Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions

Carole R. Beal, University of Arizona

Ivon M. Arroyo, University of Massachusetts, Amherst

Paul R. Cohen, University of Arizona

Beverly P. Woolf, University of Massachusetts, Amherst

Correspondence may be directed to:

Carole R. Beal Ph.D.

Cognitive Science Program, University of Arizona

Tucson AZ 85721 USA

520-626-1214

crbeal@email.arizona.edu

Abstract

Three studies were conducted with middle school students to evaluate a web-based intelligent tutoring system (ITS) for arithmetic and fractions. The studies involved pre and post test comparisons, as well as group comparisons to assess the impact of the ITS on students' math problem solving. Results indicated that students improved from pre to post test after working with the ITS, whereas students who simply repeated the tests showed no improvement. Students who had more sessions with the ITS improved more than those with less access to the software. Improvement was greatest for students with the weakest initial math skills, who were also most likely to use the multimedia help resources for learning that were integrated into the software.

Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions

In recent years, there has been increasing interest in the factors that lead K12 teachers to adopt or avoid technology-based learning activities in their classrooms (Barron, Kemker, Harmes & Kalaydjian, 2003; Belland, 2009; Duffield & Moore, 2006; Ronsisvalle & Watkins, 2005). One factor is the need to document that there are benefits for student learning (Cavanaugh, Gillan, Kromrey, Hess & Blomeyer, 2004; Dawson & Ferdig, 2006; Hew & Brush, 2007; Rice, 2009; Schrum, Thompson & Sprague, 2005). Although students often appear to be highly engaged with computer-based activities, one recent review of instructional software reported that there was little evidence that the programs actually led to higher student performance in math and reading (Dynarski et al., 2007). With growing emphasis on performance outcomes in K12 education, teachers are unlikely to integrate technology into their classrooms unless they can be confident that the activities will help students master the material.

One issue to consider is that there are different types of instructional software, ranging from relatively static electronic worksheets to interactive programs that are adaptive to the individual learner. In fact, the particular software packages that were considered in the Dynarksi et al. (2007) review tended to involve fairly traditional drill-and-practice activities. Students viewed problems and entered answers on the computer screen but in most cases the instruction did not really take advantage of the computer format in terms of interactivity, multimedia and the potential to adapt instruction for individual students. Thus, it is possible that greater benefits for learning might be observed with other types of instructional software.

One promising class of instructional software relies on artificial intelligence algorithms to customize instruction for individual students: “intelligent tutoring systems” (ITS) (Shute & Zapata-Rivera, 2007; Woolf, 2009). The design of intelligent tutoring systems reflects the

theoretical framework originally outlined by Vygotsky (1978), and subsequently extended by sociocultural theorists (Wood & Wood, 1996). A premise of this framework is that instruction should be guided by an ongoing assessment of the student's Zone of Proximal Development (ZPD) (Murray & Arroyo, 2002). The ZPD refers to the range in a student's performance that is defined by the kinds of problems that the student can solve without help, and those that he or she can solve with assistance from a teacher or tutor. The hypothesis is that students will learn most with problems that fall in his or her Zone of Proximal Development, and when assistance is provided in the form of hints, examples, or instruction. Brown referred to such tutorial assistance as "scaffolding" (Brown, Ellery & Campione, 1998).

Support for the ZPD framework is found in studies of expert human tutors (Bloom, 1984; Moschkovich, 2004; Lepper, Woolverton, Mumme & Gurtner, 1993). Skilled tutors have been found to select problems to alternate between those that the student can solve easily, and those that are more challenging. With challenging problems, skilled tutors guide the student to the solution by offering hints, posing questions and providing examples, helping the student discover the solution rather than simply telling the student the answer (Graesser, McNamara & VanLehn, 2005). Tutors also consider students' motivational states and use strategies to keep the student engaged, such as attributing the student's success to his or her own efforts (Lepper, Woolverton, Mumme & Gurtner, 1993). With this form of instruction, students have the opportunity to learn new skills, and to build confidence that they can succeed with challenging material.

