On-line Tutoring for Math Achievement Testing: A Controlled Evaluation

Carole R. Beal
University of Southern California

Rena Walles, Ivon Arroyo, and Beverly P. Woolf
University of Massachusetts-Amherst

Abstract

We report the results of a controlled evaluation of an interactive on-line tutoring system for high school math achievement test problem solving. High school students (N = 202) completed a math pre-test and were then assigned by teachers to receive interactive on-line multimedia tutoring or their regular classroom instruction. The on-line tutored students improved on the post-test, but the effect was limited to problems involving skills tutored in the on-line system (within-group control). Control group students showed no improvement. Students' use of interactive multimedia hints predicted pre- to post-test improvement, and benefits of tutoring were greatest for students with weakest initial math skills.

Introduction

Performance on high-stakes tests has become increasingly important in recent years with growing demands for accountability in education. Some improvement in academic achievement has been observed since the passage of the No Child Left Behind Act in 2002 (U.S. Dept. of Education, 2006). However, the overall performance of American students in math remains of particular concern (Gollub, Bertenthal, Labov, & Curtis, 2002). Students in the United States score well below their peers in comparable countries on international assessments of math proficiency (American Institutes for Research, 2005). In addition, group differences in math achievement persist. For example, female students still score 30-40 points lower on average than males on the SAT-Math exam, even though females receive higher grades in math classrooms (College Board, 2005; Willingham & Cole, 1997). These gender differences are paralleled by gaps associated with ethnic groups; for example, African-American and Hispanic students score less well on average than White and Asian-American students on high-stakes achievement tests (Byrnes, 2003; Martin, 2000; NAEP Mathematics Report Card, 2005).

Recent research has focused on technology-based learning systems for students' math and science learning (Carnegie Learning, 2002; Middleton & Murray, 1999; Nguyen & Kulm, 2005). Current interactive tutoring systems are designed within the theoretical framework based on the Zone of Proximal Development, specifically, that instruction that is individualized and responsive to the student's ongoing performance will be most effective (Brown et al., 1994). Such "intelligent" tutoring systems make instructional decisions using a pedagogical agent: a software component that tracks the student's estimated understanding against its model of the curriculum (Beck, Woolf, & Beal, 2000). The pedagogical agent selects individual problems that are predicted to develop specific skills as needed for the individual student, as well as problems that review and reinforce skills that are estimated to be relatively well-understood by the student. The pedagogical agent also selects scaffolding from the range of instructional

resources available for a specific topic. Such instructional resources may include text hints, dialogue with the tutoring agent, worked examples that require transfer to the current problem, and interactive multimedia modules that walk the student through the solution path to the current problem.

There has been a considerable amount of research showing that intelligent tutoring systems designed within this theoretical framework provide effective instruction. For example, extensive evaluations of the Cognitive Tutor, created at Carnegie Mellon University and used by thousands of students, show that students improve with use of the Intelligent Tutoring System ITS relative to traditional whole-class math instruction (Carnegie Learning, 2002). Evaluation studies of the Andes tutoring system for physics indicate that student learning is significantly improved relative to paper-and-pencil work (Van Lehn et al., 2005). The Auto Tutor system provides effective instruction for physics and introductory computer science, among other topics, through simulated dialogue with the student (D'Mello, Craig, Sullins, & Graesser, 2006).

By comparison, there has been relatively little work to assess whether interactive tutoring systems can also help students improve their performance on academic achievement test items, which may require novel problem solving strategies and approaches rather than the direct application of procedures learned in the classroom. For example, SAT-Math exam problems can often be solved with good problem representation, estimation, and imagery strategies that may not be introduced or emphasized in the math classroom (Byrnes & Takahira, 1993; Gallagher, 1992; Reuhkala, 2001; Willingham & Cole, 1997). Deubel (2001) found that many teachers were not convinced that existing commercial math teaching software would be of value in helping students prepare for high-stakes math assessments, suggesting a need for supplemental resources to help students learn to solve the types of items that are likely to be on math achievement tests.

