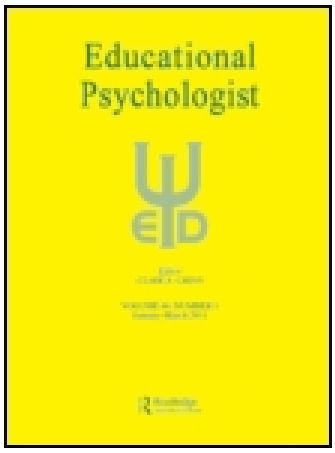

Educational Psychologist
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hedp20

The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems

KURT VanLEHN [a]
[a] Computing, Informatics and Decision Systems Engineering Arizona State University
Published online: 17 Oct 2011.

To cite this article: KURT VanLEHN (2011) The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, Educational Psychologist, 46:4, 197-221, DOI: 10.1080/00461520.2011.611369

To link to this article: http://dx.doi.org/10.1080/00461520.2011.611369

Educational Psychologist, *46*(4), 197–221, 2011

ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2011.611369

Routledge
Taylor & Francis Group

The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems

Kurt VanLehn

Computing, Informatics and Decision Systems Engineering
Arizona State University

This article is a review of experiments comparing the effectiveness of human tutoring, computer tutoring, and no tutoring. "No tutoring" refers to instruction that teaches the same content without tutoring. The computer tutoring systems were divided by their granularity of the user interface interaction into answer-based, step-based, and substep-based tutoring systems. Most intelligent tutoring systems have step-based or substep-based granularities of interaction, whereas most other tutoring systems (often called CAI, CBT, or CAL systems) have answer-based user interfaces. It is widely believed as the granularity of tutoring decreases, the effectiveness increases. In particular, when compared to No tutoring, the effect sizes of answer-based tutoring systems, intelligent tutoring systems, and adult human tutors are believed to be $d = 0.3$, 1.0, and 2.0 respectively. This review did not confirm these beliefs. Instead, it found that the effect size of human tutoring was much lower: $d = 0.79$. Moreover, the effect size of intelligent tutoring systems was 0.76, so they are nearly as effective as human tutoring.

From the earliest days of computers, researchers have strived to develop computer tutors that are as effective as human tutors (S. G. Smith & Sherwood, 1976). This review is a progress report. It compares computer tutors and human tutors for their impact on learning gains. In particular, the review focuses on experiments that compared one type of tutoring to another while attempting to control all other variables, such as the content and duration of the instruction. The next few paragraphs define the major types of tutoring reviewed here, starting with human tutoring. Current beliefs about the relative effectiveness of the types of tutoring are then presented, followed by eight common explanations for these beliefs. Building on these theoretical points, the introduction ends by formulating a precise hypothesis, which is tested with meta-analytic methods in the body of the review.

Although there is a wide variety of activities encompassed by the term "human tutoring," this article uses "human tutoring" to refer to an adult, subject-matter expert working synchronously with a single student. This excludes many other kinds of human tutoring, such as peer tutoring, cross-age tutoring, asynchronous online tutoring (e.g., e-mail or forums), and problem-based learning where a "tutor" works with small group of students. Perhaps the major reason why computer tutor developers have adopted adult, one-on-one, face-to-face tutoring as their gold standard is a widely held belief that such tutoring is an extremely effective method of instruction (e.g., Graesser, VanLehn, Rose, Jordan, & Harter, 2001). Computer developers are not alone in their belief. For instance, parents sometimes make great sacrifices to hire a private tutor for their child.

Even within this restricted definition of human tutoring, one could make distinctions. For instance, synchronous human tutoring includes face-to-face, audio-mediated, and text-mediated instantaneous communication. Human tutoring can be done as a supplement to the students' classroom instruction or as a replacement (e.g., during home schooling). Tutoring can teach new content, or it can also be purely remedial. Because some of these distinctions are rather difficult to make precisely, this review allows "human tutoring" to cover all these subcategories. The only proviso is that these variables be controlled during evaluations. For instance, if the human tutoring is purely remedial, then the computer tutoring to which it is compared should be purely remedial as well. In short, the human tutoring considered in this review

Correspondence should be addressed to Kurt VanLehn, Computing, Informatics and Decision Systems Engineering, Arizona State University, PO Box 878809, 699 South Mill Avenue, Tempe, AZ 85287-8809. E-mail: kurt.vanlehn@asu.edu

includes all kinds of one-on-one, synchronous tutoring done by an adult, subject-matter expert.

In contrast to human tutoring, which is treated as one monolithic type, two technological types of computer tutoring are traditionally distinguished. The first type is characterized by giving students immediate feedback and hints on their *answers*. For instance, when asked to solve a quadratic equation, the tutee works out the answer on scratch paper, enters the number, and is either congratulated or given a hint and asked to try again. This type of tutoring system has many traditional names, including Computer Aided-Instruction (CAI), Computer-Based Instruction, Computer-Aided Learning, and Computer-Based Training.

The second type of computer tutoring is characterized by giving students an electronic form, natural language dialogue, simulated instrument panel, or other user interface that allows them to enter the steps required for solving the problem. For instance, when asked to solve a quadratic equation, the student might first select a method (e.g., completing the square) from a menu; this causes a method-specific form to appear with blanks labeled "the coefficient of the linear term," "the square of half the coefficient of the linear term," and so on. Alternatively, the student may be given a digital canvas to write intermediate calculations on, or have a dialogue with an agent that says, "Let's solve this equation. What method should we use?" The point is only that the intermediate steps that are normally written on paper or enacted in the real world are instead done where the tutoring system can sense and interpret them. The tutoring system gives feedback and hints on each step. Some tutoring systems give feedback and hints immediately, as each step is entered. Others wait until the student has submitted a solution, then either mark individual steps as correct or incorrect or conduct a debriefing, which discusses individual steps with the student. Such tutoring systems are usually referred to as Intelligent Tutoring Systems (ITS).

A common belief among computer tutoring researchers is that human tutoring has an effect size of $d = 2.0$ relative to classroom teaching without tutoring (Bloom, 1984; Corbett, 2001; Evens & Michael, 2006; Graesser et al., 2001; VanLehn et al., 2007; Woolf, 2009). In contrast, CAI tends to produce an effect size of $d = 0.31$ (C. Kulik & Kulik, 1991). Although no meta-analyses of ITS currently exist, a widely cited review of several early ITS repeatedly found an average effect size of $d = 1.0$ (Anderson, Corbett, Koedinger, & Pelletier, 1995). Figure 1 displays these beliefs as a graph of effect size versus type of tutoring. Many evaluations of ITS have been done recently, so it is appropriate to examine the claims of Figure 1.

This article presents a review that extends several earlier meta-analyses of human and computer tutoring (Christmann & Badgett, 1997; Cohen, Kulik, & Kulik, 1982; Fletcher-Flinn & Gravatt, 1995; Fletcher, 2003; J. Kulik, Kulik, & Bangert-Drowns, 1985; J. Kulik, Kulik, & Cohen, 1980; J. A. Kulik, Bangert, & Williams, 1983; G. W. Ritter, Barnett, Denny, & Albin, 2009; Scruggs & Richter, 1985; Wasik, 1998). All the preceding meta-analyses have compared just two types of instruction: with and without tutoring. Some meta-analyses focused on computer instruction, and some focused on human instruction, but most meta-analyses focused on just one type of tutoring. In contrast, this review compares five types of instruction: human tutoring, three types of computer tutoring, and no tutoring. The "no tutoring" method covers the same instructional content as the tutoring, typically with a combination of reading and problem solving without feedback. Specific examples of no tutoring and other types of tutoring are presented later in the "Three Illustrative Studies" section.

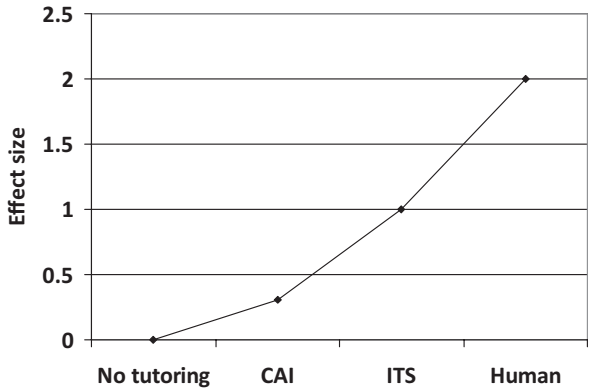

FIGURE 1 Common belief about effect sizes of types of tutoring. *Note.* CAI = Computer Aided-Instruction; ITS = Intelligent Tutoring Systems.

THEORY: WHY SHOULD HUMAN TUTORING BE SO EFFECTIVE?

It is commonly believed that human tutors are more effective than computer tutors when both teach the same content, so for developers of computer tutors, the key questions have always been, What are human tutors doing that computer tutors are not doing, and why does that cause them to be more effective? This section reviews some of the leading hypotheses.

1. Detailed Diagnostic Assessments

One hypothesis is that human tutors infer an accurate, detailed model of the student's competence and misunderstandings, and then they use this diagnostic assessment to adapt their tutoring to the needs of the individual student. This hypothesis has not fared well. Although human tutors usually know which correct knowledge components their tutees had not yet mastered, the tutors rarely know about their tutees' misconceptions, false beliefs, and buggy skills (M. T. H. Chi, Siler, & Jeong, 2004; Jeong, Siler, & Chi, 1997; Putnam, 1987). Moreover, human tutors rarely ask questions that could diagnose specific student miscon-

ceptions (McArthur, Stasz, & Zmuidzinas, 1990; Putnam, 1987). When human tutors were given mastery/nonmastery information about their tutees, their behavior changed, and they may become more effective (Wittwer, Nuckles, Landmann, & Renkl, 2010). However, they do not change their behavior or become more effective when they are given detailed diagnostic information about their tutee's misconceptions, bugs, and false beliefs (Sleeman, Kelly, Martinak, Ward, & Moore, 1989). Moreover, in one study, when human tutors simply worked with the same student for an extended period and could thus diagnosis their tutee's strengths, weaknesses, preferences, and so on, they were not more effective than when they rotated among tutees and thus never had much familiarity with their tutees (Siler, 2004). In short, human tutors do not seem to infer an assessment of their tutee that includes misconceptions, bugs, or false beliefs, nor do they seem to be able to use such an assessment when it is given to them. On the other hand, they sometimes infer an assessment of which correct conceptions, skills, and beliefs the student has mastered, and they can use such an assessment when it is given to them. In this respect, human tutors operate just like many computer tutors, which also infer such an assessment, which sometimes called an overlay model (VanLehn, 1988, 2008a).

2. Individualized Task Selection

Another hypothesis is that human tutors are more effective than computer tutors because they are better at selecting tasks that are just what the individual student needs in order to learn. (Here, "task" means a multiminute, multistep activity, such as solving a problem, studying a multipage text, doing a virtual laboratory experiment, etc.) Indeed, individualized task selection is part of one of the National Academy of Engineering's grand challenges (http://www.engineeringchallenges.org/cms/8996/9127.aspx). However, studies suggest that human tutors select tasks using a curriculum script, which is a sequence of tasks ordered from simple to difficult (M. T. H. Chi, Roy, & Hausmann, 2008; Graesser, Person, & Magliano, 1995; Putnam, 1987). Human tutors use their assessment of the student's mastery of correct knowledge to regulate how fast they move through the curriculum script. Indeed, it would be hard for them to have done otherwise, given that they probably lack a deep, misconception-based assessment of the student, as just argued. Some computer tutors use curriculum scripts just as human tutors do, and others use even more individualized methods for selecting tasks. Thus, on this argument, computer tutors should be *more* effective than human tutors. In short, individualized task selection is not a good explanation for the superior effectiveness of human tutors.

3. Sophisticated Tutorial Strategies

Another common hypothesis is that human tutors use sophisticated strategies, such as Socratic irony (Collins & Stevens, 1982), wherein the student who gives an incorrect answer is led to see that such an answer entails an absurd conclusion. Other such strategies include reciprocal teaching (Palinscar & Brown, 1984) and the inquiry method. However, studies of human tutors in many task domains with many degrees of expertise have indicated that such sophisticated strategies are rarely used (Cade, Copeland, Person, & D'Mello, 2008; M. T. H. Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Cho, Michael, Rovick, & Evens, 2000; Core, Moore, & Zinn, 2003; Evens & Michael, 2006; Fox, 1991, 1993; Frederiksen, Donin, & Roy, 2000; Graesser et al., 1995; Hume, Michael, Rovick, & Evens, 1996; Katz, Allbritton, & Connelly, 2003; McArthur et al., 1990; Merrill, Reiser, Merrill, & Landes, 1995; Merrill, Reiser, Ranney, & Trafton, 1992; Ohlsson et al., 2007; VanLehn, 1999; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). Thus, sophisticated tutorial strategies cannot explain the advantage of human tutors over computer tutors.

4. Learner Control of Dialogues

Another hypothesis is that human tutoring allows mixed initiative dialogues, so that the student can ask questions or change the topic. This contrasts with most tutoring systems, where student initiative is highly constrained. For instance, although students can ask a typical ITS system for help on a step, they can ask no other question, nor can they cause the tutor to veer from solving the problem. On the other hand, students are free to ask any question of human tutors and to negotiate topic changes with the tutor. However, analyses of human tutorial dialogues have found that although students take the initiative more than they do in classroom settings, the frequency is still low (M. T. H. Chi et al., 2001; Core et al., 2003; Graesser et al., 1995). For instance, Shah, Evens, Michael, and Rovick (2002) found only 146 student initiatives in *28 hr* of typed human tutoring, and in 37% of these 146 instances, students were simply asking the tutor whether their statement was correct (e.g., by ending their statement with "right?"). That is, there was about one nontrivial student initiative every 18 min. The participants were medical students being tutored as part of a high-stakes physiology course, so apathy is not a likely explanation for their low rate of question asking. In short, learners' greater control over the dialogue is not a plausible explanation for why human tutors are more effective than computer tutors.

5. Broader Domain Knowledge

Human tutors usually have much broader and deeper knowledge of the subject matter (domain) than computer tutors. Most computer tutors only "know" how to solve and coach the tasks given to the students. Human tutors can in principle discuss many related ideas as well. For instance, if a student finds a particular principle of the domain counterintuitive, the human tutor can discuss the principle's history, explain the

experimental evidence for it, and tell anecdotes about other students who initially found this principle counterintuitive and now find it perfectly natural. Such a discussion would be well beyond the capabilities of most computer tutors. However, such discussions seldom occur when human tutors are teaching cognitive skills (McArthur et al., 1990; Merrill et al., 1995; Merrill et al., 1992). When human tutors are teaching less procedural content, they often do offer deeper explanations than computer tutors would (M. T. H. Chi et al., 2001; Evens & Michael, 2006; Graesser et al., 1995), but M. T. H. Chi et al. (2001) found that suppressing such explanations did not affect the learning gains of tutees. Thus, although human tutors do have broader and deeper knowledge than computer tutors, they sometimes do not articulate it during tutoring, and when they do, it does not appear to cause significantly larger learning gains.