The design of intelligent tutoring systems is drawn from the ZPD theoretical framework Like a human tutor, intelligent tutoring software is designed to choose problems that are predicted to advance the student's understanding, and then to update its estimate of the student's proficiency based on the student's performance. An essential component of intelligent tutoring

systems is the availability of instructional resources such as hints, multimedia examples, tutorial dialogues and other tools that the student can access during problem solving. With these resources, the student should be able to solve increasingly challenging problems in a particular area, eventually demonstrating mastery through independent problem solving.

Some research indicates that the instruction provided by intelligent tutoring systems can be highly effective with adult learners. Intelligent tutoring systems have been widely used for training in military, business and medical applications, with positive results (Woolf, 2009). Intelligent tutoring systems have also been deployed successfully with college learners in domains such as physics and chemistry. Examples include the AutoTutor system for physics, which uses tutorial dialogues to help students master the material. Students who were assigned to use AutoTutor had better conceptual understanding and test performance than peers who simply engaged in textbook reading (Graesser et al., 2005). Positive results were also reported for the Andes tutoring system for college physics (VanLehn et al., 2005). College students who used the Quantum Chemistry intelligent tutoring system as part of their coursework had significantly better course performance than peers who did not use the software (Walsh, Moss, Johnson, Holder & Madura, 2002). Interestingly, the benefits of the Quantum Chemistry tutor were most apparent for students who had not been performing well in their chemistry course.

Encouraging results have also been reported for intelligent tutoring systems for high school students in the area of mathematics learning. The "Cognitive Tutor", an ITS for algebra, has been shown to have positive effects on high school students' algebra achievement (Ritter, Anderson, Koedinger & Corbett, 2007). The ASSISTment system for high school algebra has also been found to help students improve their end-of-year test scores (Razzaq et al., 2005). Students who used Wayang Outpost, an ITS for high school geometry, showed significant

improvement from pre- to post-test, whereas there was no improvement for students who did not use the software (Beal, Walles, Arroyo & Woolf, 2007). As with the Quantum Chemistry tutor, both the Cognitive Tutor and Wayang Outpost systems appeared to have the greatest benefits for students who had the weakest initial skills.

Although prior work suggests that intelligent tutoring systems are effective with adult, college and high school learners, there have been relatively few studies of ITS software designed for use by younger students in classroom settings. In one study, fifth grade students were assigned to use the ASSISTment tutoring system for mathematics to complete 20 homework problems at home (Mendicino, Razzaq & Heffernan, 2009). The ASSISTment software provided hints and scaffolding in response to students' problem solving errors. The ASSISTment students performed better on a post-test than peers who completed their homework in traditional paper-and-pencil form, meaning that they did not receive immediate feedback and assistance on the problems. The results were encouraging, although somewhat limited by the relatively brief nature of the intervention. Thus, there is a need for additional research on the potential benefits of intelligent tutoring systems when used in the classroom with relatively young students.

The goal of the present project was to evaluate the impact of AnimalWatch, an ITS designed for students who are mastering basic computations and fractions skills. AnimalWatch also focuses uniquely on word problem solving, considered a core component of mathematics proficiency (Kintsch & Greeno, 1985; Koedinger & Nathan, 2004; National Council of Teachers of Mathematics, 2000). In AnimalWatch, students solve word problems with authentic information about various endangered species. The ITS uses problem solving errors to estimate the student's skill with each math topic, and selects problems that the student should be able to

solve by using the integrated help resources. When the student can solve challenging problems in one topic successfully, the system will move on to a new math topic. Because students do not all solve problems in the same way, students will not necessarily see the same problems as they work with AnimalWatch, or move at the same pace through the math topics.

An evaluation of the first version of AnimalWatch, a stand-alone application that was installed directly on computer labs in the schools, focused primarily on its impact on students' mathematics self concept. The hypothesis suggested by the ZPD theoretical framework was that adaptive selection of problems and the easy availability of hints should help to sustain students' confidence as they worked with challenging material. The initial results were encouraging, in that students showed significant increases in mathematics self concept after working with the system (Arroyo, 2000). However, the initial study did not include pre and post tests of the math skills targeted in the ITS, or comparison groups whose performance might help to evaluate its impact on math problem solving.