The present study was conducted to evaluate an on-line tutoring system designed to provide students with multimedia instruction in solving high-stakes math problems (e.g., problems from the SAT-Math exam). In addition, we hoped to learn if improvement in problem solving could be attributed specifically to the multimedia instruction. An alternate possibility is that students' performance might improve simply as the result of a general halo effect of interacting with a computer, for example, by increasing students' attention to the material; prior work suggests that interaction alone may enhance learning, independent of the instructional content (Mayer & Chandler, 2001). In addition, students might improve simply from taking a test twice (i.e., the first test might help the student become familiar with the types of math items that are on achievement tests and thus to improve on the second test even without tutoring in specific problem solving strategies). Thus, one goal of the study was to learn if students who received on-line interactive tutoring would improve only on math problems that required the math skills specifically targeted by the tutoring system, or if they would show general improvement, possibly indicating that they simply were engaged and attentive due to the novelty of the interactive computer activity. In addition, a second group of students took the pre- and post-tests but did not receive on-line tutoring to check for learning-from-the-test effects.

A second goal of the present study was to evaluate the potential effectiveness of different forms of interactive scaffolding. Casey et al. (1997) note that challenging math problems can often be solved in a variety of ways. For example, students may use an algorithmic, textbook-like approach whereby they assign names to variables and create and solve equations. However, some problems, particularly items associated with math achievement tests, may also be solved with a more visual approach (e.g., using angle estimation, imagery, and visualization in order to

infer the most likely answer to a problem) (Byrnes & Takahira, 1993). Such strategies can be readily displayed through animations on the screen, suggesting that visually-oriented scaffolding might be especially helpful to students. On the other hand, there is growing evidence that although animations often enhance students' interest and attention, they do not necessarily improve learning outcomes, possibly due to increased cognitive load (Mayer, 2001). In the present study, we compared the effects of algorithmic or visually-oriented hints on learning.

Method

Participants

The participants were students in geometry classes at two high schools located in suburban areas in Western Massachusetts. Teachers at each school selected students from one class to participate in the control condition (N = 49); students in the other classes participated in the on-line tutoring condition (N = 153). Student populations included roughly equal proportions of White, African-American, and Latino/a students, with 80% qualifying for free lunch and other assistance.

Materials

Students worked with the Wayang Outpost web-based interactive tutoring system. Wayang Outpost was designed to provide individualized multimedia tutoring in how to solve SAT-Math problems involving geometry skills. (The system includes additional modules that were not used in the study.) Students viewed a series of math problems, each of which showed a figure, table, or other graphic; the problem or equation to be solved; and five answer options. Students could click on an answer option and receive feedback (correct, incorrect).

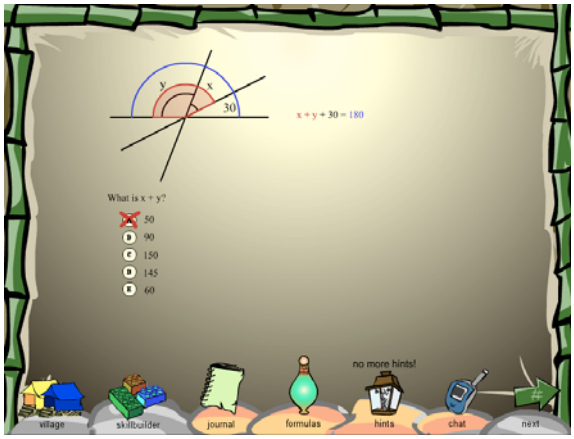

Figure 1. Screen shot of Wayang Outpost problem showing hint animation (in red) to indicate angle values need to find solution.

Students could also view a sequence of interactive hints leading to the solution for a problem by clicking the "help" icon. Each "help" click produced an additional step in the solution path, culminating in the answer. Students could view as many of the hints as they chose, or could answer the problem at any point. When students first logged into the system, they were randomly assigned by the tutoring system server to view either algorithmic or visual interactive hints. Figure 1 shows a screen shot of a problem requiring the student to find the value of a missing angle by summing the two known angles and subtracting from the degree value associated with a straight angle. In the example shown, the relevant angles are highlighted

with an animation in the visual interactive version before appearing on the screen in equation form. In the corresponding algorithmic version of the help, the values of the relevant angles are shown on the screen before moving into equation form.

There were 60 problems available at the time of the activity. For the present study, the order of problems presented to individual students was randomly determined by the tutoring system's problem selection mechanism (subject to the constraint that problems previously presented would not be selected again). Thus, students sitting at adjacent computers were unlikely to view the same math problem at the same time.