6. Motivation

The effectiveness of human tutoring perhaps may be due to increasing the motivation of students. Episodes of tutoring that seem intended to increase students' motivation are quite common in human tutoring (Cordova & Lepper, 1996; Lepper & Woolverton, 2002; Lepper, Woolverton, Mumme, & Gurtner, 1993; McArthur et al., 1990), but their effects on student learning are unclear.

For instance, consider praise, which Lepper et al. (1993) identified as a key tutorial tactic for increasing motivation. One might think that a human tutor's praise increases motivation, which increases engagement, which increases learning, whereas a computer's praise might have small or even negative effects on learning. However, the effect of human tutors' praise on tutees is actually quite complex (Henderlong & Lepper, 2002; Kluger & DeNisi, 1996). Praise is even associated with *reduced* learning gains in some cases (Boyer, Phillips, Wallis, Vouk, & Lester, 2008). Currently, it is not clear exactly when human tutors give praise or what the effects on learning are.

As another example, the mere presence of a human tutor is often thought to motivate students to learn less (the so-called "warm body" effect). If so, then text-mediated human tutoring should be more effective than face-to-face human tutoring. Even when the text-mediated tutoring is synchronous (i.e., chat, not e-mail), it seems plausible that it would provide less of a warm-body effect. However, Siler and VanLehn (2009) found that although text-mediated human tutoring took more time, it produced the same learning gains as face-to-face human tutoring. Litman et al. (2006) compared text-mediated human tutoring with spoken human tutoring that was not face-to-face (participants communicated with full-duplex audio and shared screens). One would expect the spoken tutoring to provide more of a warm-body effect than text-mediated tutoring, but the learning gains were not significantly different even though there was a trend in the expected direction.

As yet another example, Lepper et al. (1993) found that some tutors gave positive feedback to incorrect answers. Lepper et al. speculated that although false positive feedback may have actually *harmed* learning, the tutors used it to increase the students' self-efficacy.

In short, even though motivational tactics such as praise, the warm body effect, or false positive feedback are common in human tutoring, they do not seem to have a direct effect on learning as measured in these studies. Thus, motivational tactics do not provide a plausible explanation for the superiority of human tutoring over computer tutoring.

7. Feedback

Another hypothesis is that human tutors help students both monitor their reasoning and repair flaws. As long as the student seems to be making progress, the tutor does not intervene, but as soon as the student gets stuck or makes a mistake, the tutor can help the student resolve the lack of knowledge and get moving again (Merrill et al., 1992). With CAI, students can produce a multiminute-long line of reasoning that leads to an incorrect answer, and then have great difficulty finding the errors in their reasoning and repairing their knowledge. However, human tutors encourage students to explain their reasoning as they go and usually intervene as soon as they hear incorrect reasoning (Merrill et al., 1992). They sometimes intervene even when they heard correct reasoning that is uttered in an uncertain manner (Forbes-Riley & Litman, 2008; Fox, 1993). Because human tutors can give feedback and hints so soon after students make a mental error, identifying the flawed knowledge should be much easier for the students. In short, the frequent feedback of human tutoring makes it much easier for students to find flaws in their reasoning and fix their knowledge. This hypothesis, unlike the first six, seems a viable explanation for why human tutoring is more effective than computer tutoring.

8. Scaffolding

Human tutors scaffold the students' reasoning. Here, "scaffold" is used in its original sense (the term was coined by Wood, Bruner, and Ross, 1976, in their analysis of human tutorial dialogue). The following is a recent definition:

> [Scaffolding is] a kind of *guided prompting* that pushes the student a little further along the same line of thinking, rather than telling the student some new information, giving direct feedback on a student's response, or raising a new question or a new issue that is unrelated to the student's reasoning . . . . The important point to note is that scaffolding involves *cooperative execution* or *coordination* by the tutor and the student (or the adult and child) in a way that allows the student to take an increasingly larger burden in performing the skill. (M. T. H. Chi et al., 2001, p. 490)

For instance, suppose the following dialogue takes place as the student answers this question: "When a golf ball and a feather are dropped at the same time from the same place in a vacuum, which hits the bottom of the vacuum container first? Explain your answer."

1. Student: "They hit at the same time. I saw a video of it. Amazing."
2. Tutor: "Right. Why's it happen?"
3. Student: "No idea. Does it have to do with freefall? "
4. Tutor: "Yes. What it the acceleration of an object in freefall?"
5. Student: "g, which is 9.8 m/sˆ2."
6. Tutor: "Right. Do all objects freefall with acceleration g, including the golf ball and the feather?"
7. Student: "Oh. So they start together, accelerate together and thus have to land together."

Tutor Turns 2, 4, and 6 are all cases of scaffolding, because they extend the student's reasoning. More examples of scaffolding appear later.

Scaffolding is common in human tutoring (Cade et al., 2008; M. T. H. Chi et al., 2001; Cho et al., 2000; Core et al., 2003; Evens & Michael, 2006; Fox, 1991, 1993; Frederiksen et al., 2000; Graesser et al., 1995; Hume et al., 1996; Katz et al., 2003; McArthur et al., 1990; Merrill et al., 1995; Merrill et al., 1992; Ohlsson et al., 2007; VanLehn, 1999; VanLehn et al., 2003; Wood et al., 1976). Moreover, experiments manipulating its usage suggest that it is an effective instructional method (e.g., M. T. H. Chi et al., 2001). Thus, scaffolding is a plausible explanation for the efficacy of human tutoring.

9. The ICAP Framework

M. T. H. Chi's (2009) framework, now called ICAP (M. T. H. Chi, 2011), classifies observable student behaviors as interactive, constructive, active, or passive and predicts that that they will be ordered by effectiveness as

$$\text{interactive} \geq \text{constructive} > \text{active} > \text{passive}.$$

A *passive* student behavior would be attending to the presented instructional information without additional physical activity. Reading a text or orienting to a lecture would be passive student behaviors. An *active* student behavior would include "doing something physically" (M. T. H. Chi, 2009, Table 1), such as taking verbatim notes on a lecture, underlining a text, or copying a solution. A *constructive* student behavior requires "producing outputs that contain ideas that go beyond the presented information" (M. T. H. Chi, 2009, Table 1), such as self-explaining a text or drawing a concept map. An interactive behavior requires "dialoguing extensively on the same topic, and not ignoring a partner's contributions" (M. T. H. Chi, 2009, Table 1).

The ICAP framework is intended to apply to observed *student* behavior, not to the instruction that elicits it. Human tutors sometimes elicit interactive student behavior, but not always. For instance, when a human tutor lectures, the student's behavior is passive. When the tutor asks a question and the student's answer is a pure guess, then the student's behavior is active. When the tutor watches silently as the student draws a concept map or solves a problem, then the students' behavior is constructive. Similarly, all four types of student behavior can in principle occur with computer tutors as well as human tutors.

However, M. T. H. Chi's definition of "interactive" includes co-construction and other collaborative spoken activity that are currently beyond the state of the art in computer tutoring. This does not automatically imply that computer tutoring is less effective than human tutoring. Note the $\geq$ in the ICAP framework: Sometimes constructive student behavior is just as effective as interactive student behavior. In principle, a computer tutor that elicits 100% constructive student behavior could be just as effective as a human tutor that elicits 100% interactive student behavior.

Applying the ICAP framework to tutoring (or any other instruction) requires knowing the relative frequency of interactive, constructive, active, and passive student behaviors. This requires coding observed student behaviors. For instance, in a study where training time was held constant, if one observed that student behaviors were 30% interactive, 40% constructive, 20% active, and 10% passive for human tutoring versus 0% interactive, 70% constructive, 20% active, and 10% passive for computer tutoring, then ICAP would predict equal effectiveness. On the other hand, if the percentages were 50% interactive, 10% constructive, 20% active, and 20% passive for human tutoring versus 0% interactive, 33% constructive, 33% active, and 33% passive for computer tutoring, then ICAP would predict that the human tutoring would be more effective than the computer tutoring. Unfortunately, such coding has not yet been done for either human or computer tutoring.

Whereas the first eight hypotheses address the impact of tutors' behaviors on learning, this hypothesis (the ICAP framework) proposes an intervening variable. That is, tutors' behaviors modify the frequencies of students' behaviors, and students' behaviors affect students' learning. Thus, even if we knew that interactive and constructive behaviors were more common with human tutors than with computer tutors, we would still need to study tutors' behaviors in order to figure out *why.* For instance, prior to discovery of the ICAP framework, M. T. H. Chi et al. (2001) found that scaffolding was associated with certain student behaviors that would now be classified as constructive and interactive. Scaffolding is a tutor behavior referenced by Hypothesis 8. In principle, the ICAP framework could be combined with any of the eight hypotheses to provide a deeper explanation of the difference between human and computer tutoring, but the ICAP framework is not an alternative to those hypotheses.

Having listed several hypotheses, it may be worth a moment to clarify their differences. Feedback (Hypothesis 7) and scaffolding (Hypothesis 8) are distinct hypotheses but strongly related. Feedback occurs *after* a student has made a mistake or reached an impasse, whereas scaffolding is *proactive* and encourages students to extend a line of reasoning. Both are effective because they allow students to experience a correct line of reasoning wherein they do much of the reasoning themselves and the tutor assists them with the rest. Of the first eight hypotheses listed earlier, these two seem most viable as explanations for the effectiveness of human tutoring perhaps because they are simply the proactive and reactive versions of the same assistance.

Although feedback and scaffolding are forms of adaptivity and individualization, they should not be confused with the individualized task selection of Hypothesis 2. In all three cases, the tutors' decision about what activity to do next is based on the students' behavior, so tutors are adapting their behavior to the students'. However, the durations of the activities are different. In the case of individualized task selection, the tutor decides upon a relatively long activity, such as having the student solve a selected problem, read a few pages of text, study a video, and so on. In the case of scaffolding and feedback, the tutor decides whether to remain silent, to give feedback, to give a hint, to do the next step for the student, and so forth. Shute (1993) captured the distinction using the terms "macro-adaptive" for computer tutors that have sophisticated methods for individualizing task selection and "micro-adaptive" for tutors that have sophisticated methods for regulating scaffolding, feedback, and other short-duration instructional activities.

To summarize, of the hypotheses discussed here, Hypothesis 9 (the ICAP framework) is complementary with all the others and can potential provide an explanation for how the tutors' behaviors affect the students' learning. Of the other eight hypotheses—detailed diagnostic assessment, individualized task selection, sophisticated tutorial strategies, learner control of dialogues, domain knowledge, motivation, feedback, and scaffolding—only the last two seem completely free of contravening evidence. Hypothesis 6 (motivation) may explain the efficacy of human tutoring, but such an explanation would be complicated. Thus, the best hypotheses so far are that when human tutors cause larger learning gains than computer tutors, it is because they are better at scaffolding students and giving feedback that encourages students to engage in interactive and constructive behaviors as they self-repair and construct their knowledge. Similar conclusions were reached in earlier reviews (M. T. H. Chi et al., 2001; Graesser et al., 1995; Merrill et al., 1992).

The Interaction Granularity Hypothesis

The preceding sections argued that human tutoring was more effective than computer tutoring because human tutors used feedback and scaffolding more effectively than computer tutors. This section examines how different kinds of tutors give feedback and scaffolding. It argues that the key difference is the granularity of their interaction.

Granularity is a property of the user interface. Although human tutoring has no constraints on granularity of interaction, different types of computer tutors have different granularities. Granularity refers to the amount of reasoning required of participants between opportunities to interact. The larger the grain size of the user interface, the more reasoning per interaction. For instance, after an answer-based tutoring system poses a problem, the next opportunity for interaction is when the student enters an answer. Thus, the user interface is designed so that the student does all the reasoning required for solving the problem between interactions. In contrast, an ITS gives the students many opportunities to interact with the tutor while solving the problem. For instance, students may fill out a form that prompts for intermediate steps, or they may participate in a dialogue with the tutoring system. Because the answer-based tutoring system requires more reasoning between interactions, it has a larger grain size of interaction than the step-based tutoring system.

To explain some of the nuances of granularity, it is best to work with an example. Consider the following problem:

> In a basket of apples, 40% of the apples have worms, and 25% are rotted inside. Assuming independence, what is the chance that a randomly selected apple will have a worm but no rot?

Let us first consider a hypothetical typed dialogue between a human tutor and a student.

1. Student: I'm stuck
2. Tutor: Our event is an apple with a worm and no rot, right?
3. Student: Right.
4. Tutor: Can that event be decomposed into an AND or an OR or something of two events?
5. Student: I don't see how.
6. Tutor: Well, if you think of having a worm as one kind of event, and having rot as another event, then our event is both having a worm AND having no rot.
7. Student: OK
8. Tutor: Can you write that in symbols, like P(. . .)?
9. Student: P(worm & rot).
10. Tutor: Almost. Check your "rot"
11. Student: P(worm & ∼rot)
12. Tutor: Good. Do you know a rule that matches that?
13. Student: P(A&B) = P(A) * P(B)
14. Tutor: Excellent! Now are you sure that it is legal to apply this rule? What do you have to check?
15. Student: A is worm and B is ∼rot.
16. Tutor: As we've discussed before, you must always remember which rules require independence. This is

one of them. To apply P(A&B) = P(A)*P(B), events A and B must be independent.

On Turns 10 and 16, the tutor gives negative feedback. On Turns 6 and 16, the tutor gives an explanation. On all the other turns, the tutor scaffolds the student by prompting the next little bit of reasoning that needs to be done. Notice how little reasoning is covered by each interaction. This particular dialogue illustrates fine-grained interaction. Human tutoring doesn't have to be like this. In fact, the tutor or the student may give an explanation that covers a whole solution. Such a long explanation would compose a large-grained interaction. However, the "user interface" of human tutoring puts no constraints on the granularity of the interaction.