To address these limitations, three studies were conducted with an expanded version of the AnimalWatch ITS, now implemented as a web-accessible application. The studies involved pre and post test designs. In the first pilot study, we investigated the extent to which working with the ITS might be comparable to receiving instruction in a small group with a human math tutor, as suggested by the ZPD theoretical framework. In the second study, pre and post test performance was investigated for students who varied in the number of sessions that they had with the ITS. In the third study, ITS users were compared with other students who simply took the pre and post tests, to establish that repeating the tests alone was not associated with improvement.

Study 1

The goal of this pilot study was to learn if students would improve from pre to post test after working with the ITS software, and to assess if the effects would be roughly comparable to working with a human tutor.

Method

Participants

The participants were rising Grade 6 students enrolled in a summer academic skills class in Los Angeles, California, designed to help students prepare for middle school. Students were nominated for the program by their teachers on the basis of academic motivation and interest in science (i.e., the program was not remedial). The program included two hours of literacy activity and two hours of mathematics study per week. For the math study time, students were assigned to small groups (4 to 6 students) that worked entirely with a math tutor for all four sessions, or to groups that divided the math time between working with a tutor and working with AnimalWatch, i.e., one hour with the tutor and one hour with the ITS. Math tutors were middle and high school math teachers with at least five years of classroom teaching experience.

Pre and post tests

In this and the subsequent studies, there were two versions of the tests, with similar items of equivalent difficulty. Versions were counterbalanced across participants. Initial comparisons indicated that the tests were similar in difficulty and no effects were associated with test version, so the data were collapsed across this factor. The tests included problems that involved the skills tutored in the ITS: computation and fractions.

Procedure

In the first week of the program, students completed the pre test. Instructional sessions were held once per week for the next four weeks. In each session, the ITS students met with

their tutor for one hour and then walked to the computer lab where they worked with the ITS for 45 minutes (i.e., four hours with the tutor, and about three hours with the ITS, after time required to travel to the computer lab). Students in the comparison group worked for all four math sessions with their tutor (i.e., eight hours with the tutor). Comparison group activities included blackboard lessons and worksheet practice. Students worked in small groups, with assistance from their tutor. In the sixth week, students in both groups completed the post test.

Results and Discussion

Students received two scores representing the number of problems correctly solved on the pre test, and on the post test. Scores ranged from 0 to 30. Mean proportion correct scores for the students who completed both tests are shown in Table 1.

Table 1

Mean proportion correct scores for participants in Study 1, with standard deviations in parentheses

	Pre test	Post test
50% ITS 50% Small Group Tutoring N = 13	0.57 (.22)	0.73 (.17)
100% Small Group Tutoring N = 12	0.48 (.16)	0.70 (.12)

An analysis of variance (ANOVA) was conducted with Test (pre, post) as a repeated factor) and Group (ITS plus tutor, tutor) as a between subjects factor. There was a significant effect of Test, $F(1,23) = 27.816$, $p < .001$. No other effects were significant. This result indicates that both groups showed significant improvement after the program activity. More

specifically, students who worked for half of their math instructional time with the ITS improved as much as those who worked for the entire time with a human tutor.

Study 2

The results of the pilot study were encouraging with regard to the potential benefits of the software. However, conclusions were limited due to the small sample size of students who worked with AnimalWatch. In particular, it was not possible to compare the impact on students with stronger or weaker initial math skills. To address these limitations, a second study was conducted in which AnimalWatch was integrated into math classes in four schools located in the downtown area of Los Angeles. No comparison group was included, because the schools wanted all the students to use the ITS.

Method

Participants

The sample included 149 students who completed both the pre and post test, worked with the ITS for at least one session, and completed a minimum of 10 problems. The schools served highly diverse student populations, including many (46%) English Language Learners (ELLs). Students who were identified as English Learners participated in the activity, but the analyses reported here include only data from the English Primary students. (Results for the English Learners were previously reported elsewhere in the context of a separate investigation into the relation of math problem solving and ELLs' reading proficiency, Beal, Cohen & Adams, in press).

Procedure

In the first week of the study, students completed the pre test during their math instruction period. The following week, they began working with the ITS during math class, and

continued each week for four sessions, before completing the post test in the sixth week of the study. Each weekly session lasted about 40 minutes.

