Web-based Intelligent Multimedia Tutoring for High Stakes Achievement Tests

Ivon Arroyo[1], Carole Beal[2,3], Tom Murray[1], Rena Walles[2],
Beverly P. Woolf[1]

[1] Computer Science Department, University of Massachusetts Amherst
{ivon, tmurray, bev}@cs.umass.edu
[2] Department of Psychology, University of Massachusetts Amherst
{cbeal, rwalles}@psych.umass.edu
[3] Information Sciences Institute, University of Southern California
cbeal@isi.edu

Abstract. We describe Wayang Outpost, a web-based ITS for the Math section of the Scholastic Aptitude Test (SAT). It has several distinctive features: help with multimedia animations and sound, problems embedded in narrative and fantasy contexts, alternative teaching strategies for students of different mental rotation abilities and memory retrieval speeds. Our work on adding intelligence for adaptivity is described. Evaluations prove that students learn with the tutor, but learning depends on the interaction of teaching strategies and cognitive abilities. A new adaptive tutor is being built based on evaluation results; surveys results and students' log files analyses.

1 Introduction

High stakes achievement tests have become increasingly important in the past years in the United States, and a student's performance on such tests can have a significant impact on his or her access to future educational opportunities. At the same time, concern is growing that the use of high stakes achievement tests, such as the Scholastic Aptitude Test (SAT)-Mathematics exam and others (e.g., the MCAS exam) simply exacerbates existing group differences, and puts female students and those from traditionally underrepresented minority groups at a disadvantage. Studies have shown that women generally perform less well than men on the SAT-M although their academic performances in college are similar (Wainer&Steiberg, 1992). Student's performance on SAT has a significant impact on students' access to future educational opportunities such as admission to universities and scholarships. New approaches are required to help all students perform to the best of their ability on high stakes tests.

Computer-based intelligent tutoring systems (ITS) provide one promising option for helping students prepare for high stakes achievement tests. Research on intelligent tutoring systems has clearly shown that users of tutoring software can make rapid progress and dramatically improve their performance in specific content areas. Evalua-

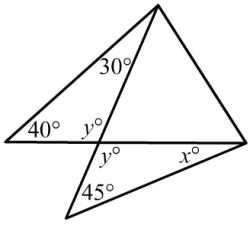
tion studies of ITS for school mathematics showed the benefits to student users in school settings (Arroyo, 2003).

This paper describes "Wayang Outpost", an Intelligent Tutoring System to prepare students for the mathematics section of the SAT, an exam taken by students at the end of high school in the United States. Wayang Outpost provides web-based access to tutoring on SAT-Math (http://wayang.cs.umass.edu). Wayang Outpost is an improvement over other tutoring systems in several ways. First, although they can provide effective instruction, few ITS have really taken advantage of the instructional possibilities of multimedia techniques in the help component, in terms of sound and animation. Second, this paper describes our work on incorporating intelligence to improve teaching effectiveness in various parts of the system: problem selection, hint selection and student engagement. Third, although current ITS model the student's knowledge on an ongoing basis to provide effective help, there have been only preliminary attempts to incorporate knowledge of student group characteristics (e.g., profile of cognitive skills, gender) into the tutor and to use this profile information to guide instruction (Shute, 1995; Arroyo et al., 2000). Wayang Outpost addresses factors that have been shown to cause females to score lower than males in these tests. It is suspected that cognitive abilities such as spatial abilities and math fact retrieval are important determinants of the score in these standardized tests. Math Fact retrieval is a measure of a student's proficiency with math facts, the probability that a student can rapidly retrieve an answer to a simple math operation from memory. In some studies, math fact retrieval was found to be an important source of gender differences in math problems (Royer et al., 1999). Other studies found that when mental rotation ability was statistically adjusted for, the significant gender difference in SAT-M disappeared (Casey et al, 1995).

2 System description

Wayang Outpost was designed as a supplement to high school geometry courses. Its orientation is to help students learn to solve math word problems typical of those on high stakes achievement tests, which may require the novel application of skills to tackle unfamiliar problems. Wayang Outpost provides web-based instruction. The student begins a session by logging into the site and receiving a problem. The setting is an animated classroom based in a research station in Borneo, which provides rich real world content for mathematical problems. Each math problem (a battery of SAT-Math problems provided by the College Board) is presented as a flash movie, with decisions about problem and hint selection made on the server (the tutor's "brain"). If the student answers incorrectly, or requests help, step-by-step guidance is provided in the form of Flash animations with audio (see figure 1). The explanations and hints provided in Wayang Outpost therefore resemble what a human teacher might provide when explaining a solution to a student, e.g., by drawing, pointing, highlighting critical parts of geometry figures, and talking, in contrast to previous ITS that relied heavily on static text and images.