Now consider a typical answer-based tutoring system. It would have a box where students can type in an answer. If 0.3 is entered, the student gets positive feedback. If 0.1 is entered, the tutor might say, "Almost. That's P(worm & rot). You want P(worm & ∼rot)." Entering any other number might cause the tutor to say, "Try again." Note that the tutor only recognizes attempts at a whole solution, which is a long line of reasoning. A typical tutor would also have two buttons. The Hint button gives some scaffolding, just like the human tutor's scaffolding, but intended to power the student's reasoning all the way from start to finish. The Explain button presents a complete solution. Again, both the hint and the explanation refer to the whole line of reasoning from problem to answer. In short, the students' reasoning, the tutor's explanations, and all the other interactions permitted by this user interface refer to a long line of reasoning, so this user interface has a large granularity of interaction.

Turning now to ITS, we need to distinguish two types. The first type of ITS will be called a *step-based* tutoring system because it is engineered to let users enter the steps that they would do when solving problems normally, without the tutoring (VanLehn, 2006). For instance, if the tutor is teaching a procedural skill such as programming a video recording system (Mark & Greer, 1995), then the user interface simulates the front panel of the device, and students use the mouse to click in times, select TV channels, and so on. On the other hand, if students normally work on paper, then the user interface may provide an electronic form for them to enter what they should write if they are showing all their work. For instance, Figure 2 shows a form appropriate for the rotten apples example, partially filled out by the student (in italics). Suppose the student gets stuck at this point and presses the Hint button. The tutor might say, "You seem to be seeking P(∼rot) but the problem provides only P(rot) = 0.75. Can you think a rule that relates them?" If the student presses the Hint button enough times, the tutor gives an explanation for the reasoning required to reach the next step, for example, "You need to apply the rule P(A) = 1-P(∼A), which in this case is P(∼rot) = 1 – P(rot) = 1 – 0.25 = 0.75" The point is that the students' entries, the tutor's hints, and the tutor's explanations all refer to a relatively short line of reasoning.

Goal	Rule	Result
P(worm & ~rot)	*P(A&B) = PA) * P(B)*	
P(worm)	*Given*	*0.40*
P(~rot)		

FIGURE 2 A form used by a hypothetical step-based tutoring system for probability.

A step-based tutoring system has a much smaller grain size than the answer-based tutoring system.

The second type of ITS will be called a *substep-based* tutor because it can give scaffolding and feedback at a level of detail that is even finer than the steps students would normally enter when solving a problem. For instance, suppose a student is programming a video recorder to start recording at 2 p.m., and the student presses the Hint button. A substep-based tutor might start by asking, "Are you familiar with military time, e.g., 1600?" If the student says yes, then the tutor would say, "This recorder uses military time, so enter 1400." If the student says no, then the tutor might provide instruction on military time using a multientry form such as this:

> To convert a *morning* time to military time, just append the hours to the minutes, so 9:30 am becomes 0930. To convert an *afternoon* time to military time, add 12 to the hours and append the minutes, so 3:30 pm becomes 1530. For this TV show, you want to enter a 2:00 pm start time. This is an _____ time, so you should add ___ to ___ hours then append ___ minutes, so you finally get _____, which is what you should enter into the recorder.

The student fills in the blanks using typing and/or menus. These entries are not *overt* when normally operating the recorder, but they probably correspond to mental inferences. Thus, they are substeps and not steps.

This example of a substep-based tutoring system uses a conventional user interface, often called a graphical user interface (GUI) or WIMP (Windows, icon, menu, and pointing) user interface. Other major types of user interface are command line, touch screen, and natural language dialogue user interfaces. In principle, the type of user interface is independent of the granularity of the user interface. However, all substep-based tutoring systems developed so far have used a natural language dialogue user interface. Indeed, the rotten apple dialogue presented earlier could be conducted by a substep-based tutoring system, and it illustrates the use of substeps. For instance, the tutor asked for the reasoning behind a step in Turn 14. Although identifying independence assumptions is critically important in probability problem solving, it is not usually done explicitly when students are just writing steps. Thus, Turn 14 illustrates what a substep looks like in the context of a natural language dialogue user interface.

Just to drive home the point that the user interface type (natural language vs. GUI vs. command line, etc.) is independent of the granularity (answer vs. step vs. substep), it

is worth mentioning that step-based tutoring systems sometimes have natural language user interfaces. A good example is the COVE tutoring system (Roberts, 2001), which is an ITS where students enter steps by uttering natural language commands (e.g., "All engines ahead full."). COVE teaches conning officers how to steer a naval vessel, so it is important that the students learn to speak the proper commands. Its steps are natural language commands, because that is what the student normally does when driving a ship. However, COVE does not conduct a dialogue with the trainees about their reasoning. For instance, it does not ask them why they decided to issue a specific command. Thus, despite its speech input, COVE is a step-based tutor and not a substep-based tutor. Although there are other speech-input *step*-based tutoring systems (e.g., Core et al., 2006; F. Ritter & Feurzeig, 1988; Stottler & Panichas, 2006; Zachary et al., 1999), there are also speech-input *substep*-based tutoring systems as well (D'Mello, King, Stolarski, Chipman, & Graesser, 2007; Litman et al., 2006; Pon-Barry, Schultz, Bratt, Clark, & Peters, 2006). The point is that the user interface type is independent of the user interface granularity.

This review classifies tutoring systems by their granularity of interaction and ignores their user interface type (natural language vs. GUI, etc.). Five grain sizes are used, and they are ranked by size as follows:

human tutoring < substep-based tutoring < step-based tutoring < answer-based tutoring

Human tutoring is finer grained than substep-based tutoring because its granularity is unconstrained. Substep-based tutoring systems have dialogues that are preplanned and sometime preauthored, so there is always a limit on their granularity.

As argued earlier, human tutoring is thought to be more effective than computer tutoring because human tutors are better at giving feedback and scaffolding. As the preceding discussion illustrates, computer tutors also give feedback and scaffold; they just do so at a larger grain size than human tutoring. This suggests that a key property that makes one type of tutoring more effective than other types of tutoring may be the granularity of the interaction—how much reasoning the user interface expects to occur between student–tutor interactions.

The *interaction granularity hypothesis* is that the effectiveness of tutoring systems (including human tutoring as a "system") increases as the granularity of interaction of the system decreases. This is presumably due to the participants interacting around smaller and smaller pieces of reasoning and pieces of knowledge, which makes the scaffolding, feedback, and perhaps even explanations more effective. Although an earlier statement of this hypothesis referred to the independent variable as "interactivity" (VanLehn et al., 2007), interaction granularity is a more accurate name for it. The interaction granularity hypothesis predicts that the effectiveness of different types of tutoring should be

human tutoring > substep-based tutoring > step-based tutoring > answer-based tutoring

The hypothesis could be extended to predict that two tutoring systems that have the same interaction granularity but are otherwise identical will produce the same learning gains. In particular, a GUI user interface should produce the same learning gains as a dialogue-based user interface if they have the same granularity. For instance, COVE should be no more effective than a version of it that used menus instead of spoken language as input. Of course, the menu-based version wouldn't give students practice in recalling and uttering the proper naval commands, so the comparison would be fair only if it were restricted to knowing which command to give at what time, even if the command were uttered improperly. This prediction has not been tested yet with COVE, but it is consistent with several experiments in mathematics (Aleven, Ogan, Popescu, Torrey, & Koedinger, 2004; Corbett, Wagner, Lesgold, Ulrich, & Stevens, 2006). These experiments showed that when a step-based tutor used a natural-language interface, it improved students' mastery of mathematical language more than the same tutor using a menu-based interface, but both methods of entering steps were equally good at improving students' problem solving skills and mathematical understanding, as predicted by the interaction granularity hypothesis.

Scope and Method of the Review

The first subsection explains the criteria for inclusion in the review. A second subsection describes the review's methods.

Inclusion and exclusion criteria. To test the hypothesis that types of tutoring are ranked by effectiveness as

human > substep-based > step-based > answer-based > no tutoring,

only experiments that involved at least two different types of tutoring were included. For instance, an experiment could compare human tutoring to no tutoring, or substep-based tutoring to answer-based tutoring.

There already exist at least a dozen reviews comparing answer-based tutoring to no tutoring (J. A. Kulik, 1994). Thus, if a study included only answer-based tutoring and no-tutoring, it was excluded.

This review covers only experiments that manipulated interaction granularity, as previously defined, while trying to control for all other variables. Most important, comparisons were excluded if the control and experimental conditions received different content. For instance, studies of Carnegie Learning's Cognitive Tutors were excluded because students in the experimental condition used a different textbook and

classroom activities than students in the comparison conditions. Given the prominence of these step-based tutoring systems, it is worth mentioning that What Works Clearinghouse (WWC) recognized that three efficacy studies of the Carnegie Learning Algebra I treatment met WWC standards of evidence, and two studies met the standards with reservations (WWC, 2009, 2010). The results of these studies were mixed. Using measures selected by WWC, the effect sizes from these five studies were 0.38 (S. Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007), –0.18 (Cabalo, Jaciw, & Vu, 2007), –0.16 (Campuzano, Dynarski, Agodini, & Rall, 2009), 0.04 (Shneyderman, 2001), and –0.07 (J. E. Smith, 2001). Carnegie Learning's (2010) review of these studies found moderately large positive effects in measures not selected by WWC. The company's review also reports moderately large positive effects in several efficacy studies not included by WWC. However, none of these studies controlled for content. The negative results reported by the WWC would be explained if the standardized exams focused on the traditional content taught in the control classrooms and the Cognitive Tutors focused on other content.

Topic has been a strong moderating variable in earlier meta-analyses of tutoring. For instance, effect sizes were smaller for reading than for science, technology, engineering or mathematics (STEM) topics (Cohen et al., 1982; Fletcher-Flinn & Gravatt, 1995; G. W. Ritter, et al., 2009). To reduce the variance across studies and thus allow a pattern to emerge, the subject matter being tutored was constrained to be a STEM topic. In particular, this review excludes studies of tutoring of reading, second language, music, and sports.

Moreover, all the studies included here used tasks that have distinguishable right and wrong solutions. If there were studies using ill-defined tasks (e.g., design tasks where the outcome variable is novelty or creativity), then their findings might differ from those found in this article.

Only studies of one-on-one tutoring were included. This excludes studies in which two or more students interacted simultaneously with the same human tutor or tutoring system. In one borderline case (Reif & Scott, 1999), two human tutors circulated among six students doing their physics homework in a conference room. Students were not interacting with each other, only with the tutors. The study was included in this review because its method more closely matched one-on-one human tutoring than the one-to-many treatments where a human tutor participates in a discussion among a small group of students.

For human tutoring, the tutors were required to be adult, subject-matter experts. Studies of peer tutoring or cross-age tutoring were excluded. No restriction was placed on the expertise of the tutor in tutoring itself because there are no established measures of such expertise apart from the learning gains of the tutees. Although it is feasible to include only experienced tutors, the pattern of the results might not be any different, because there is little evidence that experienced tutors are significantly more effective than inexperienced tutors, *given that they are all subject-matter experts* (Chae, Kim, & Glass, 2005; di Eugenio, Kershaw, Lu, Corrigan-Halpern, & Ohlsson, 2006; Fossati, 2008). The Cohen et al. (1982) meta-analysis found no relationship between tutor's experience and their effectiveness. Clark, Snow, and Shavelson (1976) found that giving subject-matter experts training and experience as tutors did not make them more effective. Although the expertise of tutors was not used as a criterion for exclusion of studies, the discussion section reopens this issue.

Only experiments with random assignment to conditions were included. Studies with random assignment of *classes* to condition were included, even though this is a weaker design than random assignment of students to condition.

Studies were included only if they reported measures that allowed an effect size to be calculated. Besides the mean and total sample size for each condition, either standard deviations, standard errors, mean standard error, the t statistic, or the F statistic were required.

To match common beliefs about maximally effective forms of human tutoring, studies of human tutoring were restricted to synchronous human tutoring. Synchronous means that there are no long delays between turns, as there are with e-mail or forums. Face-to-face tutoring qualifies as synchronous, and most studies of human tutoring in the review used face-to-face tutoring. Some studies used other methods of synchronous human tutoring. Typically, the student and the human tutor in these studies were in separate rooms or were separated by a partition. They could both see a common workspace, but they could not see each other. They communicated via full-duplex audio (e.g., speaking loudly enough to be heard over the partition) or a text "chat" interface. Most of these studies were comparing human tutoring to a substep-based tutoring, so they arranged for the human tutors to use exactly the same communication medium as the computer tutor (both spoken or both typed). These studies were included in the review because existing evidence suggests that text-based, synchronous human tutoring is slower but no less effective than face-to-face human tutoring (Siler & VanLehn, 2009) or audio-mediated human tutoring (Litman et al., 2006).

The review excludes studies of tutoring with specialized student populations, such as deaf, blind, or learning disabled students. Although there a gray area between "learning disabled" and "below grade level" but not disabled, none of the candidate studies for this review fell into that gray area.

Last, only studies published between 1975 and 2010 were included.

METHODS

This review attempts to cover all studies that fit the criteria listed in the previous sections. Two searches were

conducted—an informal search and a formal search. The informal search started by considering four well-known venues for work on intelligent tutoring systems (the *International Journal of AI in Education,* the journal *Interactive Learning Environments*, and the proceedings of two conferences, *Intelligent Tutoring Systems* and *AI in Education*). The inclusion/exclusion criteria were applied to abstracts in all volumes of these four venues. The selected articles were read, and their reference lists were searched for more relevant articles. This recursive search through reference lists continued until no new articles were found. This search yielded 87 comparisons.

The formal search started by querying ERIC using this criterion: "tutor*" AND "experiment" NOT "peer" NOT "crossage" NOT "reciprocal" NOT "reading" NOT "L2" NOT "spelling" NOT "vocabulary" NOT "autism". This yielded 1,226 citations. Applying the inclusion/exclusion criteria to the abstracts filtered out all but 44 publications. The 44 publications included two publications not found during the informal search, and they contained eight new comparisons. Next, PsycINFO was searched with a similar criterion and yielded 197 citations. Applying the inclusion/exclusion criterion to the abstracts filtered out all but 1 publication, which yielded no new comparisons. At this point, the formal search was truncated because it had failed to uncover most of the publications known to be relevant before the search was conducted and had located few new relevant ones.

Because the formal search was truncated and the informal search is not replicable, and only one coder (the author) was used in applying the inclusion/exclusion criteria, this review does not meet current standards for meta-analyses (Cooper, Hedges, & Valentine, 2009). Nonetheless, it is likely that nearly all the relevant studies have been located and correctly classified, in part because the author has worked for more than three decades in the tutoring research community.

Many experiments use multiple assessments of learning. For instance, an experiment might include both a conceptual and a quantitative posttest. This review reports the assessment with the largest effect size.

