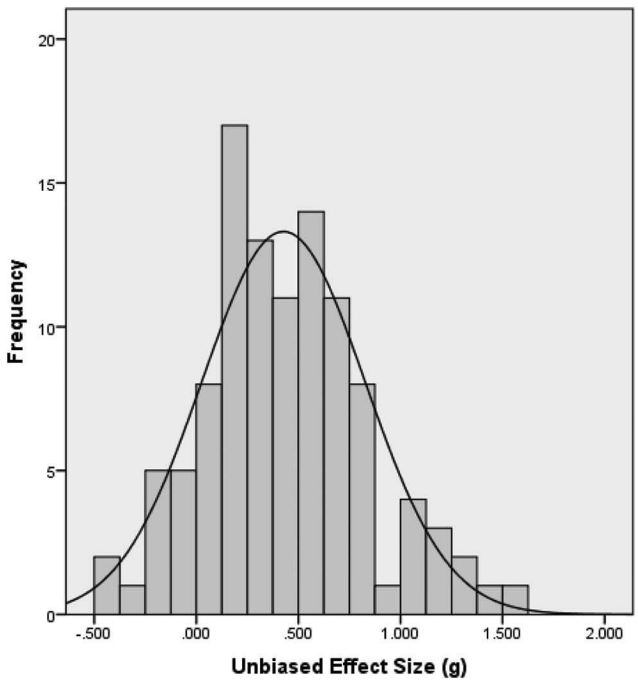

Figure 2. Distribution of 107 effect sizes ($M = 0.43$, $SD = 0.40$).

Table S2, shown in the online supplemental materials, summarizes the characteristics of each study that met the inclusion criteria, including the author(s), subject domain, grade-level of participants, type of ITS, comparison treatment, study setting, the unbiased effect size, Hedges's *g*, and 95% lower and upper confidence intervals around each unbiased effect size.

Tables 1–6 present the results organized by the research questions. These tables present results for both the fixed- and random-effects models and include the number of participants (*N*) in each category, the number of studies (*k*), the weighted mean effect size ($g+$) and its standard error (*SE*), the 95% confidence interval around the mean, and a test of heterogeneity (*Q*). Each weighted mean effect size was obtained through weighting of independent effect sizes by inverse variances.

Research Question 1: Do Students Using ITS Have Different Learning Outcomes From Students Using Other Modes of Instruction?

Table 1 shows the overall analysis of the weighted mean of all statistically independent effect sizes. Under a fixed-effect model, Table 1 shows a moderate, statistically significant effect of learning with intelligent tutors ($g = .36$, $p < .001$) with significant heterogeneity, $Q(106) = 390.52$, $p < .001$, $I^2 = .73$. Under a random-effects model, the overall weighted mean effect size was also statistically significant and moderate ($g = .41$, $p < .001$).

Table 1 also lists the breakdown of the comparison treatment instruction in all studies. It shows that the majority of the studies compared the use of intelligent tutors with large-group human instruction ($k = 66$). Under both the fixed- and random-effects models, the use of ITS produced moderate, statistically significant mean effect sizes when compared with large-group human instruction which included but was not limited to traditional classroom instruction ($g = .44$), individual computer-based instruction (CBI, $g = .57$) and the individual use of textbooks or workbooks ($g = .36$). The use of ITS did not produce statistically significant effect sizes when compared with small-group human instruction (defined as any form of synchronous instruction in groups of up to 8 students conducted with the presence of a human tutor such as problem-based learning and similar collaborative methods). Individual human instruction (i.e., human tutoring) appeared to offer a small, non-significant advantage over the use of ITS. The between-levels variance was statistically significant under both the fixed- and random-effects models ($p < .001$). Post hoc analyses found studies which compared the use of ITS to large-group human instruction had effect sizes similar to those which compared ITS to individual computer-based instruction and individual use of textbooks or workbooks, but these all had significantly higher weighted mean effect sizes than studies which compared the use of ITS to human tutoring. Taken together, these results showed that students who used ITS learned significantly more than students who used other modes of instruction except small-group and individual human tutoring.

Table 1
Overall Mean Effect and Mean Effect Sizes for Comparison Treatments

Overall effect	*N*	*k*	Effect size *g*+	Effect size *SE*	95% CI Lower	95% CI Upper	Test of heterogeneity Q_B	*df*	*p*	I^2 (%)
Fixed-effect model	14,321	107	0.36*	0.02	0.32	0.39	390.52	106	<.001	0.73
Random-effects model	14,321	107	0.41*	0.04	0.34	0.48				

	Random-effects model								Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Comparison treatments	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Type of instruction							27.54	<.001					27.35	<.001
Large-group human instruction	11,296	66	0.44*	0.05	0.35	0.53			0.37*	0.02	0.33	0.41		
Small-group human instruction	184	4	0.05	0.28	−0.50	0.61			0.10	0.16	−0.21	0.41		
Individual human instruction	404	5	−0.11	0.10	−0.31	0.10			−0.11	0.10	−0.31	0.10		
Individual CBI	1,034	15	0.57*	0.11	0.34	0.79			0.47*	0.06	0.34	0.59		
Individual textbook or workbook	1,403	17	0.36*	0.09	0.18	0.53			0.30*	0.06	0.19	0.41		

Note. CI = confidence interval; CBI = computer-based instruction.
* $p < .05$.

The significant heterogeneity in the overall result indicates there was unattributed variability in the individual effect sizes that constitute the overall result. Therefore, moderator analyses were conducted on ITS characteristics, sample characteristics and methodological features of the studies to further determine the factors that may be responsible for the variability in effect sizes.

Research Question 2: Do the Effects Associated With ITS Vary With Characteristics of the ITS?

The results in Table 2 show how different features and characteristics of ITS moderated the overall effect of learning with these systems. We examined the effects of learning with different characteristics of ITS including the type of ITS, the nature of inter-

Table 2
Weighted Mean Effect Sizes for Characteristics of Intelligent Tutoring Systems (ITS)