Pre- and post-tests of math problem solving. Paper-and-pencil tests of math problem solving proficiency were constructed from items taken from previously administered SAT-Math exams provided by the College Board. There were two forms, established in prior work to be equivalent in difficulty (Arroyo, Beal, Murray, Walles, & Woolf, 2004). Forms were counterbalanced for each student from pre- to post-test (e.g., one student received Form A for the pre-test, and Form B for the post-test; another student received Form B for the pre-test, and Form A for the post-test). Each form consisted of 21 problems: 15 geometry items assessed skills that were specifically tutored in the on-line system, and 6 algebra items assessed non-tutored skills, allowing for a within-subjects comparison of the system's impact. Problems were presented in multiple-choice format (i.e., there were five answer options for each item).

Procedure

On-line tutoring group. Students completed the paper-and-pencil pre-test of SAT-Math problem solving in their regular geometry class, proctored by their mathematics class teacher. They were given 30 minutes to work on the pre-test.

Students then worked with the on-line tutoring system for two class periods. Sessions were held in an Internet-equipped computer lab at the students' school, and scheduled during the regular mathematics class time. In the first session, students were provided with user names and passwords, logged into the system, and then directed to the tutoring module. The second ITS tutoring session took place the following day. Students were instructed to re-enter the SAT-Math tutoring module and to work on additional problems. They were allowed to work within the tutoring module until the end of the class period, or until they completed all 60 problems. Students worked with the tutoring module for approximately 50 minutes each day and completed an average of 56 problems.

The paper-and-pencil post-test was administered two days later in the regular classroom setting by the students' mathematics teachers. Students were given 30 minutes to complete the post-test. (Seventeen students were absent when the post-test was administered due to class schedule conflicts.)

Post-activity survey. After the post-test, students were asked to complete a brief paper-and-pencil survey about their perceptions of the tutoring system. There were four items focusing on how much students felt they had learned, how much they liked the system, how seriously they had tried to learn while using it, and how much they would like to use it again. Students rated their responses on a five-point Likert-type rating scale.

Control group students. Students in the control group were administered the paper-and-pencil pre- and post-tests in the same manner and on the same days as students in the experimental group. In the interim, the control group students participated in the normal mathematics class activities conducted by their teacher.

Scoring

Pre- and post-tests. The pre- and post-tests were scored for correct answers, incorrect answers, and skipped items. A scoring system similar to actual SAT-Math achievement test scoring was utilized to account for guessing: three points were given for each correct answer, one point was taken away for each incorrect answer, and 0.2 points was subtracted for each unanswered question (College Board, 2004). Each student received scores for their responses on the tutored (geometry) and untutored (algebra) items on the pre-test and on the post-test.

Interactions with tutoring system. As students worked in the tutoring module of the ITS, behavioral data were automatically recorded, including how many attempts were made to answer each problem, hints requested per problem, and time on the problem. Each student's action and latency data records on each problem were then machine-classified into one of five action patterns (Beal, Qu, & Lee, 2006). Table 1 shows the action patterns and definition rules used in the classifier. For example, if the student clicked on one or more incorrect answers in less than 10 seconds after the problem loaded on the computer screen, with inter-click intervals of less than 10 seconds, the student's record for that problem would be classified as GUESS. The latencies used in the classifier were determined by the performance of academically proficient students (i.e., if a high-achieving student requires more than 10 seconds to view a problem before responding with the correct answer, there is a high probability that students who choose an answer in less than 10 seconds have not actually read the problem and are guessing; the estimate of guessing increases with rapid clicks on incorrect answers).

Table 1: *Rules for Machine-classification of Student Actions and Latencies on Problem*

Action Pattern Classification	Definition
SOLVE Independent-accurate problem solving	Problem available for at least 10 seconds before student chooses correct answer; no interactive help is viewed.
SOLVE-ERRORS Independent-inaccurate problem solving	Problem available for 10+ seconds before student selects answer; first answer incorrect; at least 10 seconds before next answer selected; no interactive help viewed
LEARN Learn with help	Problem available for 10+ seconds before first action; interaction with at least one multimedia hint for 10+ seconds before correct answer selected
GUESS Select multiple answers without attending to problem or viewing help	Problem presented for under 10 seconds before answer selected; inter-click intervals on answers less than 10 seconds; no interactive help requested
SKIP Skip	Student does not select answer to current problem; requests new problem

Results

The overall scores of students in the control and on-line tutoring groups on the paper-and-pencil pretest were compared with a one-way analysis of variance. Results indicated that the control group students had significantly higher scores on the pre-test, $F(1,191) = 17.665$, $p <$

.001. Although the performance of the control group students was not at ceiling, these students were clearly more proficient than those selected by teachers to participate in the on-line tutoring group; thus, subsequent analyses were conducted separately for the two groups. (Results and interpretations are similar for analyses conducted with both groups included.)