Cognitive skills assessment. Past research suggests that the assessment of cognitive skills is relevant to selecting teaching strategies or external representations that yield best learning results. For instance, a study of students' level of cognitive development in AnimalWatch suggested that hints that use concrete materials in the explanations yield higher learning than those which explain the solution with numerical procedures for students at early cognitive development stages (Arroyo et al., 2000). Thus, Wayang Outpost also functions as a research test bed to investigate the interaction of gender and cognitive skills in mathematics problem solving, and in selecting the best pedagogical approach.□The site includes integrated on-line assessments of component cognitive skills known to correlate with mathematics achievement, including an assessment of the student's proficiency with math facts, indicating the degree of fluency (accuracy and speed) of arithmetic computation (Royer et al., 1999), and spatial ability, as indicated by performance on an standard assessment of mental rotation skill (Vandenberg et al., 1978). Both tests have captured gender differences in the past.



In the figure above, what is the value of x?

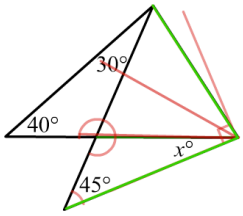
(A) 65
(B) 45
(C) 40
(D) 30
(E) 25

x + 45 + y = 180

40 + 30 + y = 180
70 + y = 180
y = 180 - 70
y = 110

x + 45 + y = 180
x + 45 + 110 = 180
x + 155 = 180
x = 180 -155
x = 25

Choose (E)



In the figure above, what is the value of x?

(A) 65
(B) 45
(C) 40
(D) 30
(E) 25

How are the rest of the angles related to x°?

x is about a third of the green angle

The green angle is a bit less than 90 degrees

x is a bit less than 90/3
x is a bit less than 30
Choose (E) for an answer

Figure 1. The computational (top) and visual (bottom) teaching strategies

Help in Wayang Outpost. Each geometry problem in Wayang is linked to two alternative types of hints, following different strategies to solving the problem: one strategy provides a computational and numeric approach and the second provides spatial transformations and visual estimations, generally encompassing a spatial "trick" that makes the problem simpler to solve. An example is shown in Figure 1. The choice of hint type should be customized for individual students on the basis of their cognitive profile, to help them develop strategies and approaches that may be more effective for particular problems. For example, students who score low on the spatial ability assessment might receive a high proportion of hints that emphasize mental rotation and estimation, approaches that students of poor spatial ability may not apply even though they are generally more effective in a timed testing situation. This is a major hypothesis we have evaluated, and the findings are described in the evaluation section.

Adventures: fantasy component. Wayang Outpost includes measures of transfer via performance on challenging multi-step math problems integrated into virtual adventures. Animated characters based on real female scientists (who serve as science, technology, engineering and mathematics role models) lead the virtual adventures. Thus the fantasy component is female-friendly and uses female role models. For example, the character based on Anne Russon (orangutan researcher, University of Toronto) takes the student across the rainforest to rescue orangutans trapped in a fire. Within the fantasy adventure, students are provided hints and shown SAT problems that are similar to the problem being solved within the adventure. The Lori Perkins character (Zoo Atlanta, Georgia) leads the "illegal logging" adventure involving the over-harvesting of rainforest teakwood, leading to flooding and loss of orangutan habitat. Students are asked to calculate a variety of items: discrepancies between the observed and permitted areas of harvest; orangutan habitat area lost to the resulting floods; perimeter distances required to detour around flooded areas; and how far to travel to reach areas with emergency cell phone access using cone models of satellite coverage.

3 Intelligence for adaptive tutoring

As the student works through a problem, performance data (e.g., latency, answer choice, hints requested) are stored in a centralized database. This raw data about student interactions with the system feed all our intelligent modules, to select problems at the appropriate level of challenge, to chooses hints that will be helpful for the student, to detect negative attitudes towards help and the tutoring system in general.

