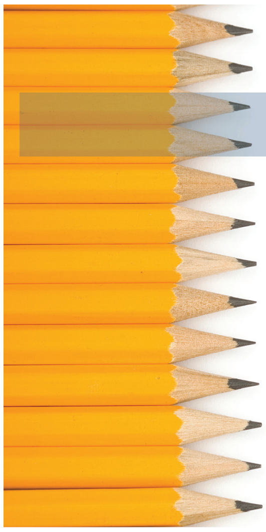

Editor: **Judy Kay,** University of Sydney, judy.kay@sydney.edu.au

Artificial Intelligence as an Effective Classroom Assistant

Benedict du Boulay, *University of Sussex*

The field of artificial intelligence in education (AIED) has been in existence for about 40 years and operated under various other names, the most common of which is intelligent tutoring systems (ITSs). The field was initially brought to wider attention by papers in a special issue of the *International Journal of Man-Machine Studies*,[1] in a book based on that special issue,[2] and in AI books of the era.[3] It continues to use techniques from AI and cognitive science to attempt to understand the nature of learning and teaching and to build systems to assist learners to master new skills or understand new concepts in ways that mimic the actions of a skilled human tutor working one-on-one with the learner. That is, such systems attempt to adapt the way they teach to the learner's knowledge, skill, and preferred ways of learning, and to consider the learners' affective trajectory as they deal with the expected setbacks and impasses of mastering new material. There is clearly some overlap with other uses of computing technology in education, although the commitment to individual adaptation through modeling different parts of the educational process is key.

For such systems to adapt to the learner and provide a personalized experience, a typical conceptual architecture has evolved. This consists of

- a model of the domain being learned, so that the system can reason about and judge whether a student's answer or a problem-solving step is appropriate;
- a model of the learner's current understanding or skill level, so that tasks of appropriate complexity can be posed;
- a model of pedagogy, so that the system can make sensible tutorial moves such as providing effective feedback or adjusting the nature of the next task; and
- one or more interfaces through which the system and the learner can communicate to explore and learn about the domain in question.

Over the years, many systems using various pedagogical techniques and topics have been built and evaluated. To illustrate the scope of the work, I discuss four diverse systems, which range from classic teaching in a formal subject and a procedural skill, to learning by creating externalized forms of knowledge for a highly conceptual learning task, to rich, natural user interaction via speech for learning complex, culture-laden skills.

The first AIED example is a system to help learners understand basic algebra by being given problems and provided with step-by-step feedback and guidance on their solution.[4] The second example is a system that helps learners gain a conceptual understanding of river ecosystems by building a concept map of that domain, as if for another learner, and having that simulated other learner take tests on the concept map's adequacy.[5] The third example is a system that helps military personnel learn and speak Arabic and understand the social and cultural norms needed to interact with people in the country in which they are operating.[6]

The fourth example illustrates the increasing importance of the interface in AIED systems and their use in informal learning environments, such as museums, as well as formal ones. Figure 1 shows Coach Mike, a pedagogical agent designed to help children visiting a museum learn about

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

Answer-, Step-, and Substep-Based Tutoring

Imagine that a student must solve the following equation:

$2(14 - x) = 23 + 3x$

An answer-based system would expect the student to do all the work offline and then provide the answer $x = 1$. If asked for a hint, the tutor can suggest broad ways of going about the problem, such as to collect all the terms in x on one side of the equation, but the tutor has no way of knowing that this advice is being followed. If the answer provided is wrong—for example, $x = 1.25$—the tutor might be able to hypothesize that the student multiplied out the bracket incorrectly, but if the answer provided is, say, $x = 14$, it probably will not offer much in the way of specific help.

In a step-based system, the student might be invited to multiply out the bracket expression as a first step, and thus will give $28 - 2x$ as the answer to that step. If a hint is requested or a wrong answer given to this step, then help can be given about how to work that step. Once the step is completed correctly, the tutor would invite an answer to the next step, such as reordering terms in the equation, and then on through further steps to the final answer.

In a substep-based system, there might be a remedial dialogue at a finer level than an individual step, for instance about what expressions such as $2x$ or $3x$ mean, if that seems warranted by the request for a hint or by a wrong step answer.