To learn if the control group students improved from pre- to post-test, an analysis of variance was conducted with test time (pre-, post-) as the within subjects factor, and test score as the dependent measure. The results indicated that there was no significant difference in the scores of the control group students on the first and second test. Mean scores and standard deviations are shown in Table 2.

Table 2: *Mean scores (standard deviations in parentheses)*

Student Group	Number of students	Geometry Pre-test	Geometry Post-test	Algebra Pre-test	Algebra Post-test
Interactive tutoring	N = 153	0.25 (7.97)	3.31 (9.35)	-0.23 (4.76)	0.57 (4.81)
Control group	N = 49	6.93 (9.16)	5.85 (11.47)	0.79 (5.37)	2.29 (5.63)

The comparison of pre- to post-test scores was repeated for students in the on-line tutoring group. The results indicated that these students showed significant overall improvement from pre-test to post-test ($M = 4.13$), $F(1,125) = 12.977$, $p < .001$. More specifically, an analysis of variance with problem type (geometry, algebra) and test (pre-, post) as within-subjects factors yielded a significant interaction between problem type and test, $F(1,126) = 6.817$, $p < .01$. Students showed improvement on the math problems involving skills tutored in the on-line system (geometry), but not on the math problems involving non-tutored skills (algebra). Mean scores and standard deviations are in Table 2.

We next considered the effects of on-line tutoring on students in relation to their prior math skills, as indicated by their performance on the pre-test. Students were divided into "high" and "low" proficiency groups based on their pre-test scores, using a median split technique. An analysis of variance was conducted with math proficiency as the grouping factor, test time (pre-, post-) as a within subjects factor, and scores on tutored-skill test items as the dependent measure. Not surprisingly, there was a main effect of proficiency, $F(1,125) = 51.33$, $p < .001$, indicating that high proficiency students solved more problems than low proficiency students. There was also an effect of test time, $F(1,125) = 46.40$, $p < .001$; as noted above, students improved on tutored-skill items from pre- to post-test. In addition, there was a significant interaction between initial math proficiency and test time, $F(1,125) = 30.697$, $p < .001$. Students with low initial math proficiency showed greater improvement from pre- to post-test than students who started the activity with greater math proficiency. This indicates that the benefits of interactive on-line tutoring were greatest for the students with relatively weak math skills.

We next attempted to relate students' behavior with the tutoring system to their pre- and post-test performance. Recall that each student's interactions with the system on each math problem were classified into one of five patterns. Proportion scores were calculated for each pattern, in relation to the number of problems completed by each student. Mean proportion

scores were 0.26 for LEARN, 0.23 for GUESS, 0.21 for SOLVE-ERRORS, 0.17 for SOLVE, and 0.09 for SKIP; remaining problems were not classified by the rules shown in Table 1 (0.04).

Students were grouped via K-means clustering on their proportion scores for GUESS, LEARN, SOLVE, and SOLVE-ERRORS (because rates for SKIP were low, these items were not considered further). The results indicated five clusters; coordinate plots are shown in Figure 2. Cluster 1 (N = 11) included students whose dominant strategy was to solve the problems independently, without errors or requesting multimedia hints. Cluster 2 (N = 22) included students who tended to guess. Cluster 3 (N = 16) students also guessed, but not as frequently as Cluster 2 students, and they also interacted with multimedia help features. Cluster 4 (N = 30) students had the highest proportion of attempting to learn through interaction with the multimedia help. Cluster 5 (N = 34) students attempted to solve the problems on their own (i.e., they did not typically interact with the help) but differed from Cluster 1 in that their initial answers were often wrong.