When an experiment has three or more conditions, each pairwise comparison produces its own effect size. For completeness, all of them are reported here. For instance, suppose an experiment has three groups: human tutoring, step-based tutoring, and no-tutoring. There are three pairwise comparisons, and each is reported in a different table in the Appendix: The human tutoring versus step-based tutoring is reported in one table, the human tutoring versus no-tutoring is reported in a second table, and the step-based tutoring versus no-tutoring is reported in a third table.

As mentioned earlier, the review compares four types of tutoring (one human, and three computer) and a no-tutoring treatment that teaches the same content as the tutoring condition. Different experiments used different no-tutoring treatments. Most were combinations of reading text and solving problems without feedback. It would be difficult to develop a moderator variable or subclassification for the no-tutoring treatments, as they are often not completely described in the original sources. Nonetheless, in the Appendix that lists the individual studies, the no-tutoring controls are briefly described.

Three Illustrative Studies

The goal of this review is to test the interaction granularity hypothesis by comparing the effectiveness of five types of instruction: human tutoring, substep-based tutoring, step-based tutoring, answer-based tutoring, and no-tutoring. To clarify the five types and the hypothesis, this section reviews several studies that each include at least three types of tutoring.

Illustration 1. Evens and Michael (2006) reported and integrated all of their studies on tutoring of medical students who were taking a class on cardiovascular physiology. They studied different methods for teaching students an operational mental model of the baroreceptor reflex, which controls human blood pressure. All students were first taught the basics of the baroreceptor reflex and then were given a training problem wherein an artificial pacemaker malfunctions and students must fill out a table describing the body's response. The table's rows denoted physiological variables (e.g., heart rate; the blood volume per stroke of the heart, etc.). The table's columns denoted time periods. The students filled each cell in the table with a +, −, or 0, thus indicating whether the row's variable was increasing, decreasing, or constant during the column's time period. Filling in a cell counts as a step, so this was a multistep task. When students had filled in a whole column, which they could do in any order they wished, they pressed a Submit button and tutoring on that column began. When all the cells in that column had correct entries, student could start on the next column.

Evens, Michael, and their colleagues first developed a step-based tutoring system, CIRCSIM. When tutoring the student on a column, it indicated each incorrect step (i.e., a cell that was filled in incorrectly) and presented a short textual explanation of the proper reasoning for that step. Because the student had no opportunity to interact with the tutor as they read the explanation, the user interface granularity was a step.

Next, Evens, Michael, and their colleagues developed a substep-based tutoring system, CIRCSIM-tutor. Instead of the tutor printing a paragraph-long explanation for each incorrect step, CIRCSIM-tutor conducted a sophisticated, typed natural language dialogue. Although the default dialogue had exactly the same content as the printed explanations of the step-based tutor, it went more deeply into a topic if the student's response indicated missing knowledge or a misconception.

The CIRCSIM team conducted several experiments. Some used a no-tutoring condition wherein students studied a text that included examples of the correct reasoning for solving the pacemaker problem. Some experiments included

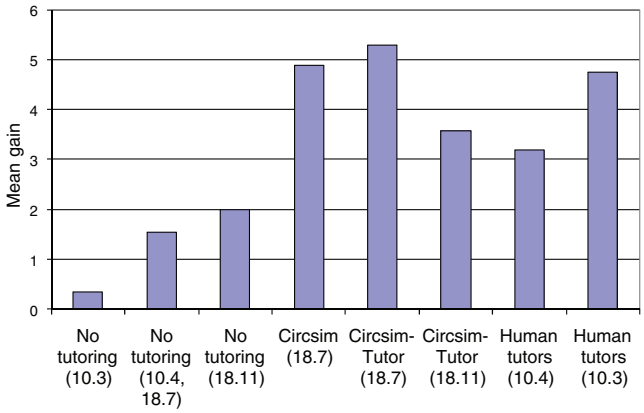

FIGURE 3 Results from Evens & Michael (2006). *Note.* Condition names are followed by the table number in Evens & Michael, as the experiments were not numbered (color figure available online).

human tutors interacting with students via typing. Although inexperienced tutors were used in one such experiment, the others used two highly experienced tutors who had studied transcripts of their own tutoring in order to improve it. If the interaction granularity hypothesis holds, then the treatments should be ordered as follows: human tutoring > CIRCSIM-tutor (a substep-based tutor) > CIRCSIM (a step-based tutor) > reading text (no tutoring).

Figure 3 summarizes the results from experiments that used the same assessments and training problems but different types of instruction. The three most interactive types of tutoring—human tutors, CIRCSIM-tutor, and CIRCSIM—tied with each other. The no-tutoring (reading) conditions differed from the CIRCSIM, CIRSIM-tutor, and human tutor conditions, but these tutoring conditions did not differ statistically from each other. This pattern is inconsistent with the interaction granularity hypothesis.

Illustration 2. VanLehn et al. (2007) studied conceptual physics tutoring. Conceptual physics is an open task domain in that it is often taught "by discussion" instead of by solving pencil-and-paper multistep tasks. For instance, students discuss "why" questions such as

> As the earth orbits the sun, the sun exerts a gravitational force on it. Does the earth also exert a force on the sun? Why or why not?

In this experiment, such question-answering tasks were treated as multistep tasks. That is, a correct and complete answer to such a question was decomposed into several key ideas. Mentioning such an idea was treated as a step. For instance, two of the key ideas to be mentioned in response to the question just presented were "Newton's third law" and "action-reaction force pair." Because the steps were in natural language, students were not expected to use exactly those words, but a complete and correct explanation should contain those two steps along with several others.

In all conditions of the experiment, students first studied a short textbook. They then worked on several tasks that each involved answering a "why" question, such as the one just listed. For each task, the students wrote a short essay as their initial answer; were tutored on missing or incorrect steps; and then read a correct, well-written essay.

The treatments differed in how students were tutored when the essay lacked a step or had an incorrect or incomplete step. There were five treatments. They are listed next in order of interaction granularity, with the finest granularity treatment first:

1. Human tutors who communicated via a text-based interface with student; with a few exceptions, the tutor was a professional physics tutor who studied transcripts of his tutoring in order to improve it.
2. Why2-Atlas, a substep-based tutoring system.
3. Why2-AutoTutor, another substep-based tutoring system.
4. A simple, step-based tutor that presented the same content as Why2-Atlas, but as text instead of dialogue.
5. A control condition that had students merely read passages from a textbook without answering conceptual questions.

If the interaction granularity hypothesis holds, then one would expect the treatments to be ordered by efficacy as $1 > 2 = 3 > 4 > 5$.

Figure 4 shows the posttest scores, adjusted for pretest scores in an analysis of covariance. The four most fine-grained types of tutoring (numbers 1–4 in the list just presented) were not reliably different from each other. All four were higher than the read-only no-tutoring condition (number 5 in the list) by approximately $d = 1.0$. Thus, the results of Experiments 1 and 2 of VanLehn et al. (2007) do not support the interaction granularity hypothesis.

Surprised by the failure of the interaction granularity hypothesis, VanLehn et al. (2007) conducted five more experiments. The experiments used different assessment methods (e.g., far transfer; retention), students with different prior knowledge (with or without a college physics course), and different numbers of training problems. With one exception, the same pattern of results was observed in every experiment. The exception was an experiment in which students who had not taken college physics were trained with materials that were designed for students who had taken college physics. In that experiment, human tutoring was more effective than the step-based tutor. Upon reviewing transcripts of the tutoring, the researchers concluded that the materials were too far above the students' current level of competence. Although reading the step-based tutor's remedial text probably didn't suffice for comprehension, the human tutor was able to help explain the content in novice terms. In partic-

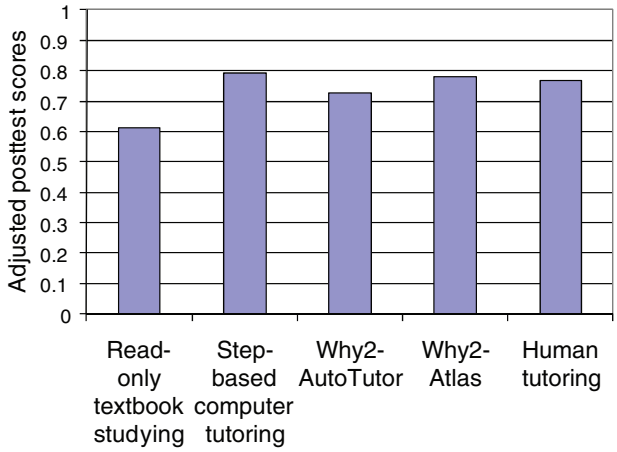

FIGURE 4 Results from VanLehn et al. (2007), Experiments 1 and 2 (color figure available online).

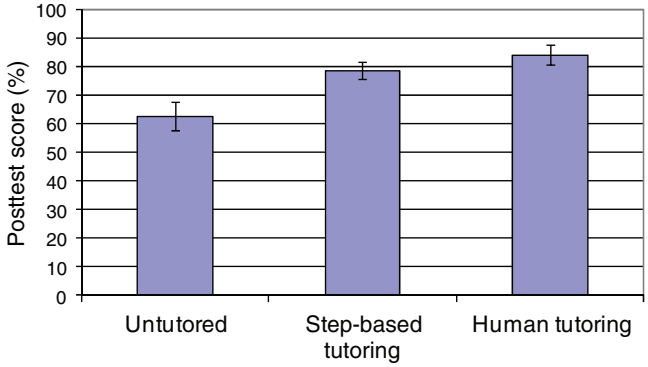

FIGURE 5 Results from Reif and Scott (1999) (color figure available online).

ular, the text assumed that readers knew about vectors, but the novices clearly did not. This interpretation was consistent with later experiments that used completely overhauled materials designed especially for students who had not taken college physics and yielded the pattern shown in Figure 4.[1]

Illustration 3. Reif and Scott (1999) compared human tutoring, step-based tutoring, and no tutoring. All students in their experiment were in the same physics class; the experiment varied only the way in which the students did their homework. There were three treatments. (1) One group of students did their physics homework problems with the aid of a human tutor. (2) Another group of students merely did their homework at home as usual. Because they received no feedback until their papers were returned, this treatment is classified as no-tutoring. (3) The remaining students did their homework on a step-based tutoring system. The system had them either solve a problem or study a solved problem (example). When solving a problem, students got immediate feedback and hints on each step. When studying an example, they were shown equations, vectors, and other inscriptions. As each was displayed, students were asked to judge whether it was correct. These judgments counted as steps. Students were given immediate feedback on their steps and could ask for hints as well. Thus, both the problem-solving and example-studying activities count as step-based tutoring.

Figure 5 shows the Reif and Scott results. The human tutors and the step-based tutor produced learning gains that were not reliably different, and yet both were reliably larger than the gains of the no-tutoring group ($d = 1.31$ for human tutoring; $d = 1.01$ for step-based tutoring).

However, the human tutors and the step-based tutor taught an effective problem-solving method (Heller & Reif, 1984) that was not taught to the no-tutoring students (F. Reif, personal communication, October 2009). Thus, part of the large benefit of tutoring over no-tutoring may be due to a difference in instructional content rather than a difference in instructional method. This review is intended to focus only on differences in method, so from this experiment, the review includes only the comparison between human tutoring and step-based tutoring. The no-tutoring condition is ignored as it appears to have used different content than the other two.

RESULTS

Unlike most meta-analyses, which compare just two types of instruction, this review compares five types of tutoring, so there are 10 possible pairwise comparisons. Table 1 shows them all. Each row presents one pairwise comparison. For

TABLE 1
Summary of Effects

Comparison	*No.*	ES	*% Reliable*	*Table*
Answer based vs. no tutoring	165	0.31	40%	From Kulik & Kulik (1991, Table 3)
Step based vs. no tutoring	28	0.76	68%	A1
Substep based vs. no tutoring	26	0.40	54%	A2
Human vs.no tutoring	10	0.79	80%	A3
Step based vs.answer based	2	0.40	50%	A4
Substep based vs. answer based	6	0.32	33%	A5
Human vs.answer based	1	−0.04	0%	A6
Substep based vs. step based	11	0.16	0%	A7
Human vs.step based	10	0.21	30%	A8
Human vs.substep based	5	−0.12	0%	A9

[1]For the read-only studying condition (number 5 in the list), these last two experiments used an experimenter-written text instead of a commercial textbook. Although the interaction granularity hypothesis predicts that no-tutoring instruction should be less effective than tutoring, students in this no-tutoring condition had equally high learning gains as students in the other conditions. It is not clear why they learned so well, but it may be due to a lack of fatigue because they finished their reading much more quickly than the other students, who also wrote essays, interacted with a tutor, and so on.

instance, human tutoring can be compared to each of the other four types of tutoring, so human tutoring appears in four rows (rows 4, 7, 9, and 10). Step-based tutoring can also be compared to the other four types of tutoring, so it appears on four rows (rows 2, 5, 6, and 9). In this manner, every type of tutoring appears in four rows.

Within a row, the number of reviewed comparisons appears first, followed by the mean of the effect sizes of those comparisons. The third number is the proportion of the comparisons that were reliable ($p < .05$). The last column of a row identifies the table in the Appendix listing all the comparisons referenced by the row.

The first row of Table 1, labeled Answer-based vs. no-tutoring, differs from the rest in that it reports a result from the C. Kulik and Kulik (1991) review. The review divided computer-based instruction into several categories, one of which is called answer-based tutoring here. Only that category's mean effect size is reported in the table. The proportion of reliable comparisons is an estimate. Although Kulik and Kulik reported that 40% of their comparisons were reliable, they did not break this down per category. However, studies of answer-based tutoring dominated the review (165 of 248 studies were classified as answer-based tutoring), so 40% is used as an approximation in Table 1's third column.