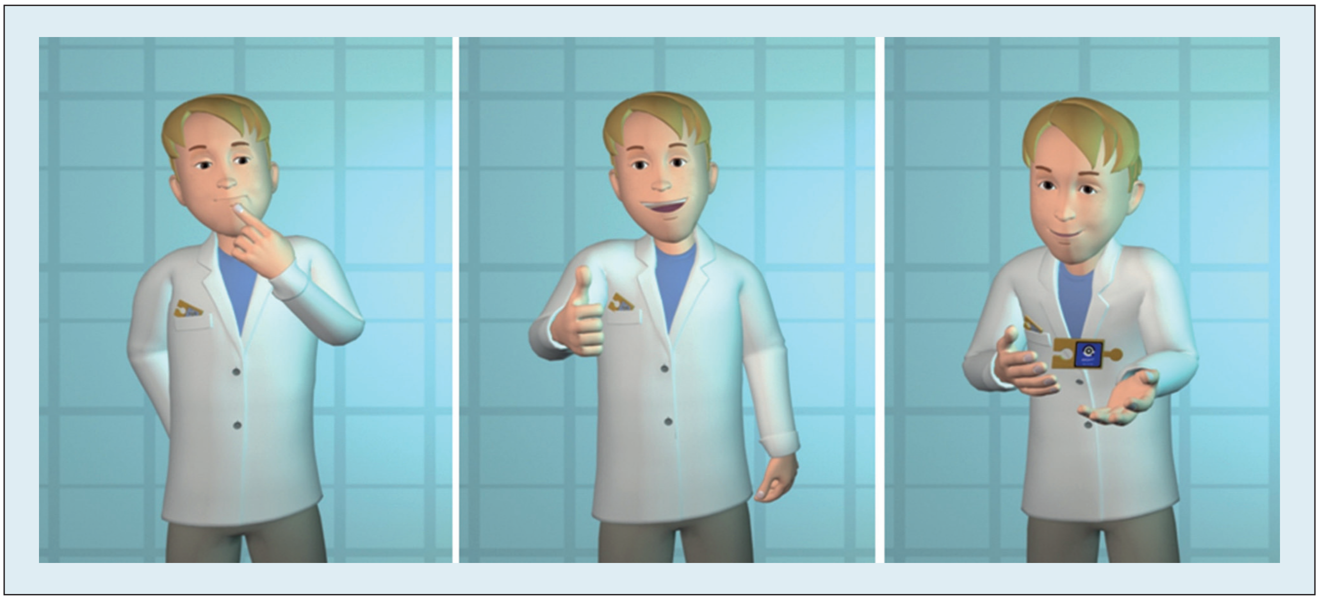

Figure 1. Coach Mike, a pedagogical agent, in three different poses.[7]

robotics. This kind of application extends the role of classroom teaching[7]:

> It means that such systems need to go beyond simply focusing on knowledge outcomes. They must take seriously goals such as convincing a visitor to engage, promoting curiosity and interest, and ensuring that a visitor has a positive learning experience. In other words, pedagogical agents for informal learning need to not only act as coach (or teacher), but also as advocate (or salesperson).

Coach Mike was designed to emulate some of the human museum curators' work, including helping orientate visitors, encouraging them to explore, and providing problem-solving challenges and support.

Some researchers have recently argued the benefits of AI systems in education,[8,9] whereas others have been more skeptical.[10] This column looks at the evidence derived from metareviews and meta-analyses conducted over the past five years. Its main focus is on the comparative effectiveness of AIED systems versus human tutoring. Note that a metareview of the use of pedagogical agents (not necessarily in AIED systems) "produced a small but significant effect on learning."[11]

This column is not intended as support for an argument about getting rid of human teachers, but rather as support for blended learning, in which some of the human teacher's work can be offloaded to AIED systems, as if to a classroom assistant.

Meta-Analysis and Metareviews

Since 2011, several metareviews and meta-analyses have attempted to determine the degree to which a whole host of systems have been educationally

Table 1. Effect sizes adapted from work by Kurt VanLehn.[12]

Comparison	No. studies	Mean effect size	Reliability (%)
Answer based vs. no tutoring*	165	0.31	40
Step based vs. no tutoring	28	0.76	68
Substep based vs. no tutoring	26	0.40	54
Human vs. no tutoring	10	0.79	80
Step based vs. answer based	2	0.40	50
Substep based vs. answer based	6	0.32	33
Human vs. answer based	1	−0.04	0
Substep based vs. step based	11	0.16	0
Human vs. step based	10	0.21	30
Human vs. substep based	5	−0.12	0

*Row 1 was taken by VanLehn from a separate study.[13]

effective. Typically, this has meant comparing them in terms of learning gains with other instructional methods, such as whole-class teaching by a human teacher or the use of a textbook without a teacher.