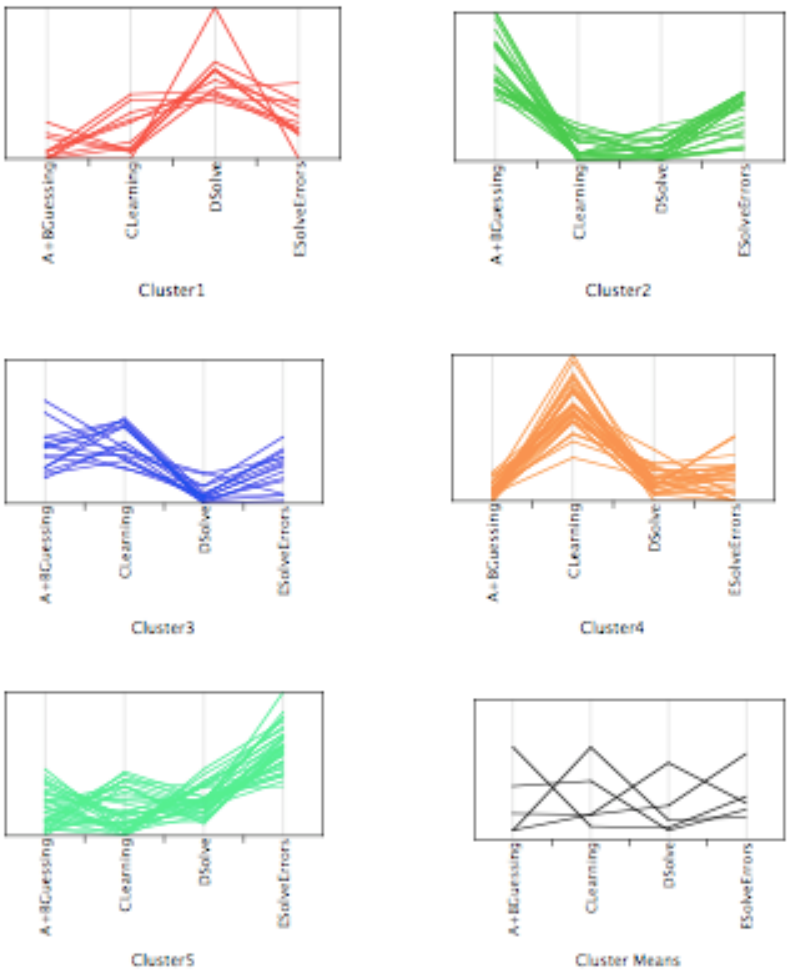

Figure 2. Coordinate plots showing action patterns associated with student clusters

An analysis of variance with Cluster as the grouping factor, test (pre-, post-) as the within-subjects factor, and scores on tutored test items as the dependent measure revealed a main effect of test, $F(1,108) = 8.499$, $p < .01$. This reflects the improvement from pre-test to post-test previously noted. In addition, there was a main effect of Cluster, $F(1,108) = 13.226$, $p < .001$.

As may be seen in Figure 3, Cluster 1 students had the highest pre-test scores, but showed the smallest improvement. These students were most likely to solve problems independently (i.e., they had the lowest level of interaction with the tutoring system). Students in the other clusters had lower pre-test scores, but tended to improve more. In particular, a contrast comparison showed that Cluster 4 students (who used multimedia help more than other students) improved more than Clusters 2, 3, and 5 students (who guessed or made errors), $F(1,108) = 7.501$, $p < .01$.