			Random-effects model						Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Moderator variables	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Type of ITS							4.18	.52					36.37	<.001
Model tracing	5,970	21	0.35*	0.07	0.22	0.47			0.25*	0.03	0.20	0.31		
Constraint-based modeling	569	7	0.24	0.16	−0.08	0.56			0.20*	0.09	0.03	0.37		
Bayesian network modeling	1,417	10	0.54*	0.10	0.35	0.73			0.52*	0.06	0.41	0.63		
Expectation and misconception tailoring	142	3	0.34	0.35	−0.35	1.02			0.24	0.18	−0.12	0.59		
Other	4,425	53	0.44*	0.06	0.32	0.56			0.44*	0.03	0.38	0.50		
Not reported	1,798	13	0.40*	0.10	0.20	0.59			0.43*	0.05	0.32	0.54		
ITS intervention							2.41	.79					32.38	<.001
Principal instruction	4,505	35	0.37*	0.07	0.23	0.51			0.32*	0.03	0.26	0.38		
Integrated class instruction	4,045	15	0.33*	0.08	0.17	0.49			0.25*	0.03	0.18	0.31		
Separate in-class activities	1,939	24	0.47*	0.10	0.27	0.67			0.53*	0.05	0.43	0.62		
Supplementary after-class instruction	933	8	0.43*	0.11	0.22	0.64			0.36*	0.07	0.23	0.48		
Homework	2,480	15	0.45*	0.07	0.32	0.59			0.46*	0.04	0.38	0.54		
Not reported	419	10	0.48*	0.13	0.23	0.74			0.47*	0.10	0.27	0.66		
Feedback provided?							4.55	.10					13.53	<.001
No	1,411	10	0.54*	0.15	0.25	0.83			0.40*	0.05	0.30	0.51		
Yes	11,728	86	0.42*	0.04	0.34	0.50			0.37*	0.02	0.33	0.41		
Not reported	1,182	11	0.21*	0.10	0.02	0.41			0.15*	0.06	0.04	0.27		
Model misconception?							0.02	.99					5.14	.08
No	1,508	21	0.40*	0.07	0.27	0.54			0.39*	0.05	0.29	0.49		
Yes	9,911	58	0.40*	0.05	0.31	0.49			0.33*	0.02	0.29	0.37		
Not reported	2,902	28	0.42*	0.10	0.23	0.61			0.43*	0.04	0.35	0.51		

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

Table 3
Weighted Mean Effect Sizes for Student and Study Characteristics

			Random-effects model						Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Moderator variables	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Grade levels							2.2	.82					25.74	<.001
Elementary school	1,496	19	0.31*	0.08	0.16	0.47			0.26*	0.05	0.16	0.37		
Middle school	810	10	0.41*	0.13	0.15	0.66			0.45*	0.07	0.31	0.59		
High school	4,355	14	0.40*	0.10	0.21	0.59			0.25*	0.03	0.18	0.31		
Postsecondary	6,767	60	0.43*	0.05	0.33	0.53			0.43*	0.03	0.38	0.48		
Mixed grades	771	3	0.61	0.32	−0.02	1.25			0.42*	0.07	0.28	0.57		
Not reported	122	1	0.33	0.18	−0.02	0.69			0.33	0.18	−0.02	0.69		
Subject domains							6.53	.48					50.67	<.001
Mathematics and Accounting	8,038	35	0.35*	0.05	0.24	0.45			0.29*	0.02	0.25	0.34		
Physics	2,890	24	0.38*	0.07	0.26	0.51			0.41*	0.04	0.33	0.49		
Computer Science	1,152	19	0.51*	0.11	0.30	0.72			0.46*	0.06	0.34	0.58		
Language and Literacy	1,075	14	0.34*	0.11	0.12	0.56			0.27*	0.06	0.15	0.39		
Chemistry	141	2	0.16	0.17	−0.17	0.48			0.16	0.17	−0.17	0.48		
Biology and Physiology	210	3	0.59*	0.27	0.07	1.11			0.51*	0.14	0.23	0.78		
Humanities and Social Science	671	8	0.63*	0.22	0.20	1.06			0.84*	0.08	0.68	1.01		
Others and Not Reported	144	2	1.23	0.96	−0.65	3.10			0.53	0.17	0.20	0.87		
Prior domain knowledge							3.45	.49					11.87	.02
Low	5,265	32	0.38*	0.06	0.27	0.49			0.37*	0.03	0.31	0.43		
Medium	1,356	17	0.28*	0.08	0.12	0.45			0.27*	0.06	0.16	0.38		
High	77	2	0.51	0.29	−0.06	1.07			0.53*	0.23	0.07	0.98		
Varied	2,699	22	0.48*	0.12	0.25	0.71			0.27*	0.04	0.19	0.34		
Not reported	4,924	34	0.46*	0.06	0.34	0.58			0.41*	0.03	0.35	0.47		

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

vention provided by the ITS, whether the ITS modeled misconceptions, as well as whether the ITS provided feedback. Most commonly, ITS were used as the principal means of instruction ($k = 35$), provided feedback to students ($k = 86$), and modeled student misconceptions ($k = 58$). Under a random-effects model, two types of ITS, constraint-based modeling and expectation and misconception tailoring, did not produce significant effects. However, ITS with model tracing, Bayesian network modeling and other types of student modeling produced statistically significant effect sizes. Although ITS with Bayesian network modeling produced a higher weighted mean effect size ($g = .54$) than model tracing ($g = .35$), constraint-based modeling ($g = .24$), and expectation and misconception tailoring ($g = .34$), the between-levels difference was not statistically significant under a random-effects model. However, statistically significant differences were detected under a fixed-effect model, $Q_B(5) = 36.37$, $p < .001$, and post hoc analyses found ITS which used Bayesian network modeling had a significantly higher weighted mean effect size than those which used model tracing and constraint-based modeling.

Table 2 also shows that ITS were effective in all the instructional roles in which they were evaluated. Adopting the categories used by Steenbergen-Hu and Cooper (2014), the instructional roles of ITS were coded as: principal instruction (the ITS was the principal means of instruction); integrated class instruction (the ITS was an integral part of regular classroom instruction); separate in-class activities (the ITS was used for separate laboratory or other exercises that took place during class time), supplementary after-class instruction, and homework (the ITS was used as part of homework assignments). Across all these categories, the use of ITS was associated with statistically significant effect sizes under both the fixed- and random-effects models. The between-levels difference was not statistically significant under a random-effects model. However, the between-levels variance was statistically significant under a fixed-effect model, $Q_B(5) = 32.38$, $p < .001$. Post hoc analyses found that studies which used ITS for separate, in-class activities and homework had significantly higher weighted mean effect sizes than those which used ITS for other purposes such as principal instruction.

Table 2 shows that, under both fixed and random-effects models, the use of ITS was associated with statistically significant effect sizes whether or not they provided feedback to students. Under both the fixed- and random-effects models, overlap in confidence intervals indicates that effect sizes were not moderated by whether or not the ITS provided feedback. Table 2 also shows that the use of ITS produced moderate, statistically significant effect sizes regardless of whether the ITS modeled misconceptions or not. The between-levels difference was not statistically significant under both the fixed and random-effects models.

Research Question 3: Do the Effects Associated With ITS Vary With Characteristics of the Students, Outcome Assessments, and Research Setting?

To answer the third research question, Tables 3, 4, and 5 show results of moderator analyses based on student and study characteristics, outcome assessments and research settings, respectively. Specifically, Table 3 shows the effects of using ITS across different grade levels, subject domains, and levels of prior knowledge.