To interpret the effect sizes in Table 1, let us first find a meaningful way to display them graphically. Figure 6a is a conventional display, which shows the effect sizes of Table 1 where each curve has a left end point at zero that represents the effect size of the treatment compared to itself. Figure 6a makes it hard to integrate the results and determine *the* effect size of each type of tutoring. For instance, suppose we are interested in the efficacy of step-based tutoring in comparison to tutoring with a lower granularity. There are

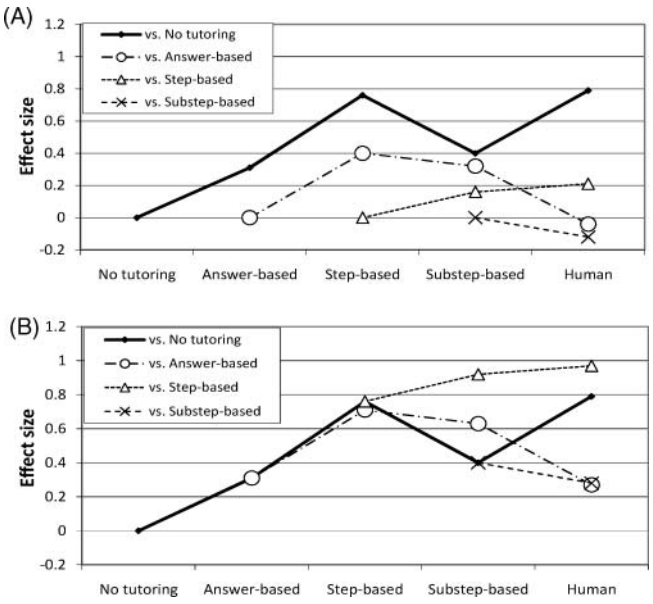

FIGURE 6 Mean effect sizes.

two relevant points: step-based tutoring versus no tutoring (0.76) and step-based tutoring versus answer-based tutoring (0.40). Although one could ignore the latter point and just take 0.76 as the effect size, that choice would ignore some evidence. One could alternatively "raise the baseline" of the latter point using this approximation:

$$\text{step-based vs. no} = \text{step-based vs. answer-based} + \text{answer-based vs. no} = 0.41 + 0.31 = 0.71$$

Using this method, the difference between step-based and no tutoring is now 0.71, which is not far from the 0.76 found when step-based tutors are compared directly to no tutoring. This "raising the baseline" method is applied to all such points in Figure 6b by raising the left end points of four curves so that they lie on the "vs. no tutoring" curve. Thus, all the points in Figure 6b are relative to a no-tutoring baseline, given the crude approximation afforded by adding effect sizes. For example, Figure 6b has four points that all estimate the effect size of human tutoring relative to the no-tutoring baseline:

- human vs. no = 0.79
- "human vs. no" = human vs. answer-based + answer-based vs. no = –0.04 + 0.31 = 0.27
- "human vs. no" = human vs. step-based + step-based vs. no = 0.21 + 0.76 = 0.97
- "human vs. no" = human vs. substep-based + substep-based vs. no = –0.12 + 0.41 = 0.29

Figure 6b suggests that the interaction granularity hypothesis is only half correct. When the granularity decreases from answer based to step based, the effectiveness increases from 0.31 to around 0.75. However, further decreases in granularity yield negligible increases in effect size. That is, there seems to be an interaction plateau—as the granularity of interaction decreases, effect sizes increase, then plateau. Based on a preliminary review, a conference talk (VanLehn, 2008b) introduced the *interaction plateau hypothesis*:

$$\text{human tutoring} = \text{substep-based tutoring} = \text{step-based tutoring} > \text{answer-based tutoring}$$

This hypothesis is approximately consistent with the effect sizes of Table 1 and Figure 6. The interaction plateau hypothesis is also consistent with the illustrative studies described earlier (Evens & Michael, 2006; Reif & Scott, 1999; VanLehn et al., 2007). Although their no-tutoring groups learned significantly less than their tutoring groups, their tutoring groups (step based, substep based, and human tutoring) were not reliably different.

Figure 6b has more surprises. First, it seems that human tutors are 0.79 sigma more effective than no tutoring and not the 2.0 sigma found in the Bloom (1984) studies. The second surprise is that step-based tutoring is almost as good as human tutoring, with a 0.76 mean effect size. Although

the newest technology, substep-based tutoring, has few statistically reliable studies and a mixture of high and low effect sizes, Figure 6b suggests that its effect size may also lie approximately on the same plateau of 0.75 to 0.80.

The main conclusions are that (a) human tutoring ($d = 0.79$) is not as effective as we once thought, (b) ITS are almost as effective ($d = 0.76$ for step based, and probably for substep based, too) as human tutoring, and (c) there is an interaction plateau rather than as steady increase in effectiveness as granularity decreases.

DISCUSSION

Given that the mean effect size for human tutoring is much less than the 2.0 standard deviations that inspired a whole field, one might wonder if there is something wrong with the human tutors in these studies. In particular, were they expert tutors? The next section examines this question and briefly reviews the literature on human tutoring expertise. The following section focuses on studies that had especially high effect sizes to discern why they were atypically successful. The remaining sections discuss theoretical implications, limitations, and recommendations for future work.

Expert Versus Novice Human Tutors

There is no accepted definition of an expert human tutor. However, in many of the reviewed studies of human tutoring, the authors characterized their tutors as experts, often based on their years of experience.

The search within the constraints mentioned earlier (e.g., STEM content, etc.) uncovered several studies that compared the effectiveness of novice and expert tutors (see Table A10 in the Appendix) along with several studies that compared their behaviors without measuring their relative effectiveness (e.g., Cromley & Azevedo, 2005; Glass, Kim, Evens, Michael, & Rovick, 1999).[2] Almost all these studies found that novice tutors tended to lecture more, and expert tutors tended to be much more interactive.

The reviewed studies contain little evidence that expert tutors were more effective than novice tutors. Table A10 lists the relevant expert–novice comparisons along with several studies that manipulated the interactivity of the tutors. Although some absolute differences in effectiveness were in the expected direction, only two of these comparisons showed a reliable difference in learning gains. Moreover, the Cohen et al. (1982) meta-analysis found no relationship between tutor's experience and their effectiveness (given that they were subject-matter experts). Clark et al. (1976) found that giving subject-matter experts training and experience as tutors did not make them more effective.

These findings are consistent with the interaction plateau hypothesis. Although expert human tutors are more interactive than novice tutors, they are often no more effective than novice tutors. Moreover, constraining human tutors to be more or less interactive than they would normally be does not have much impact on their effectiveness. Basically, once tutoring has achieved a certain interactive granularity (roughly, step-based tutoring), decreases in interaction granularity apparently provide diminishing and sometimes even negligible returns.

Why Did Some Studies Have Such Large Effect Sizes?

For 25 years, researchers have been seeking solutions for Bloom's (1984) "2 sigma problem." Although one would expect many of the studies of human tutoring to show a 2.0 effect size, only two studies did. This section discusses those two studies, which now seem like outliers.

Bloom (1984) summarized six studies of human tutoring reported in the dissertations of Anania (1981) and Burke (1983). All six studies had effect sizes close to 2.0. Of these studies, only Anania's Experiment 3 was included in this review because only it involved one-on-one tutoring. The other five experiments summarized by Bloom involved each tutor working daily with a group of three students. However, Anania's one-on-one experiment did produce an effect size of 1.95, so let us examine it more closely.

A common explanation for the effectiveness of tutors in the studies discussed by Bloom is that they were highly trained, expert tutors. However, the original sources for Bloom's review say that the tutors were "undergraduate education majors" (Anania, 1981, p. 58) who "met the experimenter each day for one week before the instruction began" (Burke, 1983, p. 85) for training on both tutoring and the task domain: probability. This suggests that the Bloom tutors were not the "super tutors" that they have sometime been thought to be.

Anania's third experiment (and the other five Bloom experiments as well) included a third condition, which was mastery learning in the classroom. That is, after students had finished classroom instruction on a unit, they took a mastery test. If they scored 80%, then they were considered to have mastered the unit and could go on to the next unit. Students who scored less than 80% had to resume studying the unit and repeat the mastery test. In all six experiments, the mastery learning students scored about 1.0 standard deviations higher on posttests than the ordinary classroom students. Moreover, the tutoring conditions of all six experiments also involved mastery learning. That is, the tutees took the same mastery tests, restudied, and so on, but they worked with a tutor instead of a classroom teacher. However, the mastery threshold for the tutoring conditions was set at 90% instead of

[2]One study (di Eugenio et al., 2006) analyzed behaviors of tutors and compared their effectiveness but was not included in the meta-analysis because effect sizes could not be computed.

80% for the classroom implementation of mastery learning (Anania, 1981, pp. 44–45). That is, the tutors were holding their students to a higher standard of mastery than the classroom teachers. This alone could account for the advantage of tutoring (2.0 effect size) over mastery learning (1.0 effect size).

The second outlier study in the studies covered by this review was one of the baroreceptor experiments of Evens and Michael (2006, Table 10.3). The experiment found an effect size of 1.95 comparing human tutors to students who read the textbook instead. The number of subjects in this study was small, so the researchers repeated the experiment a few years later with more subjects and found an effect size of 0.52 (Evens & Michael, 2006, Table 10.4). Although the mean learning gains of the tutees were approximately the same in both experiments, the first experiment's control group ($N = 9$) had a much lower mean gain (0.33) than the mean gain (1.54) of the second experiment's control group ($N = 28$). In another experiment (Evens & Michael, 2006, Table 18.11) where reading was compared to computer tutoring, the same type of control group ($N = 33$) had a gain of 2.0. Although there were minor differences in the assessments across experiments, it appears that the mean learning gain of the control condition from the first, low-powered experiment may have been unusually low, perhaps due to a sampling artifact.

At any rate, the 1.95 effect sizes of both the Anania study and first Evens and Michael study were much higher than any other study of human tutoring versus no tutoring. The next highest effect size was 0.82. In short, it seems that human tutoring is not usually 2 sigmas more effective than classroom instruction, as the six studies presented by Bloom (1984) invited us to believe. Instead, it is closer to the mean effect size found here, 0.79. This is still a large effect size, of course.

Although Bloom's 2 sigma article now appears to be a demonstration of the power of mastery learning rather than human tutoring, it inspired a generation of research on human and computer tutoring that has vastly increased our knowledge and was well worth the effort. For instance, the research generated many valuable corpora of transcribed and analyzed tutorial dialogues that have shed many insights into human tutoring. Bloom's 2 sigma challenge inspired a whole new technology, dialogue-based tutoring, that required advances in dialogue management and robust language interpretation. These and other tutors now serve as testbeds for conducting well-controlled experiments on motivation, interaction, collaboration, and many other issues (see http://www.learnlab.org for examples).

Theoretical Implications

This section constructs an explanation for the observed interaction plateau. It starts by reconsidering the two hypotheses that were deemed most plausible for explaining why human tutoring should be more effective than computer tutoring. It might seem that the two hypotheses would have to conflict with the interaction plateau hypothesis, as they were originally used to motivate the now-discredited interaction granularity hypothesis. However, with only a few revised assumptions, the two hypotheses lead to a simple explanation of the plateau.

Hypothesis 7 was that the feedback of human tutoring helps students detect and repair their knowledge. That is, human tutorial feedback facilitates self-repair. For instance, if a student makes hundreds of mental inferences when solving a problem, and an answer-based tutoring system says that the answer is incorrect, then any of the hundred inferences may be wrong. This makes it difficult for students to find the incorrect inference and repair their knowledge. The answer-based tutoring system cannot be particularly helpful, because it too has little idea about which of the hundred inferences is wrong. On the other hand, if a *human* tutor is eliciting reasoning from the student as she works, and the tutor indicates that the student's most recent utterance is wrong, then the student knows that one of the most recent inferences is incorrect. There are only a few of them at most, so self-repair is much easier. Thus, self-repair is much easier when the feedback refers to a few inferences (human tutoring) than when it refers to many inferences (answer-based tutoring). This was Hypothesis 7's argument for the interaction granularity hypothesis.

Now a step-based tutoring system gives feedback on individual steps, either immediately or when the steps are submitted. Either way, students can examine the first incorrect step and know that one of the inferences that led to it must be wrong. As long as the tutoring system ensures that there is only a little reasoning required for each step, then compared to answer-based tutoring, students should find it much easier to find and fix the inference that caused a step to be flagged as incorrect. Moreover, step-based tutoring systems usually give hints that try to make it even easier for students to self-repair their knowledge. Thus, facilitating self-repair provides one explanation for the observed interaction plateau if we assume that debugging the reasoning behind an incorrect step during step-based tutoring is not much more difficult for students than debugging the reasoning behind an incorrect utterance to a human tutor.

Hypothesis 8 was that human tutoring scaffolds students, where "scaffold" means pushing them a little further along a line of reasoning via collaborative execution (e.g., prompting) and coordination (e.g., grounding; sharing knowledge). For instance, when a human tutor says to the student, "Sounds right to me. Keep going," the tutor is indicating mutual understanding (coordination), accepting the student's reasoning (collaborative execution), and indicating who should continue the execution (collaborative execution). A step-based tutoring system also scaffolds a student, but in different way. Whenever students enter a step that the tutor marks as correct, the student knows that the tutor understood the step

(coordination) and that the tutor agrees with the reasoning (collaborative execution). When the student gets stuck, both a human tutor and a step-based tutor will offer prompts and hints to get the student moving again. If these fail to get the student unstuck, then both human and step-based tutors do some of the reasoning themselves (collaborative execution), which is called the "bottom out" hint in the ITS literature. Thus, when the student gets stuck, explicit collaborative execution occurs with both human tutoring and step-base tutoring. Little of this scaffolding occurs with answer-based tutoring systems.

Although scaffolding and encouragement of self-repair probably have direct effects on learning, they may have an equally strong indirect effect by making it more likely that students finish problems correctly having done most of the reasoning themselves. Human tutors almost always get students to finish a problem correctly (Merrill et al., 1992). Many ITS (i.e., both step-based and substep-based tutoring systems) have such strong scaffolding and support for self-repairs that students often complete problems correctly (Schofield, 1995), and some ITS even require students to correctly solve the current problem before moving on to the next (e.g., Koedinger, Anderson, Hadley, & Mark, 1997). On the other hand, answer-based tutoring systems offer such weak scaffolding and feedback that students are usually allowed to give up after several failed attempts.

This factor (i.e., self-generating a solution vs. quitting) should have a strong effect on learning. When students solve a multistep problem correctly doing most of the reasoning themselves, then they are applying hundreds of knowledge components. Each time they apply a knowledge component, they do so in a new context and thus generalize it. They access it in memory and thus strengthen it. If they fail initially to retrieve an appropriate knowledge component, then they usually construct or reconstruct it (recall that we are assuming a self-generated correct solution). Similarly, if they apply a misconception, then they eventually realize their error and apply a correct knowledge component instead. In short, when students self-generate a correct solution, they generalize, strengthen, construct, and debug all the knowledge components required by the solution. Unfortunately, when they quit early, they miss hundreds of opportunities to learn.

This explanation, that all self-generated correct solutions are equally effective, was first proposed by Anderson et al. (1995), albeit only for step-based tutoring systems. Anderson et al. hypothesized that as long as students solved a set of problems doing most of the reasoning themselves, then their learning gains would be the same regardless of what kind of step-based tutoring they had. Anderson et al. supported this hypothesis by comparing several different versions of their tutoring systems. For instance, some tutoring systems offered immediate feedback, whereas other offered delayed feedback. In most of these experiments, when students in all experimental groups were required to complete all the problems correctly, the experimental manipulations did not affect their learning gains. On the other hand, the manipulations did affect efficiency, namely, the time to complete all the problems correctly. Extending the Anderson et al. (1995) hypothesis to all types of tutoring explains the observed interaction plateau, given the assumptions above.