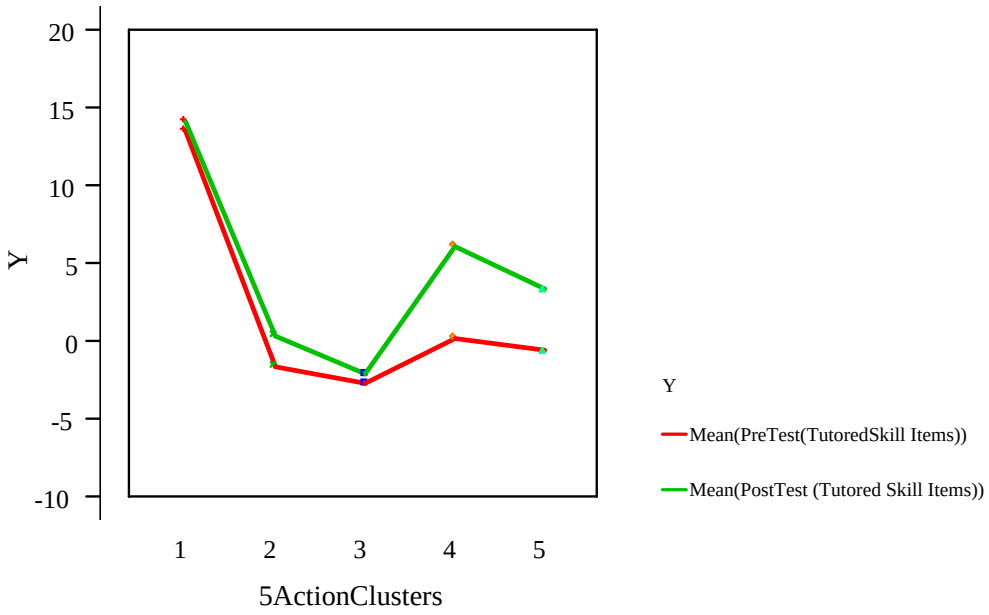

Figure 3. Mean pre-test (red line) and post-test (green line) score for student clusters representing interaction patterns with tutoring system.

Analyses were repeated with mode (algorithmic, visually-oriented multimedia help) as a between-subjects factor, but no significant effects were found. Similarly, gender did not appear as a significant factor in any of the analyses.

After the post-test, students completed a brief survey about their perceptions of the tutoring system. A MANOVA with Cluster as the grouping factor showed no significant differences in students' ratings of how seriously they had taken the activity, how much they felt they had learned, how much they liked the tutoring system, and how much they would like to use it again. However, as illustrated in Figure 4, Cluster 4 students – who were more engaged with the multimedia help features -- had the highest mean ratings in absolute terms on the four survey items.

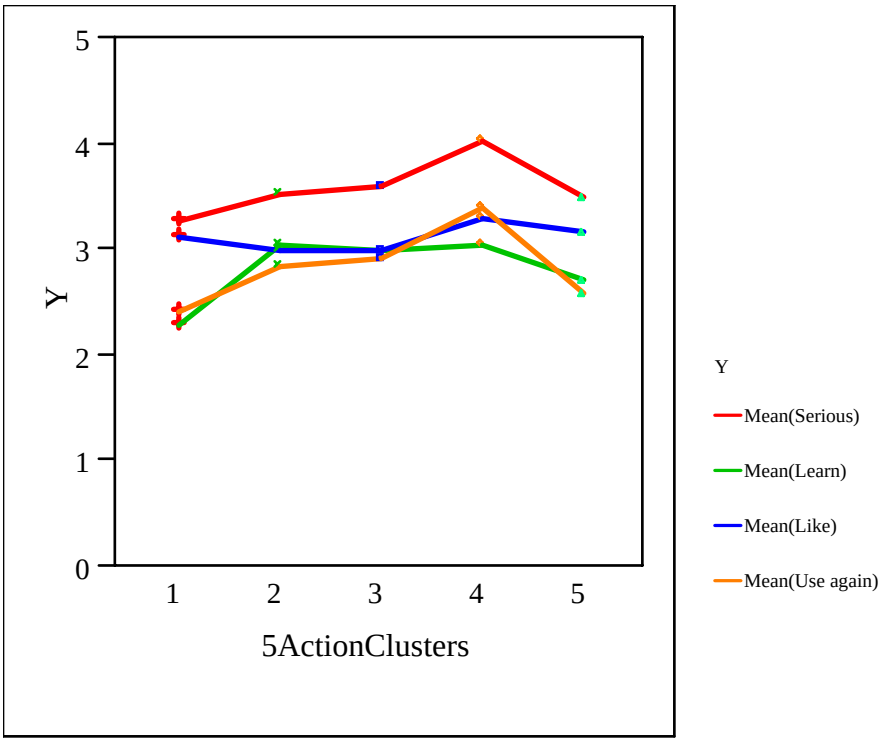

Figure 4. Mean ratings on post-activity survey for five action pattern clusters

Discussion

The present findings are consistent with previous work indicating that students benefit from interactive on-line tutoring in math (Carnegie Learning, 2002; Middleton & Murray, 1999; Nguyen & Kulm, 2005). Here, students showed significant improvement from pre- to post-test, in spite of high variability in performance. In addition, the effect did not appear to be due to a general halo effect of working with a computer: students who worked with the system significantly improved their performance on tutored (geometry) problems, without improving on control (algebra) questions.