Although the study goal was to have students work with AnimalWatch for at least four sessions, students varied considerably in the time that they actually had with the ITS, due to conditions that are not uncommon in urban schools. In particular, lockdowns due to police activity interrupted the study schedule on several occasions (during lockdown, students, staff and researchers were confined to the classroom, meaning that the laptops could not be transferred to the next class scheduled to work with the ITS). In addition, high rates of student absence, transfers, new enrollments and attrition contributed to the variability in exposure to the ITS during the six week study. The final sample included 37 students who had one session with AnimalWatch, 39 students who had two sessions, 39 students with three sessions, and 34 students who completed all four sessions.

Results and Discussion

Each student received a score for the number of problems correctly answered on the pre test, and on the post test. Scores ranged from 0 to 30. A matched pairs t test indicates that students showed significant improvement from pre to post test, $t(147) = 2.74$, $p < .01$.

Prior research suggested that intelligent tutoring systems could be especially helpful for students with relatively weak skills (Beal et al., 2007; Ritter et al., 2007; Walsh et al., 2002). To investigate this possibility, the Study 2 participants were assigned into two groups: Those scoring above the mean on the pre test (N = 78) and those scoring below the mean (N = 71). Mean proportion correct scores for these two groups in the pre and post tests are shown in Table 2.

Table 2

Mean proportion correct scores for participants in Study 2, with standard deviations shown in parentheses

	Pre test	Post test
ITS Users Low Math Skill		
N = 71	0.15 (.09)	0.25 (.58)
ITS Users Higher Math Skill		
N = 78	0.52 (.15)	0.51 (.22)

A repeated measures ANOVA with Group as the between subjects factor and Test (pre, post) as the repeated factor revealed a main effect of Group, $F(1,145) = 189.738$, $p < .0001$. This effect indicates only that students who scored low on the pre test also scored low on the post test, relative to those with stronger initial skills. There was also an effect of Test, $F(1,146) = 20.049$, $p < .001$, indicating the post test scores were higher than the pre test scores. However, this effect was qualified by a significant interaction between Group and Test, $F(1,146) = 12.049$, $p < .001$. The interaction indicates that students who scored low on the pre test improved more on the post test than students with stronger initial math skills.

To explore whether the amount of time that different students had to work with the ITS influenced performance, a regression model was fit to predict post test scores with pre test scores and the number of ITS sessions that a student had as predictors. The overall model was significant, $F(2,146) = 59.105$, $p < .001$, $r^2 = 0.44$. Not surprisingly, pre test scores accounted for the largest amount of variance in predicting post test scores, $F(1,146) = 82.287$ $p < .001$. In addition, however, the number of AnimalWatch sessions was also a significant predictor of post

test performance, $F(1,146) = 5.488$, $p < .05$. This indicates that students who had more exposure to the ITS improved more than students who had fewer sessions.

Study 3

Although the results of the first two studies showed positive effects for students who worked with the ITS, conclusions were limited because there had been no comparison group in the second study. In particular, it was possible that students might show some improvement simply by taking a second version of the test, without any intervening instruction. Students can learn from one test and thus perform better on a second test that addresses the same skills. Therefore, in the third study a quasi experimental design was adopted which included two groups: students who worked with the ITS, and students who took the pre and post test without using the ITS in the intervening interval. In addition, the software instrumentation was improved in order to record students' requests to view the integrated help resources, and the time that they spent on problem solving. The goal was to try to link improvement on the outcome measure (post test) with students' actions while working with the software.

Method

Participants

The study included students in three Grade 6 classrooms. Two classrooms were randomly assigned to work with the ITS, and the third classroom was assigned to serve as a comparison group. The comparison group classroom was given the option to use the ITS after the study was completed, as an incentive to participate. The final sample included 37 students in the ITS group, and 23 students in the comparison group.

Procedure

The study was conducted during one school week. The tests were reduced from 30 to 16 items, to facilitate completion by all students in one class period. On the first day, students in both groups completed the pre test. The ITS classes then used the software during math class for the next three days. The comparison class continued with their regular math instruction. (The topics that were covered in the comparison class were not explicitly aligned with the ITS curriculum, because the goal was simply to learn if the comparison students would improve by taking the test a second time.) All students then completed the post test in the fifth session.