VanLehn's Meta-Analysis

Kurt VanLehn analyzed papers comparing five types of tutoring: no tutoring (for example, learning with just a textbook), answer-based tutoring, step-based tutoring, substep-based tutoring, and human tutoring.[12]

The difference between answer-based, step-based, and substep-based tutoring resides in the granularity of the interaction between the tutor and student (see the sidebar, "Answer-, Step-, and Substep-Based Tutoring"). Answer-based systems can provide hints and feedback only at the level of the overall answer. Step-based systems can provide hints, scaffolding, and feedback on every step that the student makes in the problem solving. By contrast, substep-based systems work at a finer granularity level still and "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."[12] AI techniques are required to underpin both step-based and substep-based tutors, whereas answer-based systems would typically fall under the heading of computer-based or computer-assisted instruction (CAI).

Given these granularity levels, VanLehn derived 10 pairwise comparisons of effect sizes (see Table 1). The rightmost column shows the proportion of the results for that row where the individual study comparison was statistically reliable at the level $p < 0.05$.

For the purposes of this review, the most interesting comparison is that between one-on-one human tutoring and step-based tutors (effect size = 0.21). By collating all the results in Table 1, VanLehn found that step-based tutors were, within the limitations of his review, "just as effective as adult, one-on-one tutoring for increasing learning gains in STEM topics."[12] He also found that although increasing the granularity of instruction from answer-based to step-based yielded significant gains, going to the finer level of substep-based tutoring did not add further value. (Note that this latter finding was based on a small number of studies only.)

Six Metareviews

Since VanLehn's meta-analysis, six metareviews have been published, as well as a large-scale study of a specific tutor (see Table 2). In the table, the "number of comparisons" column shows the number of instances for the given comparison in that row, not the total number of studies in the overall metareview.

In a metareview of 107 studies, Wenting Ma and colleagues found similar results to VanLehn for step-based ITSs both when compared to a no-tutoring condition (that is, just a textbook; mean effect size = 0.36) and, more positively than VanLehn, when compared to large-group instruction led by a human teacher (mean effect size = 0.44).[14] They found no differences when compared to small group human tutoring or one-on-one tutoring.

The same authors analyzed 22 systems for teaching programming and also found a "a significant advantage of ITS over teacher-led classroom instruction and non-ITS computer-based instruction."[15] A larger version of a similar study involving 280 studies is currently in progress.[20]

In a metareview of 50 studies involving 63 comparisons, James Kulik and J.D. Fletcher found comparable improvements (mean effect size = 0.65),[16] but they distinguished studies that used standardized tests from those where the tests were more specifically tuned to the system providing tuition, with smaller effect sizes when standardized tests were employed. Overall, they concluded that "this meta-analysis shows that ITSs can be very effective instructional tools … Developers of ITSs long ago set out to improve on the success of CAI tutoring and to match the success of human tutoring. Our results suggest that ITS developers have already met both of these goals."[16] They also found better results for substep-based systems than VanLehn, which they ascribed to differing comparison methodologies.

Much smaller effect sizes were found by Saiying Steenbergen-Hu and Harris Cooper in their meta-analysis of pupils using ITSs in school settings.[18] Kulik and Fletcher put this down to the weaker study inclusion criteria (for example, the inclusion of answer-based systems as if they were

Table 2. Six metareviews and a large-scale study.

Row number	Metareview	Comparison	No. comparisons	Mean effect size	Standard error
1	VanLehn[12]	Step based vs. one-on-one human tutoring	10	−0.21	0.19*†
2	Wenting Ma and colleagues[14]	Step based vs. one-on-one human tutoring	5	−0.11	0.10
3	Ma and colleagues[14]	Step based vs. "large group human instruction"	66	0.44	0.05
4	John Nesbit and colleagues[15]	Step based vs. "teacher led group instruction"	11	0.67	0.09
5	James Kulik and J.D. Fletcher[16]	Step based and substep based vs. "conventional classes"	63	0.65	0.07†
6	Saiying Steenbergen-Hu and Harris Cooper (2014)[17]	Step based vs. one-on-one human tutoring	3	−0.25	0.24
7	Steenbergen-Hu and Cooper (2014)[17]	Step based vs. "traditional classroom instruction"	16	0.37	0.07
8	Steenbergen-Hu and Cooper (2013)[18]	Step based and answer based vs. "traditional classroom instruction"	26	0.09	0.01
9	John Pane and colleagues[19]	Blended learning including a step-based system vs. traditional classroom instruction	147 schools	−0.1	0.10
				0.21	**0.10**
				0.01	0.11
				0.19	0.14
10	Weighted mean	AIED system vs. one-on-one human tutoring	18	−0.19	N/A
11	Weighted mean	AIED system vs. conventional classes	182	0.47	N/A