Major difficulties in building a student model for standardized testing include the fact that we start without a clear idea of either problem difficulty or which skills should be taught. Skills are sparse across problems, so there is a high degree of uncertainty in the estimation of students' knowledge. This is different from the design of most other tutoring systems: generally, the ITS designer knows the topics to be

taught, and then needs to create the content and pedagogy. In the case of standardized testing, the content is given, without a clear indication of the underlying skills. The only clear goal is to have students improve their achievement in these types of problems. Despite clear indicators of learning have been observed, a more effective Wayang Outpost is being built by adapting the tutor's decisions in various parts of the system. We are adding artificial intelligence for adaptivity in the following tutoring decisions:

Problem selection. Problems in Wayang are expensive to build, as the help is sophisticated (using animations and sound), and each problem is extremely different from each other, thus making it hard to show a problem more than twice with different arguments, without having students get the impression that it is "the same problem again". The result is that we cannot afford the construction of hundreds or thousands of problems, so that certain problems can be used and others discarded. Because Wayang Outpost currently contains 70 distinct problems, the reality is that a sophisticated algorithm that uses skill mastery levels to determine the appropriate skills that a problem should contain is not necessary at this stage. However, we believe some form of intelligent problem selection would be beneficial. We have thus implemented an algorithm to optimize word problem "ordering", a pedagogical agent whose goal is to show a problem where the student will behave slightly worse than the average behavior expected for the problem (in terms of mistakes made and hints seen). Expected values of behavior at a problem computed from log files from prior users of the system (which used random problem selection). The agent keeps a "desired problem difficulty" factor for the next problem. The next problem selected is the one that has the closest difficulty to the desired difficulty, which changes after every solved problem: when the student behaves better than what is expected for the problem (based on log files' data of past users), the "desired problem difficulty" factor increases. Otherwise, it decreases, and thus the next problem will be easier.

Level of information in hints. When the student seeks for help, a hint explains a step in the solution. Sequences of hints explain the full solution to the problem when students keep clicking for help. However, hints have been designed to be "skipped", in that each hint contains a summary of the previous steps. Thus, skipping a hint implies providing minimal information about the step (e.g. if a student clicks for help and the first hint is skipped, the second hint shown will provide a short static summary of the first step and the full explanation for the second step in the solution using multimedia). Martin&Arroyo (2004) present the results of experiments with simulated students, which showed how a Reinforcement Learning agent can learn how to "skip" hints that don't seem useful. A more efficient Wayang Outpost will be built by providing only those hints that seem "useful". The agent learns the usefulness of hints by rewarding highly those hints that lead the student to an answer and punishing those hints that lead to incorrect answers or make the students ask for more help.

Attitudes inference. There is growing evidence that students may have non-optimal help seeking behaviors, and that they seek and react to help depending on

student motivation, gender, past experience and other factors (Aleven et al, 2003). We found that students' negative attitudes towards help and the system are detrimental to learning, and that these attitudes are correlated to specific behaviors with the tutor such as time spent on hints, problems seen per minute, hints seen per problem, standard deviation of hints asked per problem, etc. We created a Bayesian Network from students' log files and surveys about attitudes towards the system, with the purpose of making inferences of students' attitudes and beliefs *while* students use the system, and we proposed remedial actions when specific attitudes are detected (Arroyo et al., 2004).

Teaching strategy selection. Evaluation studies described in section 8 try to capture the link between the spatial and computational teaching strategies described in section 4, and different cognitive abilities (spatial ability and memory retrieval of math facts), with the idea of "macro-adapting" teaching strategies to cognitive abilities, which are diagnosed at pretest time, by selecting one teaching strategy over the other one for the whole tutoring session. Results in section 8 provide guidelines for strategy selection depending on cognitive abilities, which will be implemented and tested in schools in fall 2005.

4 Evaluation studies

We tested the relevance of students' cognitive strengths (e.g., math fact retrieval speed and mental rotation abilities) to the effective selection of pedagogies described in previous sections, to evaluate the worth of adapting help strategy selection to basic cognitive abilities of each student. As described in the previous sections, two help strategies were provided by the tutor, emphasizing either spatial or computational approaches to the solution. The question that arises immediately is whether the help component should *capitalize* or *compensat*e for a student's cognitive strengths. Is the spatial approach effective for students with high spatial ability (because it capitalizes on their cognitive strengths) or for those with low spatial ability (because it compensates for their cognitive weaknesses)? Is the computational help better for students with high mathematics facts accuracy and retrieval speed from memory (because it capitalizes on the fast retrieval of arithmetic facts), or is it better for students with low speed of math fact retrieval (because it trains them in the retrieval of facts)? Given a specific cognitive profile, what type of help should be provided to the student?