Table 4
Weighted Mean Effect Sizes for Outcome Constructs, Test Format, Knowledge Type, and Measuring Tool

			Random-effects model						Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Moderator variables	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Outcome constructs							1.12	.77					3.56	.31
Retention	3,922	33	0.35*	0.07	0.22	0.48			0.35*	0.03	0.28	0.41		
Transfer	1,683	18	0.44*	0.09	0.27	0.62			0.43*	0.05	0.33	0.52		
Mixed retention and transfer	6,371	32	0.43*	0.08	0.28	0.58			0.33*	0.03	0.28	0.38		
Not reported	2,345	24	0.42*	0.06	0.31	0.54			0.39*	0.04	0.30	0.47		
Test formats							8.79	.07					108.17	<.001
Multiple choice	1,777	18	0.26*	0.05	0.16	0.36			0.26*	0.05	0.16	0.36		
Short answer	1,170	11	0.25*	0.06	0.13	0.36			0.25*	0.06	0.13	0.36		
Mixed items	1,701	10	0.06	0.05	−0.03	0.16			0.06	0.05	−0.03	0.16		
Other	972	10	0.91*	0.07	0.77	1.05			0.91*	0.07	0.77	1.05		
Not reported	8,701	58	0.40*	0.02	0.35	0.44			0.40*	0.02	0.35	0.44		
Knowledge type							1.18	.76					30.33	<.001
Procedural	6,143	46	0.39*	0.05	0.28	0.49			0.36*	0.03	0.31	0.42		
Declarative	4,318	31	0.37*	0.07	0.23	0.51			0.26*	0.03	0.20	0.32		
Mixed procedural and declarative	777	6	0.65*	0.29	0.08	1.21			0.70*	0.08	0.55	0.86		
Not reported	3,083	24	0.43*	0.06	0.32	0.54			0.40*	0.04	0.33	0.48		
Test source							0.71	.87					14.92	<.001
Researcher developed	7,279	62	0.41*	0.05	0.32	0.50			0.40*	0.02	0.36	0.45		
Standardized	4,597	19	0.42*	0.10	0.21	0.62			0.27*	0.03	0.21	0.33		
Both	1,095	5	0.46*	0.07	0.33	0.59			0.46*	0.07	0.33	0.59		
Not reported	1,350	21	0.38*	0.07	0.24	0.52			0.34*	0.06	0.23	0.45		

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

Studies with students from kindergarten through grade 5 were grouped together under *elementary school*. Studies with students from grades 6 through 8 were categorized as *middle school* while grades 9 through 12 were categorized as *high school*. Studies conducted with university and college students were categorized as *postsecondary*. Three studies cut across these grade bands and were separately categorized as *mixed grades*. The use of ITS produced moderate statistically significant mean effect sizes at all grade levels under both the fixed and random-effects models. Table 3 further shows that the between-levels difference was not statistically significant under a random-effects model but statistically significant under a fixed-effect model, $Q_B(5) = 25.74$, $p < .001$. Post hoc analyses found studies which used ITS with students in middle school and postsecondary had significantly higher

Table 5
Weighted Mean Effect Sizes for Contextual Features

			Random-effects model						Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Moderator variables	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Setting							3.30	.07					4.29	.04
Laboratory	1,596	26	0.29*	0.07	0.15	0.43			0.26*	0.05	0.16	0.36		
Classroom	12,725	81	0.44*	0.04	0.36	0.52			0.37*	0.02	0.33	0.41		
Continent							3.59	.31					12.80	.01
North America	11,065	75	0.38*	0.04	0.29	0.46			0.33*	0.02	0.29	0.37		
Europe	1,083	18	0.51*	0.10	0.32	0.71			0.55*	0.06	0.43	0.67		
Asia	962	6	0.67*	0.20	0.28	1.06			0.42*	0.07	0.29	0.55		
Oceania	1,211	8	0.36*	0.07	0.22	0.51			0.38*	0.06	0.27	0.50		
Treatment duration							1.43	.49					4.67	.10
One hour or less	587	9	0.30	0.16	−0.01	0.62			0.18*	0.09	0.02	0.35		
Greater than one hour	7,589	59	0.39*	0.05	0.30	0.48			0.35*	0.02	0.30	0.40		
Not reported	6,145	39	0.47*	0.07	0.34	0.60			0.38*	0.03	0.33	0.43		
Study duration							3.78	.15					32.36	<.001
One month or less	2,044	32	0.34*	0.07	0.22	0.47			0.31*	0.05	0.22	0.40		
Greater than one month	9,577	53	0.38*	0.05	0.29	0.47			0.31*	0.02	0.27	0.35		
Not reported	2,700	22	0.57*	0.10	0.37	0.76			0.57*	0.04	0.49	0.66		

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

Table 6
Weighted Mean Effect Sizes for Different Methodological Features

			Random-effects model						Fixed-effect model					
			Effect size		95% CI				Effect size		95% CI			
Moderator variables	*N*	*k*	*g*+	*SE*	Lower	Upper	Q_B	*p*	*g*+	*SE*	Lower	Upper	Q_B	*p*
Random assignment							7.39	.06					65.58	<.001
Yes	5,588	34	0.31*	0.06	0.19	0.43			0.22*	0.03	0.16	0.27		
No—prior difference controlled	3,075	23	0.38*	0.07	0.25	0.50			0.35*	0.04	0.27	0.42		
No—prior difference not controlled	4,724	34	0.54*	0.06	0.42	0.67			0.55*	0.03	0.49	0.61		
Not reported	934	16	0.37*	0.11	0.15	0.58			0.30*	0.07	0.17	0.43		
Source							2.36	.50					19.73	<.001
Journal	7,171	72	0.44*	0.04	0.36	0.53			0.42*	0.02	0.37	0.47		
Conference proceeding	4,045	23	0.33*	0.08	0.18	0.49			0.29*	0.03	0.23	0.36		
Dissertation/thesis	1,419	5	0.27	0.15	−0.02	0.57			0.19*	0.05	0.08	0.30		
Technical report	1,686	7	0.46*	0.18	0.11	0.81			0.39*	0.05	0.29	0.49		
Attrition of participants							4.08	.13					30.29	<.001
None	3,191	36	0.39*	0.07	0.24	0.53			0.26*	0.04	0.19	0.33		
Some	4,075	23	0.29*	0.09	0.11	0.47			0.27*	0.03	0.20	0.33		
Not reported	7,055	48	0.48*	0.04	0.40	0.56			0.46*	0.03	0.41	0.51		

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

weighted mean effect sizes than those which used ITS in elementary and high schools.

Table 3 also shows that ITS were associated with positive moderate to large effect sizes across different subject domains. Notably, under a random-effects model, the use of ITS produced a large effect size in learning humanities ($g = .63$). For domains such as biology and physiology ($g = .59$), computer science ($g = .51$), physics ($g = .38$), and mathematics and accounting ($g = .35$), ITS produced moderate effect sizes. For chemistry as well as literacy and language learning, the use of ITS produced small and moderate mean effect sizes ($g = .16$ and $g = .34$, respectively). The between-levels variance was not statistically significant under a random-effects model but was significant under a fixed-effect model, $Q_B(7) = 50.67$, $p < .001$, indicating significant differences across subject domains. Post hoc analyses found studies which used ITS in humanities and social sciences had a significantly higher weighted mean effect size than those which used it mathematics and accounting, physics, computer science, language and literacy, and chemistry.

