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.
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).
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, Alevne, & 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 (Kodagaddam, 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, & Suraewera, 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 presentation 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 Merrienboer, 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 models 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 2011 and 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 (Grubisic, 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 use 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;
(d) were publicly available, online or in library archives;

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. However, 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 group interacted with an ITS via text and another group that interacted with the same ITS via an animated agent, the reference sections of review articles on Intelligent Tutoring Systems were manually 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).

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 ($\pm 3.3 \geq Z \geq 3.3; p < .001$) 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.

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.

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

<table>
<thead>
<tr>
<th>Moderator variables</th>
<th>Random-effects model</th>
<th>Fixed-effect model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of ITS</td>
<td>Effect size 95% CI</td>
<td>Effect size 95% CI</td>
</tr>
<tr>
<td>Model tracing</td>
<td>N k g+ SE Lower Upper</td>
<td>Qₙ df p F (%)</td>
</tr>
<tr>
<td>Principal instruction</td>
<td>4,172 13 0.44 0.09 0.17 0.46 0.05 0.10 0.20 0.59</td>
<td>2.41 .79 32.38 &lt;.001</td>
</tr>
<tr>
<td>Integrated class instruction</td>
<td>6,245 12 0.34 0.07 0.17 0.49 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Separate in-class activities</td>
<td>4,172 13 0.44 0.09 0.17 0.46 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Supplementary after-class instruction</td>
<td>5,976 21 0.35 0.07 0.22 0.47 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Not reported</td>
<td>1,798 13 0.44 0.09 0.17 0.46 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Feedback provided?</td>
<td>Effect size 95% CI</td>
<td>Effect size 95% CI</td>
</tr>
<tr>
<td>No</td>
<td>4,172 13 0.44 0.09 0.17 0.46 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Yes</td>
<td>11,728 86 0.42 0.04 0.34 0.50 0.05 0.15 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Not reported</td>
<td>1,182 11 0.21 0.10 0.02 0.41 0.05 0.15 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Model misconception?</td>
<td>Effect size 95% CI</td>
<td>Effect size 95% CI</td>
</tr>
<tr>
<td>No</td>
<td>4,172 13 0.44 0.09 0.17 0.46 0.03 0.10 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Yes</td>
<td>9,911 58 0.40 0.05 0.31 0.49 0.05 0.15 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
<tr>
<td>Not reported</td>
<td>2,902 28 0.42 0.10 0.23 0.61 0.05 0.15 0.20 0.59</td>
<td>4.45 .10 1.53 &lt;.001</td>
</tr>
</tbody>
</table>

Note. CI = confidence interval. *p < .05.
vvention 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, $Q_{(5)} = 36.37, p < .001$, 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, $Q_{(5)} = 32.38, p < .001$. 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 |
|---------------------------------------------------|---------------------------------------------------|---------------------------------|----------------------------------|
| Moderator variables                               | Random-effects model                              | Fixed-effect model              |
|                                                   | Effect size | 95% CI | Q_0 | p | Effect size | 95% CI | Q_0 | p |
| Outcome constructs                                | N | k | g+ | SE | Lower | Upper |      |     | g+ | SE | Lower | Upper |      |     |
| Retention                                         | 3,922 | 33 | 0.35 | 0.07 | 0.22 | 0.48 | 0.35 | 0.03 | 0.28 | 0.41 | 3.56 | .31 |
| 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, \( Q_0(5) = 25.74, p < .001 \). 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 |
|---------------------------------------------------|---------------------------------------------------|---------------------------------|----------------------------------|
| Moderator variables                               | Random-effects model                              | Fixed-effect model              |
|                                                   | Effect size | 95% CI | Q_0 | p | Effect size | 95% CI | Q_0 | p |
| Setting                                           | N | k | g+ | SE | Lower | Upper |      |     | g+ | SE | Lower | Upper |      |     |
| Laboratory                                        | 1,596 | 26 | 0.29 | 0.07 | 0.15 | 0.43 | 0.26 | 0.05 | 0.16 | 0.36 | 4.29 | .04 |
| Classroom                                         | 12,725 | 81 | 0.44 | 0.04 | 0.36 | 0.52 | 0.37 | 0.02 | 0.33 | 0.41 |      |     |
| Continent                                         | 5.39 | .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.
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, \( Q(3) = 50.67, p < .001 \), 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 in 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, \( Q(3) = 108.17, p < .001 \). 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, \( Q(3) = 30.33, p < .001 \), 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, \( Q(3) = 14.92, p < .001 \). 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, Qg(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, Qg(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, Qg(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 α = .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 ITSs were found to be more effective than other modes of instruction, these results do not support the direct inference that ITSs 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 ITSs 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 as the kind of intensive instructional planning and analysis that can enhance any instructional approach than to the particular features of ITSs.

What the results of this meta-analysis do provide is strong evidence that in some situations ITSs 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 ITSs, 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 ITSs 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 ITSs, a more powerful use of meta-analysis to drive those capabilities forward may be to review comparisons between ITSs. 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, Rebell, Rouinfar, & Pumontbekar, 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.

In INTELLIGENT TUTORING SYSTEMS