4.1 Experiment design

Two studies were carried out in rural and urban area schools in Massachusetts. In each of the studies, students were randomly assigned to two different versions of the system: one providing spatial help, the other providing computational help. Students took a computer-based mental rotation test and also a computer-based test that assessed a student's speed and accuracy in determining whether simple mathematics facts were true or false (Royer et al., 1999).

In the first study, 95 students were involved, 75% females. There was no pre and post-test data, so learning was captured with a *'Learning Factor'* that describes how students decrease their need for help in subsequent problems during the tutoring session, on average. This measure should be higher when students learn more. See a description of this measure (which can be higher than 100%) in (Arroyo et al., 2004). Students used Wayang Outpost for about 2 hours. Students also used the *adventures* of the system for about an hour. After that, students were given a survey asking for feedback about the system and evaluating their willingness to use the system again. The second study involved 95 students in an urban area school in Massachusetts, who used the tutoring system in the same way for about the same amount of time. These students were also given the cognitive skills pretest and a post-tutor survey asking about perceptions of the system.

4.2 Results

In the first study, we found a significant gender differences in spatial ability, specifically a significant difference in the number of correct responses (independent samples t-test, t=2, p=0.05), females having significantly less correct answers than males. Females also spent more time in each test item, though not significantly more. We did not find differences for the math fact retrieval test in this experiment, neither for accuracy nor speed. In the second study, we found a significant gender difference in math fact accuracy (females scoring higher than males). We did not find, however, a gender difference in retrieval speed in any of the two studies, differences that other authors have found (Royer, 1999). We created a variable that combined accuracy and speed to generate an overall score of math fact retrieval ability and spatial ability. By classifying students into high and low spatial and math fact retrieval ability (by splitting at the median score), we established a 2x2x2 design to test the impact of hints and cognitive abilities on students' learning, with a group size of 11-15 students.

In the Fall 2003 study, significant interaction effects were found between cognitive abilities and teaching strategies in predicting learning, based on an ANOVA (R^2=0.68). An interaction effect between mental rotation and the type of help was found (F=3.5, p=0.06, figure 2, table 1). The means in this study suggest that hints capitalize on students' mental rotation: when a student has low spatial abilities, learning is higher with computational help, and when the student has high spatial ability, hints that teach with spatial transformations produce the most learning.

Table 1. Learning as an average percent reduction of help requests in subsequent problems

	Low spatial		High Spatial	
	Spatial help	Comp. help	Spatial help	Comp. Help
Low retrieval	-10%	**47%**	**43%**	-13%
High retrieval	-9%	**325%**	**18%**	-15%

Table 2. Average percent improvement from pre to posttest (pencil and paper)

	Low spatial		High Spatial	
	Spatial help	Comp. help	Spatial help	Comp. Help
Low retrieval	**83%**	43%	**23%**	19%
High retrieval	12%	**19%**	19%	**26%**

In the second study, pre and posttest improvements were used as a measure of learning. A significant overall difference in percentage of questions answered correctly from pre- to post-test was found, $F(1,95)=20.20$, $p=.000$. Students showed an average 27% increase of their pre-test score at post-test time after 2 hours of using the tutor. An ANOVA revealed an interaction effect between type of hint, gender and math fact retrieval in predicting pre to posttest score increase ($F(1,73)=4.88$, $p=0.03$), suggesting that girls are the ones who are prone to capitalize on their math fact retrieval ability while boys are not (table 2). Girls of low math fact retrieval do not improve their score when exposed to computational hints, while they do improve when exposed to spatial hints. A similar ANOVA just for boys gave no significant interaction effect between hint type and math fact retrieval, while another one just for girls showed a stronger effect ($F(1,41)=5.0$, $p=0.03$). The effect is described in figure 3.

In the first study, the spatial dimension was more relevant than the math fact retrieval dimension, while in the second study, math fact retrieval was more important than spatial abilities, despite the fact that students had similar scores on average in the two studies. Despite these disparities, both results are consistent in that that the system should provide teaching strategies that capitalize on the students' cognitive strengths whenever there is one cognitive ability that is stronger than the other one.

Fantasy component. A second goal in our evaluation studies was to find whether the fantasy component in the adventures had differential effects on the motivation of girls and boys to use the system, given the female-friendly characteristics of the fantasy context and the female role models. After using the plain tutor with no fantasy component, we asked students whether they would want to use the system again. Students then used the adventures (SAT problems embedded in adventures with narratives about orangutans and female scientists) after using the plain tutor and we then asked them again whether they would want to use the system. In both occasions, students were asked how many more times they would like to use the Wayang system (1 to 5 scale), from would not use it again (1) to as many times as possible (5).