Many of the participants had low prior domain knowledge ($k = 32$). Under the random-effects model, all the categories of prior domain knowledge were associated with statistically significant effect sizes except high prior domain knowledge. However, the results revealed no significant differences between students with low, medium, and high prior domain knowledge. However, the certainty of this interpretation is limited by at least three factors: (a) the large number of studies that did not report the prior domain knowledge of participants ($k = 34$); (b) the small number of studies having participants with high prior knowledge ($k = 2$); and (c) the significant heterogeneity of the effect size distributions.

Table 4 shows the mean effect sizes for different outcome assessments, test formats, knowledge types and test source. The learning outcomes were coded as retention, transfer, and mixed retention and transfer, and the test formats were coded as objective format (e.g., multiple choice items), short answer, and mixed format (e.g., combinations of multiple choice and short answer). Knowledge type was coded as procedural, declarative and mixed procedural and declarative while test source was coded as researcher-developed, standardized or both. Under both the fixed- and random effects models, the use of ITS was associated with statistically significant mean effect sizes regardless of the learning outcome. The between-levels variance was not statistically significant under both models.

ITS produced statistically significant effect sizes across all test formats, except for mixed item formats. The between-levels variance was not statistically significant under the random-effects model. However, the between-levels variance was statistically significant under a fixed-effect model, $Q_B(4) = 108.17$, $p < .001$. Table 4 also shows that ITS are effective for learning all knowledge types. The between-levels variance was not statistically significant under a random-effects model but was significant under a fixed-effect model, $Q_B(3) = 30.33$, $p < .001$, indicating significant differences between knowledge types. Post hoc analyses found studies that used ITS to acquire mixed procedural and declarative knowledge had a significantly higher weighted mean effect size than those that used ITS to acquire only procedural, or only declarative knowledge. Finally, Table 4 shows that moderate, statistically significant effect sizes were obtained with researcher-developed tests, standardized tests, and tests that were both researcher-developed and standardized under both the fixed and random-effects model. However, the between-levels variance was only statistically significant under the fixed-effect model, $Q_B(3) = 14.92$, $p < .001$. Post hoc analyses found that researcher-developed tests had a significantly higher weighted mean effect size than standardized tests.

Table 5 shows the results of analyses of contextual moderator variables: the setting where the research was conducted (laboratory or classroom), the continents where study was conducted, the treatment duration, and the entire study duration. We coded "classroom" studies as those which had learning activities that were reported as part of an academic course of study or were conducted in a classroom under the supervision of an instructor. Conversely,

when learning activities were conducted solely for the purpose of research and learning was not assessed for academic credit, the setting was coded as "laboratory." Approximate median splits on duration of treatment (split at one hour) and duration of study (split at one month) were used to create two categories for each of these variables. Table 5 shows that most of the studies were conducted in the classroom ($k = 81$). Both the classroom and laboratory studies produced moderate statistically significant effect sizes, under both the fixed and random-effects models. The between-levels variance was not statistically significant under the random-effects model but marginally significant under the fixed-effect model, showing that classroom-based studies produced a higher weighted mean effect size than laboratory studies.

The majority of the studies were conducted in North America ($k = 75$). The effectiveness of ITS was evident regardless of the region where studies were conducted. Under the random-effects model, the use of ITS was associated with moderate weighted mean effect sizes in North America ($g = .38$), Europe ($g = .51$), and Oceania ($g = .36$). The effect size in Asia was larger ($g = .67$). The between-levels variance was only statistically significant under the fixed-effect model, $Q_B(3) = 12.80$, $p = .01$. Post hoc analyses found that the weighted mean effect size associated with studies conducted in Europe was larger than for studies conducted in North America.

We observed a pattern showing higher mean effect sizes for longer treatment and study durations. Results from a random-effects model showed that treatments which were less than or equal to one hour in length produced a statistically significant mean effect size ($g = .30$), as did those greater than one hour ($g = .39$). However, neither the fixed- nor random-effects model showed statistically significant, between-levels variance. Under the random-effects model, studies conducted for a month or less and those conducted for over a month were also associated with statistically significant, weighted mean effect sizes ($g = .34$ and $g = .38$). These results should be interpreted with caution because of the large number of studies that did not report the treatment and study durations.

Research Question 4: Do the Effects Associated With ITS Vary With the Methodological Features of the Research?

Table 6 shows how effect sizes varied with the methodological features of studies included in this meta-analysis. The studies were categorized according to research design, source of publication and attrition. Under both fixed- and random-effects models, learning with ITS produced moderate, statistically significant effect sizes regardless of research design. The between-levels variance was statistically significant only under the fixed-effect model, $Q_B(3) = 65.68$, $p < .001$. Post hoc analyses found that a larger mean effect size was associated with quasi-experimental designs in which prior differences were not controlled. Again, we call for caution in interpreting this result because of the high number of studies that did not explicitly report research designs.

Studies published in journals often have higher methodological quality than those presented at conferences or as dissertations. Table 6 also shows that, under a random-effects model, mean effect sizes were statistically significant for studies published in journals, conference proceedings, as well as technical reports. However, dissertation studies did not produce a statistically significant effect size. The between-levels variance was not statistically significant under random-effects model. However, it was statistically significant, $Q_B(3) = 19.73$, $p < .001$, under a fixed-effect model showing that studies published in journals had a moderate mean effect size that was significantly different from studies in conference proceedings and dissertations or theses. Finally, under both the fixed- and random-effects models, studies without attrition of participants and those with some attrition produced moderate, statistically significant effect sizes.

Are These Results Valid?

To determine whether the results reported in this meta-analysis can be regarded as valid we investigated the potential impact of publication bias. Publication bias is a plausible threat to the validity of meta-analyses because statistically significant results are more likely to be published and accessible for inclusion in meta-analyses than non-statistically significant results which may either not be reported or reported in less accessible outlets (Orwin, 1983; Rosenthal, 1979). This is often called the "file-drawer" effect. Further analysis of publication bias is particularly crucial in this meta-analysis considering that the between-levels variance of source of publication was statistically significant under a fixed-effect model showing that studies published in journals had a moderate mean effect size that was significantly different from studies in conference proceedings and dissertations or theses (see Table 6).

Two statistical tests were computed with Comprehensive Meta-Analysis to further examine the potential for publication bias. First, a "Classic Fail-Safe *N*" test was computed to determine the number of null effect studies needed to raise the *p* value associated with the average effect above an arbitrary alpha level (set at $\alpha = .05$). This test revealed that 871 additional studies would be required to invalidate the overall effect found in this meta-analysis. Orwin's Fail-Safe *N*, a more stringent publication bias test, revealed that 656 missing null studies would be required to bring the mean effect size found in this meta-analysis to a trivial level of .05. Taken together, these tests show that, with 107 analyzed studies, the number of null studies required to invalidate the overall effect size found in this meta-analysis is larger than the $5k + 10$ limit suggested by Rosenthal (1995). Hence, although there is potential for publication bias in this meta-analysis, the results of these two tests suggest that publication bias does not pose a significant threat to the validity of the findings reported in this meta-analysis.