* The standard error in row 1 is based on all 10 studies, not just the 30% that produced reliable results (see Table 1).
† Standard errors computed by this article's author.

step-based systems) used by Steenbergen-Hu and Cooper, who also noted that lower achievers seemed to do worse with ITSs than did the broad spectrum of school pupils, although Kulik and Fletcher disputed this result.[16] However, in a parallel study of university students, Steenbergen-Hu and Cooper found more positive effect sizes (in the range of 0.32 to 0.37) for ITSs as compared to conventional instruction.[17] They conclude that ITSs "have demonstrated their ability to outperform many instructional methods or learning activities in facilitating college-level students' learning of a wide range of subjects, although they are not as effective as human tutors. ITSs appear to have a more pronounced effect on college-level learners than on K–12 students."[17]

Rows 10 and 11 of Table 2 summarize the results of the metareviews, excluding the evaluation of the Cognitive Algebra Tutor, and show a weighted mean effect size of 0.47 for AIED systems versus conventional classroom teaching. We use the term *AIED system* to cover all the systems—step-based, substep-based and answer-based—looked at in the metareviews. The comparison with one-on-one human tutoring shows that AIED systems do slightly worse, with a mean effect size of −0.19. In both cases, the means are weighted in terms of the number of comparisons in the metareview, not in terms of the original *N* values in the studies themselves.

Cognitive Tutors

The Cognitive Tutor family of tutors "are found in about 3,000 schools, and over a half million students use the courses each year."[21] They represent the most successful transition, in terms of numbers of students, of AIED work from the laboratory to the classroom. They provide scaffolded help with step-by-step problem-solving in various domains, mostly mathematical, and are designed to be used in a blended learning manner, thus freeing up the teacher to work with other children while some work with the tutors. Teachers are trained to make the best of these systems' arrival in their classrooms in terms of managing all the pupils in the classroom before, during, and after the use of the tutors.[4] Individual evaluations of various Cognitive Tutors are included in the reviews already described.

A large-scale US study of the Cognitive Algebra Tutor undertook a between-schools project involving 73 high schools and 74 middle schools across seven states.[19] The schools were matched in pairs: half received the Cognitive Algebra Tutor and adjusted their teaching to include it as they saw fit, whereas the others carried on with their regular method of teaching algebra. The study ran over two years and found no significant differences on post-test scores in the first year of the study, but in the second year, the high schools that used the Cognitive Tutor showed a small but significant effect size of 0.21 (see the bolded data in row 9 of Table 2).

Note that how the Cognitive Tutor was actually used in the classrooms was not controlled, although post hoc analyses showed that teachers did not generally use the Tutor exactly as recommended by its developers.

The overall conclusion of these metareviews and analyses is that AIED systems perform better than both CAI systems and human teachers working in large classes. They perform slightly worse than one-on-one human tutors. Most of the systems taught mathematics or STEM subjects, because these are the kinds of subject for which it is easier to build the domain and student models mentioned in the introduction. Note that there was a degree of overlap between these metareviews and analyses in terms of the collections of individual evaluations from which they have drawn their conclusions.

The specific study of the Cognitive Tutor for Algebra evaluated its use as a blended addition to the regular algebra teaching in the schools in which it was tried rather than as a total replacement for the teachers, and found good results in high schools as opposed to middle schools and in the second year of the evaluation as opposed to the first year. For various reasons, the way forward for AIED systems in the classroom must be the blended model—classroom assistants, if you like—in order to provide detailed one-on-one tutoring for some students while the human teacher attends to others, as well as having overall responsibility for all the students' progress.