In short, the explanation proposed here for the interaction plateau is that human tutors, step-based tutors, and substep-based tutors all provide enough scaffolding and feedback to get students to self-generate correct solutions for most problems. Even though step-based tutoring systems require students to bridge larger gaps than the finer granularity tutoring, students are apparently able to do so most of the time. This has both direct and indirect benefits. The direct benefit is that the scaffolding and feedback that gets them to bridge gaps correctly also causes them to construct or self-repair their knowledge. The indirect benefit is that, because students keep working on a solution instead of giving up, they encounter more learning opportunities. On the other hand, when students solve problems with an answer-based tutor or with no tutor, they often cannot bridge the long gap leading all the way from the start to the finish of the solution even when they get some feedback and perhaps even some scaffolding. When they fail to bridge the gap, they miss opportunities to learn.

This explanation is consistent with M. T. H. Chi's (2009) ICAP framework, which was discussed earlier as Hypothesis 9. According to the ICAP framework, interactive and constructive student behaviors can be equally effective, whereas active and passive student behaviors are less effective. The explanation proposed here is consistent with ICAP. The explanation predicts that students working with a human tutor would exhibit mostly interactive behavior and that students working with a step-based tutor would exhibit mostly constructive behavior. On the other hand, students working with an answer-based tutor or no tutor would often exhibit guessing and quitting, which are active student behaviors at best.

Limitations and Recommendations

When it comes to making practical recommendations, the conclusions presented here must be interpreted in the light of the limitations of the review, some of which are due to the inclusion/exclusion criteria. For instance, the researchers in these studies all tried to control for content, whereas in the real world, a tutor hired to help with physics may end up coaching a student on math or reading. Moreover, these studies only measured learning gains. Tutors may also boost students' motivation and efficiency.

Another limitation is that some of the comparisons in the review have only a small number of experiments testing them. More experiments are clearly needed. In particular, direct comparisons of human tutoring with various types of computer tutoring would be especially welcome. Although

thousands of students are covered in these studies, the number of human tutors involved is considerably smaller, so generalizing to all human tutors is risky.

It is important to note that none of the field studies in this review completely replaced all classroom instruction with tutoring. Instead, they replaced or partially replaced just one activity (usually homework) with tutoring. A classroom has many instructional activities that can have significant impacts on learning gains, so upgrading just one activity does not guarantee large overall course learning gains. On the other hand, if much of the students' learning goes on during homework, then replacing paper-based homework with an ITS can have a large effect size. For instance, in 4 year-long evaluations, the learning gains of students who used a step-based physics tutoring system were $d = 0.61$ higher than the learning gains of students who did the same homework assignments on paper (VanLehn et al., 2005).

Within the limitations of this article, one recommendation is that the usage of step-based tutoring systems should be increased. Although such tutoring systems are not cheap to develop and maintain, those costs do not depend on the number of tutees. Thus, when a tutoring system is used by a large number of students, its cost per hour of tutoring can be much less than adult one-on-one human tutoring. One implication of this review, again subject to its limitations, is that step-based tutoring systems should be used (typically for homework) in frequently offered or large enrollment STEM courses.

Another implication of this review is that human tutors have room for improvement. From the decades of studies of human tutoring, a frequent observation, which is sometimes mentioned (e.g., M. T. H. Chi, 1996) but rarely given the prominence it deserves, is that human tutors miss many opportunities to help students learn. This is not surprising given that they are mere humans doing a fast-paced, real-time, complex task. Although humans can gradually improve their performance on such tasks, it can take years of intensive, deliberate, reflective practice, and moreover, frequent, specific feedback on performance seems critical for improvement (Ericsson & Lehmann, 1996). Although some professional tutors do practice tutoring for 10 or more years, their practice is not like those of professional athletes, musicians, chess players, surgeons, and others, because they probably don't get frequent, specific feedback on their successes and failures, as do many other professionals (especially athletes). For instance, it is likely that few tutors video record and analyze their performances, looking for opportunities to improve. Thus, one could argue that although the tutors in these studies were called experts and have many years of tutoring experience, they may not really be as expert as a human could be given 10 years of constant feedback and reflective practice.

Compared to improving human tutoring, it should be relatively simple to improve the performance of ITS, that is, step-based tutors and substep-based tutors. Recent studies have found many pedagogical mistakes and missed opportunities in their performance as well (Baker, 2009; Baker, de Carvalho, Raspat, Corbett, & Koedinger, 2009; Murray & VanLehn, 2006). Merely finding and fixing the pedagogical mistakes of existing ITS may produce a 2 sigma effect size.

Such analyses can be partially or even fully automated. For instance, Min Chi and her colleagues found that a machine learning technique (reinforcement learning) could be applied to log data from a substep-based tutoring system in order to adjust the parameters that controlled its pedagogical decision making (M. Chi, VanLehn, Litman, & Jordan, 2011, in press). The improved tutoring system was $d = 0.84$ more effective than the original tutoring system. In short, we may soon see self-improving tutoring systems that monitor their own processes and outcomes in order to modify their tutoring tactics and make them more effective.

In short, the bottom line is this: For ITS, although decreasing the granularity of the user interface does not seem to provide additional benefit, reengineering the tutor–student interactions may provide considerable additional benefit. For human tutors, although merely interacting more frequently with students does not seem to provide additional benefits, years of deliberate practice may allow human tutors to improve their effectiveness. It is worth remembering that no classroom teacher has been replaced by an ITS, but classroom instruction is often replaced by human tutoring, for example, in home schooling. We need both good human tutors and good ITS.

The field's future work is clear. Tutoring researchers should retain Bloom's challenge and strive to develop both computer and human tutors that are 2 standard deviations more effective than no tutoring.

The Take-Home Points

For more than 20 years, researchers in tutoring have held a mental image something like Figure 1: Effect sizes increase monotonically as the interaction granularity of tutoring decreases and culminate in Bloom's $d = 2.0$ for human tutoring. As discussed earlier, Bloom's $d = 2.0$ effect size seems to be due mostly to holding the tutees to a higher standard of mastery. That is, the tutees had to score 90% on a mastery exam before being allowed to continue to the next unit, whereas students in the mastery learning classroom condition had to score 80% on the same exam, and students in the classroom control took the exams but always went on to the next unit regardless of their scores. So the Bloom (1984) article is, as Bloom intended it to be, a demonstration of the power of mastery learning rather than a demonstration of the effectiveness of human tutoring.

If the familiar image of Figure 1 is no longer supported by Bloom's studies, then what is a more accurate image? Figure 6b presents the effect sizes reviewed here. It shows that effectiveness increases from 0.31 (answer-based tutoring) to 0.76 (step-based tutoring), then seems to hit a plateau. Further

decreases in user interface granularity (substep-based tutoring; human tutoring) do not increase effectiveness. Although more experimental data would be welcome, the interaction plateau of Figure 6b appears to be the best image so far of the relative effectiveness of different types of tutoring.

Perhaps most important, this progress report also shows that ITS are, within the limitations of this article, just as effective as adult, one-on-one human tutoring for increasing learning gains in STEM topics. Lest there be any misunderstanding due to the unfortunate choice of "tutoring" as part of the name of such systems, none of the studies reported here even attempted to replace a classroom teacher with ITS even though it is not uncommon for a human tutor to replace a classroom teacher. As argued earlier, ITS should be used to replace homework, seatwork, and perhaps other activities but not to replace a whole classroom experience. Nonetheless, within their limited area of expertise, currently available ITS seem to be just as good as human tutors.

ACKNOWLEDGMENTS

I am grateful for the close readings and thoughtful comments of Michelene T. H. Chi, Dexter Fletcher, Jared Freedman, Kasia Muldner, and Stellan Ohlsson. My research summarized here was supported by many years of funding from the Office of Naval Research (N00014-00-1-0600) and National Science Foundation (9720359, EIA-0325054, 0354420, 0836012, and DRL-0910221).

REFERENCES

Aleven, V., Ogan, A., Popescu, O., Torrey, C., & Koedinger, K. R. (2004). Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In J. Lester, R. M. Vicari & F. Paraguaca (Eds.), *Intelligent tutoring systems: Seventh International Conference, ITS 2004* (pp. 443–454). Berlin, Germany: Springer.

Anania, J. (1981). *The effects of quality of instruction on the cognitive and affective learning of students. Unpublished doctoral dissertation*, University of Chicago, Chicago, IL.

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

Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods, 40,* 672–694.

Azevedo, R., Greene, J. A., & Moos, D. C. (2007). The effect of a human agent's external regulation upon college students' hypermedia learning. *Metacognition and Learning, 2,* 67–87.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56,* 45–72.

Baker, R. S. J. d. (2009). Differences between intelligent tutor lessons, and the choice to go off-task. In *Proceedings for the 2nd International Conference on Educational Data Mining* (pp. 11–20). Córdoba, Spain: Universidad de Córdoba.

Baker, R. S. J. d., de Carvalho, A. M. J. A., Raspat, J., Corbett, A., & Koedinger, K. R. (2009). Educational software features that encourage or discourage "gaming the system." *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 475–482). Amsterdam, The Netherlands: IOS.

Bausell, R. B., Moody, W. B., & Walzl, F. H. (1972). A factorial study of tutoring versus classroom instruction. *American Educational Research Journal, 9,* 591–597.

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.

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., & Lester, J. C. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Intelligent tutoring systems: 9th International Conference, ITS2008* (pp. 239–249). Berlin, Germany: Springer.

Burke, A. J. (1983). *Student's potential for learning contrasted under tutorial and group approaches to instruction. Unpublished doctoral dissertation*, University of Chicago, Chicago, IL.

Cabalo, J. V., Jaciw, A., & Vu, M.-T. (2007). *Comparative effectiveness of Carnegie Learning's Cognitive Tutor Algebra I curriculum: A report of a randomized experiment in the Maui School District*. Palo Alto, CA: Empirical Education.

Cade, W. L., Copeland, J. L., Person, N., & D'Mello, S. K. (2008). Dialogue modes in expert tutoring. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Intelligent tutoring systems: 9th International Conference, ITS2008* (pp. 470–479). Berlin, Germany: Springer.

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

Carnegie Learning. (2010). *Cognitive tutor effectiveness*. Retrieved from http://www.carnegielearning.com/static/web_docs/2010_Cognitive_Tutor_Effectiveness.pdf

Chae, H. M., Kim, J. H., & Glass, M. S. (2005, May). *Effective behaviors in a comparison between novice and expert algebra tutors*. Paper presented at the Proceedings of the Sixteenth Midwest Artificial Intelligence and Cognitive Science (MAICS–2005) Conference, Dayton, OH.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21,* 99–135.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (in press). An evaluation of pedagogical tutorial tactics for a natural language tutoring systems: A reinforcement learning approach. *International Journal of Applied Artificial Intelligence*.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* S33–S49.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73–105.

Chi, M. T. H. (2011). *The ICAP Module: Guidelines for teachers to increase students' engagement with learning. Unpublished funded proposal to IES*, Arizona State University, Tempe.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–342.

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22,* 363–387.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533.

Cho, B.-I., Michael, J. A., Rovick, A. A., & Evens, M. W. (2000). An analysis of multiple tutoring protocols. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems: 5th International Conference* (pp. 212–221). Berlin, Germany: Springer.

Christmann, E., & Badgett, J. (1997). Progressive comparison of the effects of computer-assisted on the academic achievement of secondary

students. *Journal of Research on Computing in Education, 29,* 325–338.

Clark, C. M., Snow, R. E., & Shavelson, R. J. (1976). Three experiments on learning to teach. *Journal of Teacher Education, 17,* 174–180.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

Collins, A., & Stevens, A. (1982). Goals and strategies for inquiry teachers. In R. Glaser (Ed.), *Advances in instructional psychology, Vol. 2* (pp. 65–119). Hillsdale, NJ: Erlbaum.

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

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *User modeling: Proceedings of the Eighth International Conference* (pp. 137–147). Berlin, Germany: Springer.

Corbett, A., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko, A. Sears, M. Beaudouin-Lafon, & R. Jacob (Eds.), *Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems* (pp. 245–252). New York, NY: ACM Press.

Corbett, A., Wagner, A. Z., Lesgold, S., Ulrich, H., & Stevens, S. M. (2006). The impact of learning of generating vs. selecting descriptions in analyzing algebra example solutions. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *The 7th International Conference of the Learning Sciences* (pp. 99–105). Mahwah, NJ: Erlbaum.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88,* 715–730.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 67–74). Morristown, NJ: Association of Computational Linguistics.

Core, M. G., Traum, D. R., Lane, H. C., Swartout, W., Gratch, J., van Lent, M., & Marsella, S. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation, 82,* 685–702.

Craig, S. D., Driscoll, D. M., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning and pedagogical agents. *Journal of Educational Multimedia and Hypermedia, 13,* 163–183.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 563–589.

Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes, 40,* 83–113.

di Eugenio, B., Kershaw, T., Lu, X., Corrigan-Halpern, A., & Ohlsson, S. (2006, May). *Toward a computational model of expert tutoring: A first report*. Paper presented at the 19th International FLAIRS conference, Melbourne Beach, FL.

D'Mello, S. K., King, B., Stolarski, M., Chipman, P., & Graesser, A. C. (2007, October). *The effects of speech recognition errors on learner's contributions, knowledge, emotions, and interaction experience*. Paper presented at the SLATE, Farmington, PA.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305.

Evens, M., & Michael, J. (2006). *One-on-one tutoring by humans and machines*. Mahwah, NJ: Erlbaum.

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

Fletcher-Flinn, C. M., & Gravatt, B. (1995). The efficacy of computer-assisted instruction (CAI): A meta-analysis. *Journal of Educational Computing Research, 12,* 219–241.

Forbes-Riley, K., & Litman, D. (2008). Dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In L. Dybkjaer & W. Minker (Eds.), *Recent trends in discourse and dialogue* (vol 39, pp. 275–304). Berlin, Germany: Springer.

Fossati, D. (2008, June). *The role of positive feedback in intelligent tutoring systems*. Paper presented at the 46th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop, Columbus, OH.

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, UK.

Fox, B. A. (1991). Cognitive and interactional aspects of correction in tutoring. In P. Goodyear (Ed.), *Teaching knowledge and intelligent tutoring* (pp. 149–172). Norwood, NJ: Ablex.