Discussion

Summary of the Results

The overall result of our meta-analysis is that ITS outperformed, in aggregate, the other modes of instruction to which it was compared in evaluative studies. Moderator analysis found that using ITS was associated with significantly higher achievement outcomes than using each of the other modes of instruction except small-group human tutoring and individual human tutoring, and the difference in learning outcomes between ITS and these two forms of human tutoring was not statistically significant. ITS was also associated with greater achievement regardless of whether it

was used as the principal means of instruction, as an integral part of classroom instruction, to support in-class activities such as laboratory exercises, for supplementary after-class instruction, or as part of assigned homework. In analyzing 18 other moderator variables related to characteristics of the ITS, students, research setting, outcome assessments, and research methods, we found no substantive, significant differences among levels of the moderators under a random-effects model.

Comparison With Previous Quantitative Reviews

In broad terms, our results agree with prior reviews by VanLehn (2011), who investigated the relative effectiveness of ITS in STEM subjects, and Steenbergen-Hu and Cooper (2014), who investigated its use in all college-level subjects. Our mean effect size for postsecondary education ($g = .43$) was only slightly greater than the mean effect size ($g = .37$) reported by Steenbergen-Hu and Cooper (2014). However, our mean effect sizes for levels of K–12 education were all markedly greater than those reported in the meta-analysis of ITS effects in K–12 mathematics by Steenbergen-Hu and Cooper (2013). In discussing the contrasting results of their meta-analyses on the use of ITS by K–12 mathematics students and college students, Steenbergen-Hu and Cooper (2014) speculated that "ITS may function better for more mature students who have sufficient prior knowledge, self-regulation skills, learning motivation, and experiences with computers" (p. 342). Our analysis, which directly compared ITS effect sizes at four levels of schooling found no evidence for that hypothesis.

Our comparison of ITS with one-to-one human tutoring produced a non-significant mean effect size ($g = -.11$) similar to the effect sizes reported by VanLehn (2011) and Steenbergen-Hu and Cooper (2014) for human tutoring as a control condition. Unlike the previous reviews, our meta-analysis coded small-group human tutoring as a separate category of control treatment for which we found a nonsignificant mean effect size of $g = .05$ under a random-effects model.

Unlike Steenbergen-Hu and Cooper (2014), we separated from the main analysis studies with no-treatment control conditions. Whereas we found $g = 1.23$ relative to no-treatment controls, Steenbergen-Hu and Cooper found $g = .90$, and VanLehn (2011) found .40 and .76 for differing levels of interaction granularity. Although these discrepancies could be cause for concern, the important fact which serves as a reality check on the ITS evaluation enterprise is that in every review the mean effect sizes which compare ITS to no-treatment controls are greater or equal to the largest mean effect sizes which compare ITS to an alternate form of instruction.

Quality of Reporting

We found considerable room for improvement in how fundamental features of the primary research were reported. Basic statistics such as means and standard deviations were not reported in about a third of the studies, and reliabilities of outcome measures were reported in only a few cases. In many studies, reporting was also insufficient for methodological features such as attrition, whether participants were randomly assigned to treatments, format and provenance of achievement tests, and duration of treatment. The challenge in raising the quality of research reporting in an inter-disciplinary field such as ITS can be attributed to the lack of a shared understanding of methodological standards among researchers and editors, and the dissemination of ITS evaluation research in a remarkably eclectic body of journals including *Issues in Accounting Education*, *Thinking Skills and Creativity*, and *Methods of Information in Medicine*. This challenge, inherent to interdisciplinary research reporting, is reflected in the statement of scope and standards of the *International Journal of Artificial Intelligence in Education*, which remarks

> if a paper presents a behavioural study of students using some system to support claims about improved learning, then it must conform to the standards developed in behavioural science. . . . On the other hand, it is not reasonable to expect that authors will meet all the standards of all disciplines outside their main focus. (*International Journal of Artificial Intelligence in Education*, n.d., para. 5)

We advocate that journal editors specify their requirements for reporting research design, sample size, attrition, score reliabilities, means and standard deviations for quantitative educational research and also publish articles that inform their readership on how contemporary methodological practices such as "the new statistics" (Cumming, 2012) relate to their discipline.

One difficulty we encountered in conducting this meta-analysis was the lack of common terminology for describing and reporting the designs of individual ITS as well as inconsistent practices in selecting which ITS features should be described in an evaluation report. For example, some researchers reported that their ITS used model tracing but not did indicate whether misconceptions were modeled, whether knowledge tracing was used, and whether the ITS adaptively selected problems. Often, the method used for student modeling was not described in relation to other ITS and important features of the system's design and behavior were not reported. We speculate that developing a taxonomy of ITS design that could underpin a reporting standard would accelerate advances in ITS research. Certainly, more precise reporting of ITS design that draws from a common conceptual framework and terminology would greatly assist meta-analysts to pick apart design features and compare their effects on learning outcomes.

Can Evaluation Research Contribute to a Theory of ITS Design?

Each of the studies we examined evaluated a single ITS consisting of a complex set of inter-related features, many of which were necessarily unreported by or even unknown to the primary authors. In most cases the evaluations were performed to bear on the decision of whether to deploy the ITS or as a holistic evaluation of a software engineering project. Rarely were the studies we analyzed conceived as research into a theoretical question about the relationship between ITS and learning outcomes. Nevertheless, a meta-analysis of evaluation studies can categorize the primary studies according to theoretically significant features and observe relationships between the features and learning outcomes that were not considered in the primary work. Although none of the primary studies we analyzed compared a version of their ITS that modeled student misconceptions with another version that did not, we were able to code for misconception modeling as a moderator variable and assess its influence on effect size. As it turned out, none of the ITS characteristics we coded including type of ITS, misconception

modeling and feedback, were found to reliably influence effect size under a random-effects model (although a fixed-effect model found that one type of ITS, Bayesian network modeling, had a significantly greater influence on effect size). It is notable that when we reread the studies in which the ITS did not provide response feedback we found that in each case the primary adaptive feature was individualized task selection, an observation that suggests individualized task selection may offer benefits comparable to the well-established, positive effects of feedback on learning (Hattie & Timperley, 2007).

Most of the research we analyzed reported independent and dependent variables, but not intervening process variables (i.e., measures observed during the learning process) that might help to explain observed effects or lack thereof. Although recent work in educational data mining indicates that process variables are gaining a more prominent position in the study of ITS (Winne & Baker, 2013), such variables are rarely reported in empirical evaluations of ITS effectiveness. When a process variable was reported in the studies we analyzed, it was often only meaningful in the context of the particular learning task of the study that reported it. As a consequence, when we determine that ITS outperformed other methods of computer-based instruction there is little we can do in a meta-analysis to account for the effect at the level of computer-student interaction. Similarly, if we seek an explanation for how Bayesian network modeling might outperform other ITS designs, there are no common, interface-level data that could show whether the technical distinctions among the major types of ITS are manifested across the research base in consistent, differentiated patterns of interaction with students.

