SREE Spring 2013 Abstract Title Page

Not included in page count.

Title: Results of the efficacy study on the impact of the AnimalWatch System on middle school students' mathematics performance

Authors and Affiliations:
Shandy Hauk, WestEd
Bob Allen, WestEd
Katie D'Silva, WestEd
Weiling Li, WestEd
Cathy Ringstaff, WestEd

Abstract Body

Background / Context:

There is a clear need to ensure that middle school students are prepared to succeed in algebra with thorough proficiency in basic computation, fractions, ratios and proportions and other algebra readiness topics (Stein, Kaufman, Sherman & Hillen, 2011; Wimberly & Noeth, 2005). One possibility is the use of tutoring software for mathematics learning. Such software is designed with the goal of providing individualized instruction to the learner via the use of features such as integrated explanations, videos, and other multimedia resources. Using software gives students the opportunity to direct their own learning and may provide a motivational advantage. Software also provides integrated tracking (e.g., problems completed, number completed correctly and so on) that can inform both students and teachers about progress in real time. These features have offered promise for more effective learning (Bloom, 1984; Cohen et al., 1982; Graesser, Conley & Olney, 2011; VanLehn, 2011; Woolf, 2009).

Purpose / Objective / Research Question / Focus of Study:

Several studies suggest that there may be benefits to the use of tutoring software specifically for mathematics learning (Arroyo, Woolf, Royer, Tai & English, 2010; Barrow, Markman & Rouse, 2009; Beal, Arroyo, Cohen & Woolf, 2010; Ritter, Kulikowich, Lei, McGuire & Morgan, 2007). Although results from these studies are encouraging, conclusion are limited because they have typically involved fairly brief interventions and the use of study-specific tests as outcome measures. In contrast, research on tutoring software using more rigorous experimental methodologies and standardized outcome measures has not demonstrated a consistent benefit. A review of four widely-used mathematics tutoring systems conducted by the What Works Clearinghouse concluded that there was no compelling evidence of better student mathematics performance from the systems (Campuzano, Dynarski, Agodini & Rall, 2009). Thus, there is a need for systematic research on the mathematics learning outcomes supported by tutoring systems, and on how teacher implementation may influence effectiveness. The RCT study reported here was designed to address this need through a rigorous evaluation of the impact of one tutoring system for algebra readiness on students' mathematical proficiency: The AnimalWatch System (AWS). The system offers mathematics problems amid authentic environmental science material on tracking and monitoring the status of endangered species (hence the system's name). AWS mathematics content is algebra readiness topics such as number sense, computation, fractions, decimals, percentages, proportions, and rational numbers.

Setting:

Setting for the initial Practice Year and full-scale Study Year was California public schools, in 62 and 69 Grade 6 classrooms, respectively (half used AWS, half practiced business as usual). AWS software use was substituted for regular curriculum. Each teacher decided what to replace and when to do the substitution. An online AWS Teacher Handbook offered content alignment information for standards and guidelines for content substitution decision-making. AWS was used in computer labs or by way of computer carts (classroom sets of laptops distributed and collected for in-class use). These public school computers ran systems and software between 3 and 10 years old. AWS was designed to it would work on these systems.

Population / Participants / Subjects:
The study included two years of data collection (Practice Year and Study Year). Some teachers participated in both years. *Practice year*. The Grade 6 student sample (*N* = 1,473) consisted of 884 in the Treatment classes and 589 in the Control classes. About 20% were identified by their districts as English learners. In most districts, a letter sent home to caregivers notified them of study activities and offered an opt-out form to abstain from their child's scores being included in analysis; in one district, active caregiver consent was obtained for all participants.
Study Year. The student sample (*N* = 2,116) consisted of 1140 in the Treatment classes and 976 in the Control classes. About 20% were identified by their districts as English learners. Student and parent/caregiver consent worked as it had in the Practice Year.

Intervention / Program / Practice:
AWS offers practice with mathematizing and solving word problems, generally considered a central component of mathematics proficiency (Koedinger & Nathan, 2004). The system also offers Skill Builders as decontextualized computational practice. The word problems are grouped by mathematics topic into 16 Learning Objectives. A Learning Objective includes either 12 or 18 word problems organized into 6-item problem sets, each focusing on a different animal theme and mathematical concept. Each problem is presented with a relevant photograph of animals. Each includes two hints that are automatically delivered in response to an incorrect answer, along with a suite of multimedia help resources (e.g., short video lessons, worked examples or animated widgets demonstrating the particular operation involved in the problem). Students submit up to three answers for each problem, and have the option to click on "too hard" to skip a problem (the problem is counted as incorrect). All materials are aligned with the state mathematics content standards. A detailed online teacher handbook provides resources, lesson plans, sample problems and other materials to support implementation by teachers.

Research Design:
The study employed a school-level blocked randomized controlled trail (RCT) design. Teachers were recruited and signed participation consent forms before being randomly assigned to either the Treatment or Control condition. Treatment teachers participated in a two-hour online professional development module designed to introduce them to the software and its features. Teachers integrated AWS into their instruction throughout the school year. Software use replaced approximately 15% of mathematics instructional time across the school year. Assessments of student learning included state and standardized, nationally-used, middle school mathematics tests as well as a project-designed measure.

Data Collection and Analysis:
Data collected for the study included student performance on state and nationally validated measures, a project-designed proximal measure, student and school demographic information, teacher preparation and knowledge measures, surveys of enacted teacher practice, observations of low, medium, and high fidelity implementations, and teacher reflective logs on the nature of their experiences (feasibility and fidelity) and perceptions of efficacy.