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems*. Hillsdale, NJ: Erlbaum.

Frederiksen, N., Donin, J., & Roy, M. (2000, June). *Human tutoring as a model for computer tutors: Studying human tutoring from a cognitive perspective*. Paper presented at the Modeling Human Teaching Tactics and Strategies: Workshop W1 of ITS2000, Montreal, Canada.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., . . . Craig, S. D. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instructional Science, 37,* 487–493.

Glass, M. S., Kim, J. H., Evens, M., Michael, J., & Rovick, A. (1999, April). *Novice vs. expert tutors: A comparison of style*. Paper presented at the Proceedings of the Midwest Artificial Intelligence and Cognitive Science Society, Bloomington, IN.

Gott, S. P., Lesgold, A., & Kane, R. S. (1996). Tutoring for transfer of technical competence. In B. G. Wilson (Ed.), *Constructivist learning environments* (pp. 33–48). Englewood Cliffs, NJ: Educational Technology.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments and Computers, 36,* 180–193.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359–387.

Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39–41.

Hastings, P., Arnott-Hill, E., & Allbritton, D. (2010). Squeezing out gaming behavior in a dialog-based ITS. In V. Aleven, H. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems 2010* (pp. 204–213). Berlin, Germany: Springer-Verlag.

Heffernan, N. T., Koedinger, K. R., & Razzaq, L. (2008). Expanding the Model-Tracing Architecture: A 3rd generation intelligent tutor for algebra symbolization. *International Journal of Artificial Intelligence in Education, 18,* 153–178.

Heller, J. I., & Reif, F. (1984). Prescribing effective human problem-solving processes: Problem descriptions in physics. *Cognition and Instruction, 1,* 177–216.

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin, 128,* 774–795.

Herrmann, T. (1982, August). *Effective tutoring in a PSI course: Person vs. computer*. Paper presented at the annual convention of the American Psychological Association, Washington, DC.

Hume, G., Michael, J., Rovick, A., & Evens, M. (1996). Hinting as a tactic in one-on-one tutoring. *Journal of the Learning Sciences, 5*(1), 23–49.

Jeong, H., Siler, S., & Chi, M. T. H. (1997). Can tutors diagnose students' understanding? In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 959). Mahwah, NJ: Erlbaum.

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.

Johnson, H., & Johnson, P. (1992, January). *Different explanatory dialogue styles and their effects on knowledge acquisition by novices*. Paper presented at the Procceedings of the Twenty-fifth Hawaii International Conference on System Science, Kauai, HI.

Johnson, S. D., Flesher, J. W., Ferej, A., & Jehn, J.-C. (1992). *Application of cognitive theory to the design, development, and implementation of a computer-based troubleshooting tutor*. Berkeley, CA: National Center for Research in Vocational Education.

Johnson, S. D., Flesher, J. W., Jehng, J.-C., & Ferej, A. (1993). Enhancing electrical troubleshooting skills in a computer-coached practice environment. *Interactive Learning Environments, 3,* 199–214.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education, 13,* 79–116.

Katz, S., Connelly, J., & Wilson, C. (2007). Out of the lab and into the classroom: An evaluation of reflective dialogue in Andes. In R. Luckin & K. R. Koedinger (Eds.), *Proceedings of AI in Education, 2007* (pp. 425–432). Amsterdam, The Netherlands: IOS.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback intervention on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin, 112,* 254–284.

Koedinger, K. R., & Anderson, J. R. (1993). Effective use of intelligent software in high school math classrooms. In P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial intelligence in education: Proceedings of the World Conference on AI in Education*. Charlottesville, VA: AACE.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*(1), 30–43.

Kulik, C., & Kulik, J. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7,* 75–91.

Kulik, J. A. (1994). Meta-analytic studies of findings on computer-based instruction: An updated analysis. In E. L. Baker & H. F. O'Neil (Eds.), *Technology assessment in education and training* (pp. 9–33). Mahwah, NJ: Erlbaum.

Kulik, J. A., Bangert, R. L., & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. *Journal of Educational Psychology, 75,* 19–26.

Kulik, J., Kulik, C., & Bangert-Drowns, R. (1985). Effectiveness of computer-based education in elementary schools. *Computer in Human Behavior, 1,* 59–74.

Kulik, J., Kulik, C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Educational Research, 50,* 524–544.

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

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135–158). New York, NY: Academic.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J.-L. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 75–105). Hillsdale, NJ: Erlbaum.

Litman, D., Rose, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence and Education, 16,* 145–170.

Mark, M. A., & Greer, J. E. (1995). The VCR tutor: Effective instruction for device operation. *The Journal of the Learning Sciences, 4,* 209–246.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7,* 197–244.

Mendicino, M., & Heffernan, N. T. (2007). *Comparing the learning from intelligent tutoring systems, non-intelligent computer-based versions and traditional classroom instruction*. Manuscript submitted for publication.

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

Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and Instruction, 13,* 315–372.

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences, 2,* 277–306.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated agents? *Cognition and Instruction, 19,* 177–213.

Murray, R. C., & VanLehn, K. (2006). A comparison of decision-theoretic, fixed-policy and random tutorial action selection. In M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems: 8th International Conference ITS 2006* (pp. 114–123). Berlin, Germany: Springer.

Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., & Kershaw, T. C. (2007). Beyond the code-and-count analysis of tutoring dialogues. In *Artificial intelligence in education* (pp. 349–356). Amsterdam, The Netherlands: IOS.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117–175.

Parvez, S. M. (2007). *A pedagogical framework for integrating individual learning style into an intelligent tutoring system* Unpublished doctoral dissertation, Lehigh University, Lehigh, PA.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence and Education, 16,* 171–194.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*(1), 13–48.

Razzaq, L., Mendicino, M., & Heffernan, N. T. (2008). Comparing classroom problem-solving with no feedback to Web-based homework assistance. In B. P. Woolf, E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Intelligent tutoring systems: 9th International Conference, ITS2008* (pp. 426–437). Berlin, Germany: Springer.

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

Ritter, F., & Feurzeig, W. (1988). Teaching real-time tactical thinking. In J. Psotka, L. D. Massey, & S. A. Mutter (Eds.), *Intelligent tutoring systems: Lessons Learned* (pp. 285–301). Hillsdale, NJ: Erlbaum.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research, 79*(1), 3–38.

Ritter, S., Kulikowich, J., Lei, P., 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. Young (Eds.), *Supporting learning flow through integrative technologies* (pp. 12–20). Amsterdam, The Netherlands: IOS Press.

Roberts, B. (2001, November). *COVE—-A shiphandling trainer with an attitude*. Paper presented at the Interservice/Industry Training, Simulation and Education Conference, Orlando, FL.

Rose, C. P., Aleven, V., Carey, R., & Robinson, A. (2005). A first evaluation of the instructional value of negotiable problem solving goals on the exploratory learning continuum. In G. I. Mcalla & C.-K. Looi (Eds.), *Proceedings of the Artificial Intelligence in Education Conference* (pp. 563–570). Amsterdam, The Netherlands: IOS Press.

Rose, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2001). A comparative evaluation of Socratic versus didactic tutoring. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 897–902). Mahwah, NJ: Erlbaum.

Schofield, J. W. (1995). *Computers, classroom culture and change*. Cambridge, UK: Cambridge University Press.

Scott, L. A., & Reif, F. (1999). Teaching scientific thinking skills: Students and computers coaching each other. In S. P. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 285–293). Amsterdam, The Netherlands: IOS Press.

Scruggs, T. E., & Richter, L. (1985). Tutoring learning disabled students: A critical review. *Learning Disabled Quarterly, 8,* 286–298.

Shah, F., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes, 33*(1), 23–52.

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

Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence in Education, 4*(1), 61–93.

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world. *Interactive Learning Environments, 1,* 51–76.

Siler, S. (2004). *Does tutors' use of their knowledge of their students enhance one-to-one tutoring?* University of Pittsburgh, Pittsburgh, PA.

Siler, S., Rose, C. P., Frost, T., VanLehn, K., & Koehler, P. (2002, June). *Evaluating knowledge construction dialogues (KCDs) versus minilesson within Andes2 and alone*. Paper presented at the Workshop on Dialogue-Based Tutoring at ITS 2002, Biaritz, France.

Siler, S., & VanLehn, K. (2009). Learning, interactional, and motivational outcomes in one-to-one synchronous computer-mediate versus face-to-face tutoring. *International Journal of Artificial Intelligence and Education, 19,* 73–102.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13,* 551–568.

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

Smith, S. G., & Sherwood, B. A. (1976). Educational uses of the PLATO computer system. *Science, 192,* 344–352.

Stankov, S., Rosic, M., Zitko, B., & Grubisic, A. (2008). TEx-Sys model for building intelligent tutoring systems. *Computer and Education, 51,* 1017–1036.

Steuck, K., & Miller, T. M. (1997, March). *Evaluation of an authentic learning environment for teaching scientific inquiry skills*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Stottler, R., & Panichas, S. (2006, December). *A new generation of tactical action officer intelligent tutoring system (ITS)*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conferece (I/ITSEC), Orlando, FL.

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, ITS 2002* (pp. 377–387). Berlin, Germany: Springer-Verlag.

Swanson, J. H. (1992, April). *What does it take to adapt instruction to the individual? A case study of one-to-one tutoring*. Paper presented at the American Education Research Association, San Francisco, CA.

Timms, M. J. (2007). Using Item Response Theory (IRT) to select hints in an ITS. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education* (pp. 213–221). Amsterdam, The Netherlands: IOS.

VanLehn, K. (1988). Student modeling. In M. Polson & J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 55–78). Hillsdale, NJ: Erlbaum.

VanLehn, K. (1999). Rule learning events in the acquisition of a complex skill: An evaluation of Cascade. *Journal of the Learning Sciences, 8,* 179–221.

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

VanLehn, K. (2008a). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). New York, NY: Erlbaum.

VanLehn, K. (2008b). The Interaction Plateau: Answer-based tutoring < step-based tutoring = natural tutoring (abstract only). In B. P. Woolf, E. Aimeur, R. Nkambou, & S. P. Lajoie (Eds.), *Intelligent tutoring systems 2008* (p. 7). Berlin, Germany: Springer-Verlag.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3–62.

VanLehn, K., Lynch, C., Schultz, K., Shapiro, J. A., Shelby, R. H., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education, 15,* 147–204.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Human tutoring: Why do only some events cause learning? *Cognition and Instruction, 21,* 209–249.

Wasik, B. A. (1998). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly, 33,* 266–291.

Weerasinghe, A., & Mitrovic, A. (2006). Facilitating deep learning through self-explanation in an open-ended domain. *International Journal of Knowledge-based and Intelligent Engineering Systems, 10,* 3–19.

What Works Clearinghouse. (2009). *WWC intervention report: Cognitive Tutor Algebra I.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_cogtutor_app_072809.pdf

What Works Clearinghouse (2010). *WWC intervention report: Carnegie learning curricula and cognitive tutor software*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_cogtutor_app_083110.pdf

Wittwer, J., Nuckles, M., Landmann, N., & Renkl, A. (2010). Can tutors be supported in giving effective explanations? *Journal of Educational Psychology, 102*(1), 74–89.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17,* 89–100.

Woolf, B. P. (2009). *Building intelligent interactive tutors*. Burlington, MA: Morgan Kaufman.

Zachary, W., Cannon-Bowers, J., Bilazarian, P., Krecker, D., Lardieri, P., & Burns, J. (1999). The Advanced Embedded Training System (AETS): An intelligent embedded tutoring system for tactical team training. *International Journal of Artificial Intelligence in Education, 10,* 257–277.

APPENDIX
TABLE A1
Step-Based Tutoring Compared to No Tutoring

d	*Citation and Comments*
1.34*	(VanLehn et al., 2007, Exp. 2) The SBT coached qualitative physics explanation generation using text as remediation. NT was reading a textbook.
–0.32	(VanLehn et al., 2007, Exp. 6) The SBT coached qualitative physics explanation generation using text as remediation. NT was an experimenter-written text.
1.01*	(Moreno, Mayer, Spires, & Lester, 2001, Exp. 1) The SBT (Design-a-Plant) coached combinations of stems, roots and leaves. NT was click-through text and graphics.
0.98*	(Moreno et al., 2001, Exp. 2) Nearly the same as Exp. 1.
0.68*	(Moreno et al., 2001, Exp. 3) Nearly the same as Exp. 1, but NT was an agent who lectured.
0.61*	(VanLehn et al., 2005) The SBT (Andes) coached physics problem solving for a whole semester; NT was paper-based homework; results aggregated experiments over 4 years.
0.98*	(Corbett & Anderson, 2001). The SBT coached Lisp coding. NT was doing coding exercises without any feedback in standard Lisp environment.
1.00*	(Anderson et al., 1995, p. 177) The SBT coached Lisp coding. NT was doing coding exercises without any feedback in standard Lisp environment.
1.00*	(Anderson et al., 1995, p. 183, para. 1) The SBT coached geometry theorem proving. NT was paper-and-pencil homework.
1.00*	(Koedinger & Anderson, 1993) The SBT (ANGLE) coached geometry problem solving. NT was paper and pencil. This result was for classes taught by a project member.
–0.23	(Koedinger & Anderson, 1993) Same as above, but this result is for classes taught by a nonproject teacher.
0.61	(Mendicino, Razzaq, & Heffernan, 2009) The SBT (ASSISTments) coached solving of math problems from standardized tests. NT is paper-based homework discussed in class the next day.
0.28	(Shute & Glaser, 1990) The SBT (Smithtown) coached microeconomics inquiry. NT was classroom instruction.
1.35*	(Mark & Greer, 1995) The SBT coached programming a video recorder. NT was just doing each step as told to do so.
1.27*	(Gott, Lesgold, & Kane, 1996) The SBT (Sherlock II) coached troubleshooting. The NT was experience in troubleshooting a test station.
0.70*	(Timms, 2007) The SBT (FOSS) coached speed, distance, time problem solving. NT was doing the same exercises without any feedback.
0.94*	(Parvez, 2007) The SBT (DesignFirst-ITS) coached UML design. NT was solving UML design problems without feedback.
0.81*	(Stankov, Rosic, Zitko, & Grubisic, 2008) The SBT (TEx-Sys) had students construct or debug semantic network representations of introductory college computer science knowledge. NT had students engage in "traditional teaching and learning processes" for the same duration (14 weeks).
0.60	Ditto, 8th-grade chemistry for 10 weeks
0.75	Ditto, 8th-grade physics for 7 weeks
1.23*	Ditto, College introductory computer science for 14 weeks
0.71	Ditto, College BASIC programming for 14 weeks
1.36*	Ditto, 6th-grade mathematics for 7 weeks
0.17	Ditto, 8th-grade mathematics for 7 weeks
0.34	Ditto, 5th-grade mathematics for 5 weeks
1.17*	(S. D. Johnson, Flesher, Ferej, & Jehn, 1992; S. D. Johnson, Flesher, Jehng, & Ferej, 1993) The SBT (Technical Troubleshooting Tutor) coached students troubleshooting simulated avionic electrical systems. The NT students solved the same troubleshooting problems using real instead of simulated electrical systems.
0.57*	(B. G. Johnson, Phillips, & Chase, 2009) The SBT (Quantum Accounting Tutor) coached students as they analyzed transactions. The NT students used paper and their textbook to solve the same problems.
0.47*	(Steuck & Miller, 1997) The SBT (ISIS) taught scientific inquiry in the context of high school biology. This was a yearlong field study involving 84 sections and 1,547 students. Roughly half the sections used the tutor, and the other half had regular biology labs instead.
0.76	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). Exp. = Experiment; NT = no tutoring; SBT = step-based (computer) tutoring; UML = Unified Modeling Language.