In contrast, there was no change in the performance of the control group students. One issue is that the control group students generally performed much better at the pre-test than the tutoring system students. Selection of classes for the conditions was conducted by teachers who, it turned out, felt that classes with more low-achieving students would benefit from the on-line tutoring system and therefore selected these classes for participation. However, the important point is that the control group did not improve on either geometry or algebra problems. Thus, the improvement seen in the tutoring group was not simply due to general improvement over the school week, or the effect of re-testing.

One important issue in intervention studies is the role of time-on-task. In the present case, the tutoring and control groups participated for equal amounts of class time. However, the time was allocated to different activities: students in the control group continued to receive their regular classroom instruction, whereas for the tutoring group students, class instruction was replaced by interactive on-line tutoring. The interactive tutoring was aligned directly with the pre- and post-test outcome measures (i.e., it focused on teaching solutions to math problems from high-stakes achievement test items). In contrast, the classroom instruction provided to the

control group students did not specifically target this type of problem solving and thus, it may not be surprising that they did not show any improvement. What was not measured in the present study was whether there was a cost for the tutoring group students in terms of having regular class instruction replaced by the interactive on-line tutoring. Including an assessment of classroom learning in addition to the instrument used to assess high-stakes achievement test item problem solving would help to resolve this question.

A second goal of the research was to compare the impact of algorithmic help features to more visually-oriented animations. However, there was no indication that the type of multimedia help influenced either the students' performance (pre- and post-test changes), or their interactions with the tutoring system (as assessed by action pattern classifications). One possible reason is that the two versions of the hint sequence for a math problem only diverged after the first 2-3 hints were viewed. Students in the current study rarely asked for enough help to see these mode-specific hints. Therefore, it is at least possible that students would respond differently given more exposure to the specific types of help. However, additional work will be required to evaluate this possibility.

Another finding was that the on-line tutoring activity seemed to have the most benefit for students with the weakest math proficiency. More specifically, students with lower pre-test scores showed greater improvement than those with stronger initial skills. In addition, improvement was related directly to students' use of the multimedia help features: students who had the highest use of multimedia help features improved from pre- to post-test significantly more than other students. Conversely, students with higher pre-test scores were more likely to solve problems in the tutoring system without viewing help or making errors – yet these students did not show any improvement simply from solving problems. Thus, it seems that the students with weaker initial skills were most likely to engage in interaction with the tutoring system and, as a result, to improve their skills. These students also had the most positive perceptions of the tutoring system, as indicated by responses on the brief post-activity survey.

Our interpretation about the impact of interactive tutoring on low-proficiency students is somewhat constrained by the high variability in scores, and the low overall level of performance. The study was conducted in schools with generally low levels of academic achievement, and the students selected by teachers for participation in the tutoring activity were generally not doing as well as their peers in math. In absolute terms, even the "high proficiency" students in the study did not perform very well on our measures of math skill, and the improvement observed as the result of interactive tutoring, although significant in statistical terms (and to the classroom teachers), was hardly dramatic. Still, effective educational interventions usually have the greatest benefit for those students who were already doing well to begin with: the "rich get richer" effect (Ceci & Papierno, 2005). The present results thus suggest that interactive learning systems may have great potential to reach the students who are struggling the most in the traditional classroom.

The next step in the research is to learn what prompts some students to choose to use the interactive help features, whereas other students with similar skills decide to work independently or to guess. We did find that, not surprisingly, students with higher math skill were more likely to solve problems independently and correctly than students with lower skill. However, lower-skill students were equally likely to guess, to attempt to solve problems on their own (with errors), or to use multimedia help. More generally, the largest cluster included students who tended to keep trying to solve the math problems without viewing the help, with the result that they made many errors on problem after problem; this cluster included equal numbers of high

and low skill students. Thus, proficiency alone does not predict how students will interact with the system. One possibility is that students' beliefs about the domain, including their self-efficacy and their attributions about learning may play a role in their decision to interact with the system or to work independently (and unsuccessfully). For example, students may feel that accessing help features may somehow reflect poorly on their inherent ability, and may thus attempt to avoid seeking help even though it would be of benefit to them (Leder, Pehkonen, & Torner, 2002; Pajares, 2002). Assessing students' beliefs about their ability in relation to their behavior with interactive tutoring systems may lead to the design of interventions that will encourage students to use such systems more effectively.