What Meta-Analysis Can Tell Us About ITS

Despite the great variety of conditions under which ITS were found to be more effective than other modes of instruction, these results do not support the direct inference that ITS should replace other modes currently in use. In the studies we reviewed there was the rational tendency to evaluate an ITS relative to the goals and scope of the project which created it. Because a project's goals were determined by what might feasibly be accomplished by an ITS, the studies had an inbuilt bias toward instructional conditions and roles in which ITS could compete favorably with other modes. Another possible source of bias is that the development of an ITS, like any major instructional design project, typically involves more detailed attention to learning goals, materials and activities than the more typical instructional practices represented by the control conditions in the analyzed evaluations. It may be that the significant effect sizes we found were due as much the kind of intensive instructional planning and analysis that can enhance any instructional approach than to the particular features of ITS.

What the results of this meta-analysis do provide is strong evidence that in some situations ITS can successfully complement and substitute for other instructional modes, and that these situations exist at all educational levels and in many common academic subjects. The results do not reliably support any specific explanations for the effectiveness of ITS, but they are consistent with an attribution to the most frequently implemented ITS features enabled by student modeling, namely highly individualized task selection, prompting and response feedback. That ITS were found to be relatively effective whether or not they modeled frequent student misconceptions suggests the need for comparative research on the conditions under which misconception modeling adds value to individualized instruction.

This meta-analysis and previous reviews by Steenbergen-Hu and Cooper (2013, 2014) examined evaluation research in which the use of ITS was compared to a variety of other modes of instruction. While reviews of this type are useful in marking general progress in the capabilities of ITS, a more powerful use of meta-analysis to drive those capabilities forward may be to review comparisons between ITS. This strategy would be especially informative when analyzing studies that compare two or more versions of the same ITS such that each version represents a theoretically informed design variation. VanLehn (2011) adopted elements of this strategy to investigate the effects of interaction granularity on learning outcomes, and full meta-analyses comparing different versions of the same systems could be used to investigate many other potentially effective ITS features such as animated pedagogical agents (Baker et al., 2006), misconception modeling (Myneni, Narayanan, Rebello, Rouinfar, & Pumtambekar, 2013) and metacognitive prompts (Wu & Looi, 2012). We believe that such strategic application of meta-analysis to the end products of ITS research, development, and evaluation can inform and advance the design science of ITS.

References

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

*Abu-Naser, S. S. (2009). Evaluating the effectiveness of the CPP-Tutor, an Intelligent Tutoring System for students learning to program in C++. *Journal of Applied Sciences Research, 5,* 109–114.

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis on the cognitive correlates of bilingualism. *Review of Educational Research, 80,* 207–245. doi: 10.3102/0034654310368803

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology, 104,* 250–263. doi:10.1037/a0026147

*Aist, G., Mostow, J., Tobin, B., Burkhead, P., Cuneo, A., Junker, B., & Sklar, M. B. (2001). Computer-assisted oral reading helps third graders learn vocabulary better than a classroom control—About as well as one-on-one human-assisted oral reading. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education: AI-ED in the wired and wireless future* (pp. 267–277). San Antonio, TX: IOS Press.

*Albacete, P. L., & VanLehn, K. A. (2000). Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 25–30). Mahwah, NJ: Erlbaum.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

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

Anderson, J. R., & Lebière, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

*Arbuckle, W. J. (2005). *Conceptual understanding in a computer assisted Algebra I classroom* (Unpublished doctoral dissertation). Retrieved from ProQuest database. (UMI No. 320318)

*Argotte, L., Arroyo-Figueroa, G., & Noguez, J. (2011). SI-APRENDE: An intelligent learning system based on SCORM learning objects for training power systems operators. In K. G. Mehrotra, C. Mohan, J. C. Oh, P. K. Varshney, & M. Ali (Eds.), *Developing concepts in applied*

intelligence (Studies in Computational Intelligence, Vol. *363,* pp. 33–38). doi:10.1007/978-3-642-21332-8_5

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

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

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13,* 111–127. doi:10.2190/9LMD-3U28-3A0G-FTQT

Baker, R. S., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., . . . Beck, J. E. (2006). Adapting to when students game an intelligent tutoring system. In M. Ikeda, K. D. Ashley, & T. Chan (Eds.), *Intelligent tutoring systems* (pp. 392–401). Berlin, Germany: Springer Berlin-Heidelberg.

Barr, A., Beard, M., & Atkinson, R. C. (1976). The computer as a tutorial laboratory: The Stanford BIP Project. *International Journal of Man-Machine Studies, 8,* 567–596. doi:10.1016/S0020-7373(76)80021-1

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

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

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* doi:10.1002/9780470743386

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155–192. doi:10.1207/s15516709cog0202_4

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

Carbonell, J. (1970). AI in CAI: An artificial intelligence approach to computer aided instruction. *Science, 167,* 190–202.

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

*Chambers, B., Abrami, P., Tucker, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2008). Computer assisted tutoring in success for all: Reading outcomes for first graders. *Journal of Research on Educational Effectiveness, 1,* 120–137. doi:10.1080/19345740801941357

*Chen, C. M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education, 51,* 787–814. doi:10.1016/j.compedu.2007.08.004

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using item response theory. *Computers & Education, 44,* 237–255. doi:10.1016/j.compedu.2004.01.006

Chen, L. H. (2011). Enhancement of student learning performance using personalized diagnosis and remedial learning system. *Computers & Education, 56,* 289–299. doi:10.1016/j.compedu.2010.07.015

*Chien, T. C., Yunus, A. S. M., Ali, W. Z. W., & Bakar, A. R. (2008). The effect of an Intelligent Tutoring System (ITS) on student achievement in algebraic expression. *International Journal of Instruction, 1,* 25–38.

*Chin, D. B., Dohmen, I. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L. (2010). Preparing students for future learning with Teachable Agents. *Educational Technology Research and Development, 58,* 649–669. doi:10.1007/s11423-010-9154-5

Cohen, P. A., & Dacanay, L. S. (1992). Computer-based instruction and health professions education: A meta-analysis of outcomes. *Evaluation & the Health Professions, 15,* 259–281. doi:10.1177/016327879201500301

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., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12,* 371–417. doi:10.1023/A:1021258506583

*Conati, C., & VanLehn, K. (1999). Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. In S. P. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 297–304). Amsterdam, the Netherlands: IOS Press.

*Conati, C., & Zhao, X. (2004). Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game. In N. J. Nunes & C. Rich (Eds.), *Proceedings of the International Conference on Intelligent User Interfaces (IUI-04)* (pp. 6–13). New York, NY: ACM Press.

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.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science, 34,* 399–422. doi:10.1007/s11251-005-5774-2

*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 Berlin-Heidelberg.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent Tutoring Systems. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human–computer interaction* (pp. 849–874). Amsterdam, the Netherlands: Elsevier.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Dede, C. (1986). A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man-Machine Studies, 24,* 329–353. doi:10.1016/S0020-7373(86)80050-5

Denson, N. (2009). Do curricular and co-curricular diversity activities influence racial bias? A meta-analysis. *Review of Educational Research, 79,* 805–838. doi:10.3102/0034654309331551

*Fossati, D., Di Eugenio, B., Brown, C., & Ohlsson, S. (2008). Learning linked lists: Experiments with the iList system. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Intelligent Tutoring Systems: 9th International Conference, ITS 2008* (pp. 80–89). Berlin, Germany: Springer.

*Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., Chen, L., & Cosejo, D. (2009, July). *I learn from you, you learn from me: How to make iList learn from students*. Paper presented at the AIED 2009: The 14th International Conference on Artificial Intelligence in Education, Brighton, United Kingdom.

Graesser, A. C., Chipman, P., Haynes, B., & Olney, A. (2005). AutoTutor: An Intelligent Tutoring System with mixed-initiative dialogue. *IEEE Transactions on Education, 48,* 612–618. doi:10.1109/TE.2005.856149

*Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., . . . TRG. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1–5). Boston, MA: Cognitive Science Society.

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

[*]Graff, M., Mayer, P., & Lebens, M. (2008). Evaluating a web based Intelligent Tutoring System for mathematics at German lower secondary schools. *Education and Information Technologies, 13,* 221–230. doi: 10.1007/s10639-008-9062-z

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.

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

Halpern, D. F., Millis, K., Graesser, A. C., & Butler, H. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity, 7,* 93–100. doi: 10.1016/j.tsc.2012.03.006

[*]Han, K., Lee, E., & Lee, Y. (2010). The impact of a peer-learning agent based on pair programming in a programming course. *IEEE Transactions on Education, 53,* 318–327. doi:10.1109/TE.2009.2019121

Hannum, W. (1986). Techniques for creating computer-based instructional text: Programming languages, authoring languages, and authoring systems. *Educational Psychologist, 21,* 293–314. doi:10.1207/s15326985ep2104_4

Hartley, J. R., & Sleeman, D. H. (1973). Towards more intelligent teaching systems. *International Journal of Man-Machine Studies, 5,* 215–236. doi:10.1016/S0020-7373(73)80033-1

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–112. doi:10.3102/003465430298487

[*]Hayes-Roth, B., Saker, R., & Amano, K. (2010). Automating individualized coaching and authentic role-play practice for brief intervention training. *Methods of Information in Medicine, 49,* 406–411. doi: 10.3414/ME9311

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. doi:10.1037/1082-989X.3.4.486

[*]Heffernan, N. T. (2003). *Web-based evaluations showing both cognitive and motivational benefits of the Ms. Lindquist Tutor*. Paper presented at the 11th International Conference on Artificial Intelligence in Education, Sydney, Australia.

[*]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 2007 Annual Meeting of the American Educational Research Association (AERA), Chicago, IL.

Hughes, M. G., Day, E., Wang, X., Schuelke, M. J., Arsenault, M. L., Harkrider, L. N., & Cooper, O. D. (2013). Learner-controlled practice difficulty in the training of a complex task: Cognitive and motivational mechanisms. *Journal of Applied Psychology, 98,* 80–98. doi:10.1037/a0029821

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

[*]*International Journal of Artificial Intelligence in Education*. (n.d.). Journal scope and standards. Retrieved from http://ijaied.org/journal/scope/

[*]Jeremic, Z., Jovanovic, J., & Gasevic, D. (2009). Evaluating an Intelligent Tutoring System for design patterns: The DEPTHS experience. *Educational Technology & Society, 12,* 111–130.

[*]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

Kodaganallur, V., Weitz, R. B., & Rosenthal, D. (2005). A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal of Artificial Intelligence in Education, 15,* 117–144.

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

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

[*]Kozierkiewicz-Hetmańska, A., & Nguyen, N. T. (2011). A method for learning scenario determination and modification in Intelligent Tutoring Systems. *International Journal of Applied Mathematics and Computer Science, 21,* 69–82. doi:10.2478/v10006-011-0005-2

[*]Kumar, A. N. (2002). Model-based reasoning for domain modeling in a web-based Intelligent Tutoring System to help students learn to debug C ++ programs. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of Intelligent Tutoring Systems (ITS-2002)* (LNCS 2363, pp. 792–801). Berlin, Germany: Springer-Verlag Berlin-Heidelberg.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi: 10.1080/01638539809545028

[*]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

[*]Lanzilotti, R., & Roselli, T. (2007). An experimental evaluation of Logiocando, an intelligent tutoring hypermedia system. *International Journal of Artificial Intelligence in Education, 17,* 41–56.

[*]Lesta, L., & Yacef, K. (2002). An intelligent teaching-assistant system for logic. In S. Cerri & F. Paraguo (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems, ITS'02* (pp. 421–431). Biarritz, France: Springer-Verlag.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Martin, F., Klein, J. D., & Sullivan, H. (2007). The impact of instructional elements in computer-based instruction. *British Journal of Educational Technology, 38,* 623–636. doi:10.1111/j.1467-8535.2006.00670.x

[*]McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In G. Biswas, S. Bull, J. Kay & A. Mitrovic (Eds.), *AIED 2011* (LNAI 6738, pp. 222–229). Berlin, Germany: Springer-Verlag Berlin-Heidelberg.

[*]McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171. doi:10.2190/1RU5-HDTJ-A5C8-JVWE

Millán, E., Loboda, T., & Pérez-de-la-Cruz, J. L. (2010). Bayesian networks for student model engineering. *Computers & Education, 55,* 1663–1683. doi:10.1016/j.compedu.2010.07.010

[*]Mills-Tettey, G. A., Mostow, J., Dias, M. B., Sweet, T. M., Belousov, S. M., Dias, M. F., & Gong, H. (2010). Improving child literacy in Africa: Experiments with an automated reading tutor. *Information Technologies and International Development, 6,* 1–19.

[*]Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education, 13*(2–4), 171–195.

Mitrovic, A., & Djordjevic-Kajan, S. (1995). Interactive reconstructive student modeling: A machine-learning approach. *International Journal of Human-Computer Interaction, 7,* 385–401. doi:10.1080/10447319509526132

Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring. *IEEE Intelligent Systems, 22,* 38–45. doi:10.1109/MIS.2007.74

*Mitrovic, A., Martin, B., Suraweera, P., Zakharov, K., Milik, N., Holland, J., & McGuigan, N. (2009). ASPIRE: An authoring system and deployment environment for constraint-based tutors. *International Journal of Artificial Intelligence in Education, 19,* 155–188.

*Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database language. *International Journal on Artificial Intelligence in Education, 10*(3–4), 238–256.

Mitrovic, A., Williamson, C., Bebbington, A., Mathews, M., Suraweera, P., Martin, B., . . . Holland, J. (2011). Thermo-Tutor: An intelligent tutoring system for thermodynamics. In *Global Engineering Education Conference (EDUCON), 2011 IEEE* (pp. 378–385). Amman, Jordan: IEEE.

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

*Mostow, J., Aist, G., Bey, J., Burkhead, P., Cuneo, A., Junker, B., . . . Wilson, S. (2002). *Independent versus computer-assisted reading: Equal-time comparison of sustained silent reading to an automated reading tutor that listens*. Paper presented at the 9th Annual Meeting of the Society for the Scientific Study of Reading, Chicago, IL.