TABLE A2
Substep-Based Tutoring Compared to No Tutoring

d	*Citation and Comments*
1.34*	(VanLehn et al., 2007, Exp. 2) SSBT (Why2-AutoTutor) coached qualitative physics explanation generation. NT was reading a textbook.
–0.21	(VanLehn et al., 2007, Exp. 6) SSBT was Why2-Atlas. NT was reading targeted text.
–0.44	(VanLehn et al., 2007, Exp. 6) SSBT was Why2-AutoTutor. NT was reading targeted text.
–0.32	(Siler, Rose, Frost, VanLehn, & Koehler, 2002, Exp. 2) SSBT was KCDs; NT was reading targeted text (mini-lessons).
–0.32	(Siler et al., 2002, Exp. 3) SSBT was KCDs; NT was reading targeted text (mini-lessons).
0.96*	(Lane & VanLehn, 2005) SSBT (PROPEL) taught novice programming. NT was click-through text.
0.35*	(Graesser et al., 2004, Table 1, Exp. 1) SSBT (AutoTutor 1.1) taught computer literacy; NT was rereading the textbook.
0.46	(Graesser et al., 2004, Table 1, Exp. 2) SSBT (AutoTutor 1.2) taught computer literacy; NT was reading the textbook.
0.45	(Graesser et al., 2004, Table 1, Exp. 3) SSBT (AutoTutor) compared to reading a reduced textbook.
0.76*	(Arnott, Hastings, & Allbritton, 2008) The RMT taught research methods via a combination of SSBT and ABT. NT was different sections of the same class but without the occasional tutoring assignments.
1.21*	(Hastings, Arnott-Hill, & Allbritton, 2010) Same as above but different population.
0.54*	(Craig, Driscoll, & Gholson, 2004, Exp. 1) SSBT (AutoTutor) taught computer literacy. NT was watching a video of an AutoTutor session.
0.71*	(Craig et al., 2004, Exp. 2) SSBT (AutoTutor) taught computer literacy. NT was watching a video of an AutoTutor session.
0.53	(Craig, Sullins, Witherspoon, & Gholson, 2006, Exp. 1) SSBT (AutoTutor) taught computer literacy. NT was watching a video of an AutoTutor session.
0.28	(Craig et al., 2006, Exp. 2) SSBT (AutoTutor) taught computer literacy. NT was watching a video of an AutoTutor session.
–0.07	(Craig, et al., 2006, Exp. 1) SSBT (AutoTutor) taught computer literacy. NT is the AutoTutor agent lecturing and not asking questions.
0.59*	(Mendicino & Heffernan, 2007, Exp. 1) SSBT (Ms. Lindquist) coached solving of algebra word problems in class. NT was paper-and-pencil problem solving without feedback in class.
0.54	(Mendicino & Heffernan, 2007, Exp. 2) SSBT (Ms. Lindquist) used as homework. NT was paper-and-pencil problem solving without feedback in class. Aggregation of SSBT and CAI conditions reported in (Razzaq, Mendicino, & Heffernan, 2008).
0.65*	(Katz, Connelly, & Wilson, 2007, Exp. 2) SSBT was discussion of answers to postproblem reflection questions; NT was extra problem solving.
0.12	(Gholson et al., 2009) SSBT (AutoTutor) taught computer literacy to 8th graders; NT was same content presented as a monologue.
–0.20	(Gholson et al., 2009) SSBT (AutoTutor) taught computer literacy to 9th graders; NT was same content presented as a monologue.
0.00	(Gholson et al., 2009) SSBT (AutoTutor) taught conceptual physics to 10th graders; NT was same content presented as a monologue.
–0.13	(Gholson et al., 2009) SSBT (AutoTutor) taught conceptual physics to 11th graders; NT was same content presented as a monologue.
1.67*	(Evens & Michael, 2006, Table 18.7) SSBT (an early version CIRCSIM-Tutor) coached medical students on cardiophysiology; NT was reading a textbook.
0.46*	(Evens & Michael, 2006, p. 356) SSBT (final version of CIRCIM-Tutor) coached medical students on cardiophysiology; NT was reading a textbook.
0.49*	(Mendicino & Heffernan, 2007, Exp. 3) SSBT (Ms. Lindquist) was used in class to coach algebra word problem solving. NT was paper-and-pencil problem solving without feedback in class.
0.40	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). ABT = answer-based (computer) tutoring; Exp. = Experiment; KCD = Knowledge Construction Dialogue; NT = no tutoring; RMT = Research Methods Tutor; SSBT = substep-based (computer) tutoring.

TABLE A3
Human Tutoring Versus No Tutoring

d	*Citation and Comments*
1.95*	(Anania, 1981, Exp. 3) HT taught cartography one-on-one. NT was ordinary class.
0.67*	(Azevedo, Greene, & Moos, 2007) NT was studying hypertext on blood circulation. HT coached students as they did so.
0.65*	(Swanson, 1992) HT taught students how to run an optics experiment. NT was watching a demonstration.
0.66*	(M. T. H. Chi et al., 2008) HT coached physics problem solving. NT was students solving problems with a video of a tutoring session that they could watch and copy from.
0.82*	(M. T. H. Chi et al., 2008) HT coached physics problem solving. NT was students solving problems with a textbook available.
0.55*	(Azevedo, Moos, Greene, Winters, & Cromley, 2008) NT was studying hypertext on blood circulation. HT coached students as they did so.
0.38*	(Bausell, Moody, & Walzl, 1972) HT taught exponential notation. NT was classroom instruction for exactly the same amount of time.
1.95*	(Evens & Michael, 2006, Table 10.3) HT were experts who taught medical students about the baroreceptor reflex; NT was reading a textbook.
0.52	(Evens & Michael, 2006, Table 10.4) Same as above, with larger number of subjects ($N = 53$ here; $N = 15$ above).
–0.24	(Evens & Michael, 2006, Table 10.6) Same as above, with novice tutors instead of expert tutors. Control condition from Table 10.3 used as comparison.
0.79	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). Exp. = Experiment; HT = human tutoring; NT = no tutoring.

TABLE A4
Step-Based Tutoring Versus Answer-Based Tutoring

d	*Citation and Comments*
0.17	(Suraweera & Mitrovic, 2002, Exp.1) SBT (KERMIT) taught database design tutor. ABT was a version that only gave final answer feedback.
0.63*	(Suraweera & Mitrovic, 2002, Exp. 2) Same as Exp. 1 but with longer training times.
0.40	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). Exp. = Experiment; SBT = step-based (computer) tutoring.

TABLE A5
Substep-Based Tutoring Versus Answer-Based Tutoring

d	*Citation and Comments*
0.56*	(Heffernan, Koedinger, & Razzaq, 2008) SSBT (Ms. Lindquist) coached solving algebra word problems. ABT gave feedback on final answer but did not require correction of it.
0.41	(Mendicino & Heffernan, 2007, Exp. 1) SSBT (Ms. Lindquist) coached algebra word problems. ABT just gave feedback on final answer but did not require correction of it. Results briefly reported in Heffernan et al. (2008).
0.45	(Mendicino & Heffernan, 2007, Exp. 2) Same as Exp. 1, but tutoring done as homework instead of classroom work
0.49	(Mendicino & Heffernan, 2007, Exp. 3) Same as Exp. 1, with 3 classroom teachers instead of 1.
0.34*	(Arnott et al., 2008) The RMT taught half its modules on research methods using SSBT and half using ABT.
–0.34	(Hastings et al., 2010) Same as above; different population.
0.32	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). ABT = answer-based (computer) tutoring; Exp. = Experiment; RMT = Research Methods Tutor; SSBT = substep-based (computer) tutoring.

TABLE A6
Human Tutoring Versus Answer-Based Tutoring

d	*Citation and Comments*
−0.04	(Herrmann, 1982) In the context of a mastery-learning instruction, human or computer tutors give feedback on incorrect answers by pointing students to the appropriate section of the instructional resources.

TABLE A7
Substep-Based Tutoring Versus Step-Based Tutoring

d	*Citation and Comments*
−0.10	(VanLehn et al., 2007, Exp. 1) SSBT (Atlas) coached qualitative physics explanation. SBT explained the steps that they missed & requiring them to try again.
−0.37	(VanLehn et al., 2007, Exp. 1) Same as Exp.1 with AutoTutor as the SSBT.
0.41	(VanLehn et al., 2007, Exp. 3) Same as Exp. 1, but with more extensive assessments. SSBT was AutoTutor.
0.55	(VanLehn et al., 2007, Exp. 5) Same as Exp. 1 but with physics novices. SSBT was Atlas.
0.23	(VanLehn et al., 2007, Exp. 5) Same as Exp. 1, but with physics novices. AutoTutor as SSBT
0.11	(VanLehn et al., 2007, Exp. 6) Same as Exp. 5 but with simplified materials. SSBT was Atlas.
−0.13	(VanLehn et al., 2007, Exp. 6) Same as Exp. 5 but with simplified materials. SSBT was AutoTutor.
−0.11	(Weerasinghe & Mitrovic, 2006) SSBT (Kermit-SE) taught database design with substep-based remediation. SBT (Kermit) was identical but used canned text remediation.
0.11	(Siler et al., 2002, Exp. 1) SSBT (Andes-KCD) used substep-based help while teaching physics. SBT (Andes) used mini-lessons (text) instead of dialogues.
0.24	(Evens & Michael, 2006, Table 18.7) SSBT (CIRCSIM-Tutor) used substep-based remediation of errors made in a physiology activity. SBT (CIRCSIM) used text instead of dialogue.
0.80	(Katz et al., 2007, Exp. 1) SSBT was a substep-based discussion of post-problem reflection questions. SBT was same questions followed by canned text explanation.
0.16	Mean effect size

Note. Exp. = Experiment; SBT = step-based (computer) tutoring; SSBT = substep-based (computer) tutoring.

TABLE A8
Human Tutoring Versus Step-Based Tutoring

d	*Citation and Comments*
0.24	(Fossati, Di Eugenio, Brown, & Ohlsson, 2008) HT and SBT (iList-1) taught students how to manipulate linked-list data structures.
0.08	(Fossati et al., 2009) HT and SBT (iList-3) taught students how to manipulate linked-list data structures.
−0.34	(Katz et al., 2003, Exp. 2) HT discussed reflection questions after solution of a physics problem. SBT presented such questions and a text answer and explanation.
−0.28	(VanLehn et al., 2007, Exp. 1) HT coached students to improve their answers to qualitative physics questions. SBT presented correct answers as click-through text.
1.57*	(VanLehn et al., 2007, Exp. 4) Same as Exp. 1, but students had not yet taken college physics. HT was spoken.
0.85*	(VanLehn et al., 2007, Exp. 4) Same as above, but HT was typed.
−0.07	(VanLehn et al., 2007, Exp. 4) Same as Exp. 4 with spoken tutoring, but difficulty of material was reduced.
0.62*	(Rose, Aleven, Carey, & Robinson, 2005) HT and SBT coached students designing Carnot engines. Data from one human tutor were excluded.
−0.59	(Evens & Michael, 2006, Tables 18.7 and 10.4) HT coached medical students doing a cardiophysiology exercise. SBT (CIRCSIM) presented text when students made errors.
0.03	(Reif & Scott, 1999; Scott & Reif, 1999) SBT (PAL) had students alternate between problem solving and checking an example's steps for errors. HT coached students doing problem solving.
0.21	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). Exp. = Experiment; HT = human tutoring; SBT = step-based (computer) tutoring.

TABLE A9
Human Tutoring Versus Substep-Based Tutoring

d	*Citation and Comments*
0.09	(VanLehn et al., 2007, Exp. 1) SSBT (Why2-AutoTutor) and HT coached qualitative physics explanations.
−0.17	(VanLehn et al., 2007, Exp. 1) Same as above, but SSBT was Why2-Atlas.
−0.07	(VanLehn et al., 2007, Exp. 5) SSBT (Why2-Atlas) and HT coached qualitative physics explanations to students who had not taken college physics.
0.25	(VanLehn et al., 2007, Exp. 5) Same as above, but SSBT was Why2-AutoTutor.
−0.71*	(Evens & Michael, 2006, Tables 18.7 and 10.4) SSBT (CIRCSIM-Tutor) and HT coached solution of a cardiophysiology problem.
−0.12	Mean effect size

Note. Effect sizes are bold and followed by an asterisk if the comparison was statistically reliable ($p < .05$). Exp. = Experiment; HT = human tutoring; SSBT = substep-based (computer) tutoring.

TABLE A10
Expert or Highly Interactive Human Tutoring Compared to Novice or Less Interactive Human Tutoring

d	*Citation and Comments*
0.04	(Fossati, 2008) Expert versus novice tutors.
NS	(Chae et al., 2005) Expert versus novice tutors.
0.11	(H. Johnson & Johnson, 1992) Socratic versus Didactic tutoring by the same people.
1.00	(Rose, Moore, VanLehn, & Allbritton, 2001) Socratic versus Didactic tutoring by the same person.
3.01	(Evens & Michael, 2006, Table 10.6) Expert versus novice tutors.
NS	(M. T. H. Chi et al., 2001) Compared untrained tutors (who tended to lecture) to the same tutors constrained to only prompt the students.
S	(di Eugenio et al., 2006) An expert tutor was more effective than two less expert tutors, but he appeared to teach different content than them. The novice tutors, by the way, were no more effective than several Step-based Tutors.

Note. Reliable differences are bold. S and NS mean significant and nonsignificant, respectively, but effect size cannot be computed.