Of course, good post-test results are not the only criteria for judging whether an educational technology will be or should be adopted.[10] However, the overall message of these evaluations is that blending AIED technology with other forms of teaching is beneficial, particularly for older pupils and college-level students studying STEM subjects. ■

Acknowledgments

This column is an adapted and enlarged version of a letter to the editor of the *International Journal of Artificial Intelligence in Education*.[22]

References

1. J.S. Brown, R.R. Burton, and A.G. Bell, "SOPHIE: A Step Towards a Reactive Learning Environment," *Int'l J. Man Machine Studies*, vol. 7, 1975, pp. 675–696.
2. T. O'Shea, "A Self-Improving Quadratic Tutor," *Intelligent Tutoring Systems*, D. Sleeman and J.S. Brown, eds., Academic Press, 1982.
3. J.S. Brown and R.R. Burton, "Multiple Representations of Knowledge for Tutorial Reasoning," *Representation and Understanding*, D.G. Bobrow and A. Collins, eds., Academic Press, 1975, pp. 311–349.
4. K.R. Koedinger et al., "Intelligent Tutoring Goes to School in the Big City," *Int'l J. Artificial Intelligence in Education*, vol. 8, no. 1, 1997, pp. 30–43.
5. K. Leelawong and G. Biswas, "Designing Learning by Teaching Agents: The Betty's Brain System," *Int'l J. Artificial Intelligence in Education*, vol. 18, no. 3, 2008, pp. 181–208.
6. W.L. Johnson, "Serious Use of a Serious Game for Language Learning," *Int'l J. Artificial Intelligence in Education*, vol. 20, no. 2, 2010, pp. 175–195.
7. H.C. Lane et al., "The Effects of a Pedagogical Agent for Informal Science Education on Learner Behaviors and Self-Efficacy," *Proc. 16th Int'l Conf. Artificial Intelligence in Education*, 2013, pp. 309–318.
8. R. Luckin et al., *Intelligence Unleashed: An Argument for AI in Education*, Pearson, 2016.
9. B.P. Woolf et al., "AI Grand Challenges for Education," *AI Magazine*, vol. 34, no. 4, 2014, pp. 66–84.
10. N. Enyedy, "Personalized Instruction: New Interest, Old Rhetoric, Limited Results, and the Need for a New Direction for Computer-Mediated Learning," policy brief, National Education Policy Center, 2014.
11. N.L. Schroeder, O.O. Adesope, and R.B. Gilbert, "How Effective Are Pedagogical Agents for Learning? A Meta-Analytic Review," *J. Educational Computing Research*, vol. 49, no. 1, 2013, pp. 1–39.
12. K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educational Psychologist*, vol. 46, no. 4, 2011, pp. 197–221.
13. C.-L.C. Kulik and J.A. Kulik, "Effectiveness of Computer-Based Instruction: An Updated Analysis," *Computers in Human Behavior*, vol. 7, nos. 1–2, 1991, pp. 75–94.
14. W. Ma et al., "Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis," *J. Educational Psychology*, vol. 106, no. 4, 2014, pp. 901–918.
15. J.C. Nesbit et al., "How Effective Are Intelligent Tutoring Systems in Computer Science Education?" *Proc. IEEE 14th Int'l Conf. Advanced Learning Technologies*, 2014, pp. 99–103.
16. J.A. Kulik and J.D. Fletcher, "Effectiveness of Intelligent Tutoring Systems: A

Meta-Analytic Review," *Rev. Educational Research*, vol. 86, no. 1, 2016, pp. 42–78.
17. S. Steenbergen-Hu and H. Cooper, "A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning," *J. Educational Psychology*, vol. 106, no. 2, 2014, pp. 331–347.S.
18. Steenbergen-Hu and H. Cooper, "A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on K–12 Students' Mathematical Learning," *J. Educational Psychology*, vol. 105, no. 4, 2013, pp. 970–987.
19. J.F. Pane et al., "Effectiveness of Cognitive Tutor Algebra I at Scale," *Educational Evaluation and Policy Analysis*, vol. 36, no. 2, 2014, pp. 127–144.
20. J.C. Nesbit et al., "Work in Progress: Intelligent Tutoring Systems in Computer Science and Software Engineering Education," 122nd Am. Soc. Eng. Education Ann. Conf., 2015, paper 11861.
21. K.R. Koedinger and V. Aleven, "An Interview Reflection on 'Intelligent Tutoring Goes to School in the Big City,'" *Int'l J. Artificial Intelligence in Education*, vol. 16, no. 1, 2016, pp. 13–24.
22. B. du Boulay, "Recent Meta-Reviews and Meta-Analyses of AIED Systems," *Int'l J. Artificial Intelligence in Education*, vol. 26, no. 1, 2016, pp. 536–537.

Benedict du Boulay is an emeritus professor of artificial intelligence and a member of the Human Centered Technology Research Group at the University of Sussex. Contact him at b.du-boulay@sussex.ac.uk.