Results and Discussion

For each student, answers on the pre and post tests were scored as correct or incorrect. Correct answers were summed to yield scores ranging from 0 to 16. An ANOVA was conducted with Group (ITS, comparison) as a between subjects factor and Test (pre, post) as a within subjects factor. There were no significant effects.

Although there was no overall improvement for the ITS users, the results of Study 2 suggested that the ITS might be most helpful to students with relatively poor math skills. To learn if a similar effect might be observed again, the ITS students were divided into two groups: those who scored above and below the mean (M = 10.16 out of 16 possible points) on the pretest. Mean proportion correct scores are shown in Table 3.

Table 3

Mean proportion correct scores for Study 3 with standard deviations in parentheses

	Pre test	Post test
ITS Users Low Math Skills N = 7	0.45 (.16)	0.54 (.12)

ITS Users Higher Math Skills N = 25	0.76 (.07)	0.71 (.15)
Comparison Low Math Skills N = 16	0.47 (.11)	0.46 (.16)
Comparison Higher Math Skills N = 7	0.88 (.09)	0.86 (.10)

An ANOVA with Group (lower math skill, higher skill) as the between subjects factor and Test (pre, post) as the within subjects factor produced a main effect for Group, $F(1,35) = 38.822$, $p < .001$. There was also a main effect of Test, $F(1,35) = 5.697$, $p < .05$, indicating that post test scores were higher than pre test scores. This effect was qualified by a significant interaction between Group and Test, $F(1,35) = 8.907$, $p < .01$, indicating that the low math skill students who worked with the ITS showed significant improvement from pre to post test, whereas the higher math skill students did not. This is similar to the effect observed in Study 2.

A similar analysis on the comparison group students scoring above and below the pre test mean showed only a main effect of Group, $F(1,21) = 56.391$, $p < .001$. There was no effect of Test and no interaction between Group and Test. Thus, for the comparison group students, students with stronger skills on the pre test also scored higher on the post test than the students with weaker initial skills, but there was no improvement from test to test for either group.

Information about how the students worked with AnimalWatch is shown in Table 4. Students with relatively low pre test scores completed approximately the same number of problems overall as their peers with stronger math skills. However, the students with lower

initial math skills viewed more of the help resources and took about 12 seconds longer per problem than the students with better math skills. Thus, the students who improved the most were also the students who were mostly likely to use the resources for learning that were available in AnimalWatch.

Table 4

Behavior with ITS for Study 3 students grouped by math skill; comparison group students used the ITS after the post test

	Mean Number Problems Presented	Mean Problems Answered Correctly	Mean Unique Hint Resources Viewed	Mean Seconds per Problem
ITS Low Math Skills	52.7	37.8	6.71	79.1
ITS Higher Math Skills	53.5	31.1	4.00	63.8
Comparison Low Math Skills	72	24	5.6	46
Comparison Higher Math Skills	60	51	2.8	57

In the week after the study was completed, the students who had been in the comparison classroom were given the opportunity to work with the ITS. Their problem solving data are

included in Table 4. Two comparison group students are excluded: one viewed over 200 problems (whereas the average for other students in the same amount of time was 67 problems) and another viewed only 20 problems but took over 4 minutes for each one (whereas the other students averaged 49.84 seconds per problem). The pattern of results is similar to that observed for the original users in that the students with relatively poor math skills viewed more help resources than the higher skills students.

General Discussion

The research was designed was to learn if a web-based intelligent tutoring system for pre algebra could improve middle school students' problem solving skills when evaluated using pre and post test comparisons. Although students often appear to enjoy working with computers, it is important to look for evidence that the experience is associated with better performance on the skills targeted in the software. Three studies were conducted with middle school students who worked with the AnimalWatch ITS. All three studies included pre and post tests. In addition, all three studies included group comparisons to help interpret any effects associated with the ITS.

With regard to pre and post test comparisons, the results from the three studies indicated that students who worked with the ITS showed improvement in their math problem solving. More specifically, students whose skills were relatively poor, indicated by scores below the mean on the pre test, were most likely to improve on the post test (Study 2 and 3). It is possible that this might have been a statistical artifact, in that low scores tend to improve more than higher scores. However, the results of Study 3 argue against a simple "regression to the mean" explanation: Students who improved the most were also those who were most likely to look at the multimedia help resources integrated into the software. Also, students in the comparison

group in Study 3 who scored below the mean on the pre test did not show any improvement when they took the post test.