Students' standardized mathematics test performance. Data on mathematics proficiency included scale scores on the end-of-year California state tests (mathematics and English/Language Arts) as well as pre- and post-test scores on the Mathematics Diagnostic Testing Project (MDTP) assessment. Complete data were not available for all students. Scores for the previous year's state test were available for 84% of the sample.

Study-specific Topic Quizzes. Researchers created a project-related assessment to allow a fine-grained look at students' mathematical performance. Each topic quiz had three types of items: two word problems like those in AWS (e.g., featuring the endangered Wild Dog), two word problems situated in a bakery context (an alternate "real-world" setting), and two abstract computational items (e.g., 30 is 50% of what number?). The topics tested were the same for each subset of questions. For example, there could be three questions about computing the area of a rectangle: one question had a picture of a rectangle and of Wild Dogs and the item asked about the area of a rectangular habitat, the bakery question involved finding the length of a new floor given the area of the remodeled bakery's floor.

Teacher surveys and logs. Teachers completed an extensive survey on their curricular practices along with short weekly online logs to document their usage and report concerns or comments about implementation. Near the end of the school year, classroom observations were also conducted with a subset of teachers who had exhibited high, medium or low fidelity to the implementation guidelines based on the audit trail of their class' use of the system and their self-reports in the logs.

Feasibility of Implementation and Fidelity of Implementation. The developer gave guidelines for what was considered faithful implementation. Researchers operationalized this in four ways: AWS audit-trail information on frequency and length of use, teacher self-reports in logs, classroom observations, and structured post-observation interviews with teachers.

Findings / Results:

Results will be presented for the two years of the evaluation. Year 1 findings included:

- *MDTP*. Gain scores representing the normalized change from pre- to post-test were created for each student, and then compared by condition. The results showed that students in the Treatment group had significantly greater score gains than their Control group peers (14 versus 10 percentage points).
- *State achievement test.* The analytic strategy focused on using the previous year's math test score as a predictor of math test scores at the end of the study year, along with condition (Treatment, Control). As expected, the prior year's math score was a strong predictor of the study year math score. In addition, there was a significant effect of condition, indicating that Treatment group students scored higher than Control students.
- *Study-specific math tests.* Gain scores representing the change from pre- to post-test were created for each student, and then compared by condition. Students completing a high number of Learning Objectives had a gain of almost 9 percentage points more than control, while students who used the system a moderate number of weeks (from 8 to 11 across 18 weeks) saw an increase in their gain of almost 5 percentage points more than

the control group. Among students qualified for free and reduced-price lunch, AWS provided a 3% larger gain (21.2% vs 24.4%) than for comparable control group students.

- *Software use as a predictor of improvement*. The number of software problems completed during the year was a significant predictor of improvement on MDTP and state test, independent of students' pre-test score. There was a significant relation between the number of problems on which the student viewed the help features, and improvement from pre to post-test on the MDTP, $F(1,571) = 18.434$, $p < .001$. In addition, the percentage of word problems on which students viewed the integrated help features in the software varied significantly across Treatment Group teachers. Teachers whose students used more of the software resources effectively were also more likely to have students who performed better than would have been expected based on their prior year's test score.

Year 2 findings are not yet available but data have been collected and will be analyzed in time for the March 2013 meeting.

Conclusions:
Initial results from the first year of this RCT study were modestly promising with regard to the potential of the AWS tutoring technology to support mathematics learning. There was a positive effect on test performance, with students who used the software during the first year showing more improvement than those who did not. The effect was small but consistent across tests of mathematics proficiency. Given that software use accounted for no more than 15% of instructional time, finding any impact on distal measures of mathematics was encouraging.

Not surprisingly, the impact of the software appeared to be linked to variations in implementation across classrooms. More specifically, students who completed more Learning Objectives and whose teachers encouraged use of the integrated instructional resources in the software showed greater improvement. That is, active components in the Treatment condition appeared to include certain aspects of effective implementation by teachers, not simply the presence of software in the classroom.

In addition to the presentation of empirical findings, the key theme of the proposed SREE presentation will be to compare results across the evaluation years, highlight any mixed results, and outline "lessons learned" from the comparison. More specifically, the Year 1 findings suggest that teacher implementation of particular aspects of the software may be critical to the effectiveness of an intervention. Comparisons with Year 2 will help to pinpoint whether teacher experience (i.e., becoming more familiar with the software after a year of practice) supports improved implementation, or if the generally positive findings from the first year reflect some initial enthusiasm for participation that might have faded by the second year of the study.

References

Arroyo, I., Woolf, B. P., Royer, J. M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Lecture notes in computer science, 6094/2010* (pp. 423–432).

Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy, 1,* 52–74.

Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning, 9,* 64–77.

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.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009*). Effectiveness of reading and mathematics software products: Findings from two student cohorts* (NCEE 2009-4041). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Graesser, A. C., Conley, M. W., & Olney, A. (2011). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *American Psychological Association handbook*. Washington, DC: American Psychological Association.

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.

OCED. (2010*). PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science*. DOI: 10.1787/19963777

Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. *Supporting Learning Flow through Integrative Technologies, 162,* 1320.

Stein, M. K., Kaufman, J. H., Sherman, M., & Hillen, A. F. (2011). Algebra: A challenge at the crossroads of policy and practice. *Review of Educational Research, 81,* 453–492.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P.C. & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81,* 4–28.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46,* 197–221.

Wimberly, G. L., & Noeth, R. J. (2005). *College readiness begins in middle school*. Iowa City, IA: ACT, 2005.

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