References

American Institutes for Research. (2005). *Reassessing U. S. international mathematics performance: New Findings from the 2003 TIMSS and PISA*. Washington DC: American Institutes for Research.

Arroyo, I., Beal, C., Murray, T., Walles, R., & Woolf, B. (2004). Web-based multimedia tutoring for high stakes achievement tests. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Eds.). *Proceedings of the 7th International Conference on Intelligent Tutoring Systems: Lecture Notes in Computer Science 3220,* pp. 468-477. Berlin: Springer-Verlag.

Beal, C. R., Qu, L., & Lee, H. (2006, July). Classifying learner engagement through integration of multiple data sources. *Proceedings of the 21st National Conference on Artificial Intelligence*. Menlo Park CA: AAAI Press.

Brown, A. L., Ellery, S., & Campione, J. (1994). Creating Zones of Proximal Development electronically. In J. Greeno & S. Goldman (Eds.), *Thinking practices: A symposium in mathematics and science education*, pp. 341-468. Hillsdale NJ: Erlbaum.

Byrnes, J. P. (2003). Factors predictive of mathematics achievement in White, Black and Hispanic 12th graders. *Journal of Educational Psychology, 95*, 316-326.

Byrnes, J. P., & Takahira, S. (1993). Explaining gender differences on SAT-Math items. *Developmental Psychology, 29*, 805-810.

Carnegie Learning, Inc. (2002, May). Results from Moore OK. *Cognitive Tutor Research Report OK-01-01*. Pittsburgh PA.

Casey, M., Nuttall, R., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33*, 669-680.

Ceci, S. J., & Paierno, 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.

College Board. (2005). *SAT Math scores for 2005 highest on record*. New York: The College Board.

College Board. (2004). *How the test is scored*. Retrieved November 11, 2004, from http://www.collegeboard.com/student/testing/sat/scores/understanding/howscored.html

Deubel, P. (2001). The effectiveness of mathematics software for Ohio Proficiency Test preparation. *Journal of Research on Technology in Education, 33*, 5.

D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting affective states expressed through an emote-aloud procedure from Auto Tutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education, 16*, 3-28.

Gallagher, A. (1992). *Sex differences in problem solving used by high scoring examinees on the SAT-M*. College Board Report No. 92-2, ETS RR No. 92-33. New York: College Board Publications.

Gollub, J. P., Bertenthal, M. W., Labov, J. B., & Curtis, P. C. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U. S. high schools.* Washington DC: National Academy Press.

Leder, G. C., Pehkonen, E., & Torner, G. (2002). *Beliefs: A hidden variable in mathematics education.* Dordrecht: Kluwer.

Martin, D. B. (2000). *Mathematics success and failure among African-American youth*. Mahwah NJ: Erlbaum.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology, 93*, 390-397.

Middleton, B. M., & Murray, R. K. (1999). The impact of instructional technology on student academic achievement in reading and mathematics. *International Journal of Instructional Media, 26*, 109–116.

National Assessment of Educational Progress. (2005). *The Nation's Report Card: Mathematics.* Washington DC: National Center for Education Statistics.

Nguyen, D. M., & Kulm, G. (2005). Using web-based practice to enhance mathematics learning and achievement. *Journal of Interactive On-line Learning, 3*, 1-16.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 533-578.

Reuhkala, M. (2001). Mathematical skills in ninth-graders: Relationship with visuo-spatial abilities and working memory. *Educational Psychology, 21*, 387-399.

U.S. Dept. of Education. (2006). *No Child Left Behind is working*. Retrieved July 6, 2006, from http://www.ed.gov/nclb/overview/importance/nclbworking.html

Willingham, W., & Cole, N. (Eds.). (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Author note:

The Wayang Outpost web-based tutoring software used in this research may be accessed at < http://k12.usc.edu >. Development of Wayang Outpost was supported by grants from the National Science Foundation (HRD 0120809; REC 0411886). The conclusions described here do not necessarily represent those of the funding agency. We would like to thank the teachers, staff and students of the participating schools for their support of the research.