One interpretation of the results is that it is the students who are not doing particularly well in math who are most likely to take advantage of the instructional resources for learning that are available in the software, and to show significant test improvement as a result. Additional support for this interpretation is provided by the finding that students with weak math skills who were in the comparison group were most likely to look at the integrated help resources when they had the opportunity to work with the ITS. This result may not seem particularly surprising, yet educational interventions typically have most benefit for those students who are already doing well to begin with (Ceci & Papierno, 2005). Similar benefits for students with relatively weak skills have been reported for other intelligent tutoring systems, including the Cognitive Tutor for algebra and the Quantum Tutor for chemistry. These converging findings suggest that computer-based instruction may be an attractive alternative for students who are struggling with the material and who may be reluctant to ask for help in the traditional classroom context.

Although the results from the studies were generally encouraging, it should be noted that the effects associated with the ITS were not particularly strong. When students improved, the change was on the order of a few items more correct on the post test. The relatively weak effects may have been due to the limited scale of the intervention: at best, students worked for three or four sessions with the ITS and thus did not always reach the most challenging material. Support for this interpretation is found in Study 2, in which the effects were found to increase as students had more opportunity to work with the ITS. An important challenge for future research will be to evaluate the effects of software when integrated into classroom instruction over a more

sustained period. In addition, a delayed post test would help to assess whether the benefits are sustained over time.

On the positive side, the effects associated with the ITS intervention were similar in scale to those associated with being tutored in a small group by an experienced human math teacher (Study 1). However, it should be emphasized that none of the studies involved a direct comparison of the ITS and a human teacher. In the first study, the ITS students still received half of their instruction from an experienced math teacher. Also, in the third study, the students in the comparison group did not study the same math topics that were covered by the ITS. The goal was only to establish whether repeating the tests would be associated with improvement in scores, not to compare ITS and teacher-led instruction on the same material. Overall, the results suggest that ITS software may be a useful addition to K12 classroom instruction, particularly for struggling students.

References

Barron, A. E., Kemker, K., Harmes, C., & Kalaydjian, K. (2003). Large-scale research study on technology in K-12 schools: Technology integration as it relates to National Technology Standards. Journal of Research on Technology in Education, 35, 489-507.

Beal, C. R., Cohen, P. R., & Adams, N. (2009, in press). Reading proficiency and mathematics problem solving by high school English Language Learners. Urban Education, 44, in press.

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

Beck, J., Arroyo, I., Woolf, B. P., & Beal, C. R. (1999). An ablative evaluation. In Proceedings of the 9th International Conference on Artificial Intelligence in Education, pp. 611-613. Amsterdam: IOS Press.

Belland, B. (2009). Using the theory of habitus to move beyond the study of barriers to technology integration. Computers and Education, 52, 353-364.

Bloom, B. S. (1984). The two-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.

Brown, A. L., Ellery, S., & Campione, J. (1998). Creating Zones of Proximal Development electronically. In J. Greeno & S. Goldman (Eds.), Thinking practices: A symposium in mathematics and science education, pp. 341-368. Hillsdale NJ: Erlbaum.

Cavanaugh, C., Gillan, K.J., Kromrey, J., Hess, M., & Blomeyer, R. (2004). The effects of distance education on K-12 student outcomes: A meta-analysis. Retrieved June 19, 2009 from the North Central Regional Educational Laboratory Web site < http://www.ncrel.org/tech/distance/k12distance.pdf >.

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. American Psychologist, 60, 149-160.

Dawson, K., & Ferdig, R. (2006). Commentary: Expanding notions of acceptable research evidence in educational technology: A response to Schrum et al. Contemporary Issues in Technology and Teacher Education, 6 (Online Serial).

Duffield, J. A., & Moore, J. A. (2006). Lessons learned from PT3. TechTrends: Linking Research and Practice to Improve Learning, 50, 54-57.

Dynarsky, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, B., Murphy, R., Penuel, W., Javitz, H., Emery, D., & Sussex, W. (2007). Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor and iSTART. Educational Psychologist, 40, 225-234.

