Expert Versus novice Human tutors

Abstract

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 givi

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
many of the studies of human tutoring to show

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

(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

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