Myneni, L. S., Narayanan, N. H., Rebello, S., Rouinfar, A., & Pumtambekar, S. (2013). An interactive and intelligent learning system for physics education. *IEEE Transactions on Learning Technologies, 6,* 228–239. doi:10.1109/TLT.2013.26

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76,* 413–448. doi:10.3102/00346543076003413

Ohlsson, S. (1994). Constraint based user modeling. In J. E. Greer & G. McCalla (Eds.), *Student modeling: The key to individualized knowledge-based instruction* (pp. 167–189). doi:10.1007/978-3-662-03037-0_7

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157–159. doi:10.2307/1164923

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

*Patel, K. A., & Russell, D. (2000). A multi-institutional evaluation of Intelligent Tutoring Tools in Numeric Disciplines. *Educational Technology and Society, 3,* 66–74.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

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

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

*Pinkwart, N., Ashley, K., Lynch, C., & Aleven, V. (2009). Evaluating an Intelligent Tutoring System for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education, 19,* 401–424.

*Poulsen, R. (2004). *Tutoring bilingual students with an automated reading tutor that listens: Results of a two-month pilot study* (Unpublished master's thesis). DePaul University, Chicago, IL.

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

*Ramadhan, H. A. (2000). DISCOVER: An intelligent discovery programming system. *Cybernetics & Systems: An International Journal, 31,* 87–114. doi:10.1080/019697200124937

*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. W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In T. Hirashima, U. Hoppe, & S. S. C. Young (Eds.), *Frontiers in artificial intelligence and applications, 162: Supporting learning flow through integrative technologies* (pp. 13–20). Amsterdam, the Netherlands: IOS Press.

Roll, I., Baker, R. S., Aleven, V., & Koedinger, K. (2004). A metacognitive ACT-R model of students' learning strategies in Intelligent Tutoring Systems. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 854–856). Berlin, Germany: Springer-Verlag.

*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.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R. (1995). Critiquing Pygmalion: A 25-year perspective. *Current Directions in Psychological Science, 4,* 171–172. doi:10.1111/1467-8721.ep10772607

*Rowe, N. C., & Schiavo, S. (1998). An intelligent tutor for intrusion detection on computer systems. *Computers and Education, 31,* 395–404. doi:10.1016/S0360-1315(98)00049-9

Santopietro, G. (2011). Effects of computerized formative assessment on student learning in economics. *Journal of Learning in Higher Education, 7,* 7–15.

*Schulze, K. G., Shelby, R. N., Treacy, D. J., Wintersgill, M. C., VanLehn, K., & Gertner, A. (2000). Andes: An active learning Intelligent Tutoring System for Newtonian physics. *Themes in Education, 1,* 115–136.

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

*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., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility* (ETS Research Report, RR-07-26). Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2007/hsmy

Shute, V. J., & Psotka, J. (1996). Intelligent Tutoring Systems: Past, present, and future. In D. Jonassen (Eds.), *Handbook of research for educational communications and technology* (pp. 570–600). New York, NY: Macmillan.

Sleeman, D. H., & Brown, J. S. (Eds.). (1979). Intelligent tutoring systems [Special issue]. *International Journal of Man–Machine Studies, 11,* 1–3. doi:10.1016/S0020-7373(79)80002-4

Sleeman, D. H., & Brown, J. S. (Eds.). (1982). *Intelligent Tutoring Systems*. London, England: Academic Press.

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

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,* 97–128. doi:10.3102/0034654310378174

Sottilare, R., Graesser, A., Hu, X., & Holden, H. (Eds.). (2013). *Design recommendations for Intelligent Tutoring Systems*. Orlando, FL: U.S. Army Research Laboratory.

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

*Stankov, S., Rosić, M., Žitko, B., & Grubišić, A. (2008). TEx-Sys model for building Intelligent Tutoring Systems. *Computers & Education, 51,* 1017–1036. doi:10.1016/j.compedu.2007.10.002

Steenbergen-Hu, S., & Cooper, H. (2013, September 9). A meta-analysis of the effectiveness of Intelligent Tutoring Systems on K–12 students' mathematical learning. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0032447

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

*Suraweera, P., & Mitrovic, A. (2002). Kermit: A constraint-based tutor for database modeling. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.), *Intelligent Tutoring Systems: 6th International Conference* (pp. 377–387). Berlin, Germany: Springer-Verlag.

*Suraweera, P., & Mitrovic, A. (2004). An Intelligent Tutoring System for entity relationship modelling. *International Journal of Artificial Intelligence in Education (IJAIED), 14*(3–4), 375–417.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn & Bacon.

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

*Tsiriga, V., & Virvou, M. (2004). Evaluating the intelligent features of a web-based intelligent computer assisted language learning system. *International Journal on Artificial Intelligence Tools, 13,* 411–425. doi:10.1142/S0218213004001600

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., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons learned. *International Journal of Artificial Intelligence in Education, 15,* 1–47.

*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

*Veermans, K. H., de Jong, T., & van Joolingen, W. R. (2000). Promoting self-directed learning in simulation based discovery learning environments through intelligent support. *Interactive Learning Environments, 8,* 229–255. doi:10.1076/1049-4820(200012)8:3;1-D;FT229

Veldkamp, B. P., Matteucci, M., & Eggen, T. J. H. M. (2011). Computerized adaptive testing in computer assisted learning? In S. De Wannemacker, G. Clarebout, & P. De Causmaecker (Eds.), *Interdisciplinary approaches to adaptive learning. A look at the neighbours* (Vol. *126,* pp. 28–39). doi:10.1007/978-3-642-20074-8_3

Wang, H. C., Rosé, C. P., & Chang, C. Y. (2011). Agent-based dynamic support for learning from collaborative brainstorming in scientific inquiry. *International Journal of Computer-Supported Collaborative Learning, 6,* 371–395. doi:10.1007/s11412-011-9124-x

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66,* 98–112. doi:10.1016/j.ijhcs.2007.09.003

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

*Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development, 60,* 987–1013. doi:10.1007/s11423-012-9263-4

Winne, P. H., & Baker, R. S. J. D. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining, 5,* 1–8.

*Wisher, R. A., MacPherson, D. H., Abramson, L. J., Thorndon, D. M., & Dees, J. J. (2001). *The virtual sand table: Intelligent tutoring for field artillery training* (ARI Research Report No. 1768). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

*Woo, C. W., Evens, W. M. W., Freedman, R., Glass, M., Seop Shim, L., Zhang, Y., . . . Michael, J. (2006). An Intelligent Tutoring System that generates a natural language dialogue using dynamic multi-level planning. *Artificial Intelligence in Medicine, 38,* 25–46. doi:10.1016/j.artmed.2005.10.004

Wu, L., & Looi, C. K. (2012). Agent prompts: Scaffolding for productive reflection in an intelligent learning environment. *Journal of Educational Technology & Society, 15,* 339–353.

Received March 28, 2013
Revision received May 1, 2014
Accepted May 3, 2014 ■