Hew, K. F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. Education Technology Research and Development, 55, 223-252.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. Psychological Review, 92, 109-129.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. Journal of the Learning Sciences, 13, 129-164.

Lepper, M. R., Woolverton, M., Mumme, D., & Gurtner, J. (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.

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, 331-350.

Moschkovich, J. (2004). Appropriating mathematical practices: A case study of learning to use and explore functions through interaction with a tutor. Educational Studies in Mathematics, 55, 49-80.

Murray, T., & Arroyo, I. (2002). Toward measuring and maintaining the Zone of Proximal Development in adaptive instructional systems. Proceedings of the 6th International Conference on Intelligent Tutoring Systems. In S.A. Cerri, G. Gouardères & F. Paraguaçu (Eds.), Lecture Notes in Computer Science 2363, pp. 133-145. Springer Berlin.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston VA: Author.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T. E., Upalekar. R, Walonoski, J. A., Macasek. M. A., & Rasmussen, K. P. (2005). The Assistment project: Blending assessment and assisting. In C. K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) Proceedings of the 12th International Conference on Artificial Intelligence In Education (pp. 555-562). Amsterdam: IOS Press.

Rice, K. (2009). Priorities in K-12 distance education: A Delphi study examining multiple perspectives on policy, practice and research. Education Technology and Society, 12, 163-177.

Ritter, S., Anderson, J. R., Koedinger, K., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. Psychonomic Bulletin and Review, 14, 249-255.

Ronsisvalle, T., & Watkins, R. (2005). Student success in online K-12 education. Quarterly Review of Distance Education, 6, 117-124.

Schrum, L., Thompson, A., & Sprague, D. (2005). Advancing the field: Considering acceptable evidence in technology research. Contemporary Issues in Technology and Teacher Education, 5 (Online Serial).

Shute, V., & Zapata-Rivera, D. (2007, March). Adaptive technologies. Research Report RR-7-05. Educational Testing Service. Retrieved Sept. 21, 2009 < http://www.ets.org/research/researcher/RR-07-05.html >.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychology processes. Cambridge MA: Harvard University press.

Walsh, M., Moss, C. M., Johnson, B. G., Holder, D. A., & Madura, J. D. (2002). Quantitative impact of a cognitive modeling intelligent tutoring system on student performance in balancing chemical equations. The Chemical Educator, 7, 379-383.

Wood, D., & Wood, H. (1996). Vygotsky, tutoring and learning. Oxford Review of Education, 22, 5-16.

Woolf, B. P. (2009). Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Burlington MA: Morgan Kaufman Publishers.

Acknowledgments

We would like to thank the teachers, principals and staff at the following programs and schools for their support: the University of Southern California's Neighborhood Academic Initiative; the Green Dot Public Schools in Los Angeles CA; and the Deerfield Middle School in Deerfield MA. We would also like to thank the following people for their contributions to the project: Joseph Beck, David Marshall, and David Hart at the University of Massachusetts Amherst; Erin Shaw, Jean-Philippe Steinmetz, Mike Birch and Teresa Dey at the University of Southern California; Wesley Kerr at the University of Arizona; and Niall Adams at Imperial College London. The AnimalWatch project has been supported by grants from the National Science Foundation (HRD 9555737, 9714757) and the Institute of Education Sciences (R305K0500086, R305K090197). The views expressed in this article are not necessarily those of the sponsoring agencies.

About the Authors

Carole R. Beal is Professor of Cognitive Science at the University of Arizona. Her work focuses on the design, deployment and evaluation of online instruction for K12 students. She can be contacted at < crbeal@email.arizona.edu >.

Ivon M. Arroyo is Senior Postdoctoral Fellow in the Computer Science Department at the University of Massachusetts Amherst. Her research is in the area of intelligent tutoring systems for math and science. She can be contacted at < ivon@cs.umass.edu >.

Paul R. Cohen is Professor and Head of the Computer Science Department at the University of Arizona. He directs research projects on intelligent systems, data mining and machine learning. His email address is < cohen@cs.arizona.edu >.

Beverly P. Woolf is Research Professon in the Computer Science Department at the University of Massachusetts Amherst. She directs research projects on distance education and tutoring systems for K12 and university students. Her email address is < bev@cs.umass.edu >.