In the first study, we found a significant gender difference in willingness to return to use the fantasy component of the system (independent samples t-test, $t=2.2$, $p=0.04$), boys willing to return to the "adventures" less than girls. This effect was repeated in the second study (t-test, $t=2.2$, $p=0.03$). This suggests that girls enjoyed the adventures more than boys did, possibly because girls may have felt more identified with female characters, as there is no significant difference in willingness to return

to the plain tutor section with no fantasy component. Again, the adventures section seems to capture females' attention more than males, while the plain tutor attracts both genders equaly. However, significant independent samples t-tests indicated that girls liked the overall system more, took it more seriously, thought the help was useful more than males, heard the audio in the explanations more.

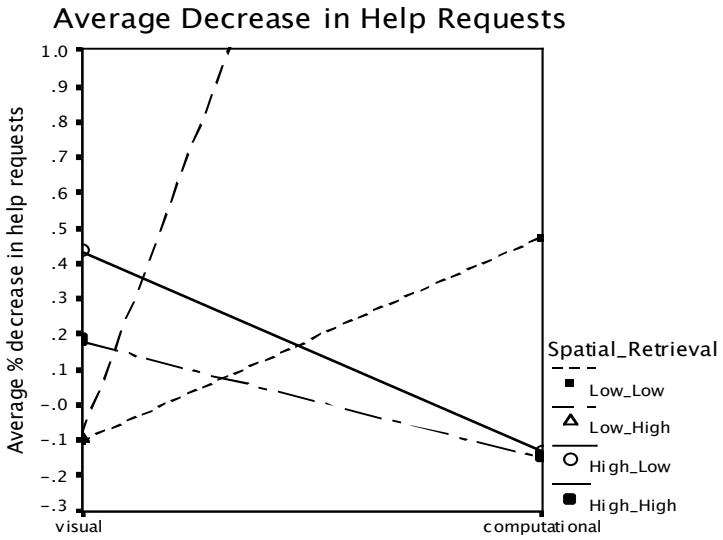


Figure 2. Learning with two different teaching strategies in the Fall 2003 study.

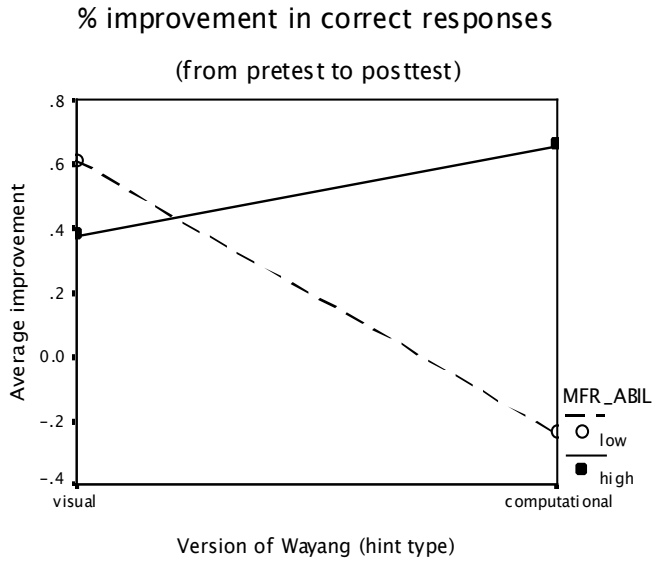


Figure 3. Learning with two different teaching strategies in the 2004 study (girls only).

5 Summary

We have described Wayang Outpost, a tutoring system for the mathematics section of the SAT (Scholastic Aptitude Test). We described how we are adding intelligence for adaptive behavior in different parts of the system. Girls are especially motivated to use the *fantasy* component. The tutor was beneficial for students in general, with high improvements from pre to posttest. However, results suggest that adapting the provided hints to students' cognitive skills yields higher learning. Students with low-spatial and high-retrieval profiles learn more with computational help (using arithmetic, formulas and equations), and students with high-spatial and low-retrieval profiles, learn more with spatial explanations (spatial tricks and visual estimations of angles and lengths). These abilities may be diagnosed with pretests before starting to use the system. Future work involves evaluating the impact of cognitive skills training on students' achievement with the tutor, and evaluating the intelligent adaptive tutor.

Acknowledgements. We gratefully acknowledge support for this work from the National Science Foundation, HRD/EHR #012080. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

References

Arroyo, I.; Beck, J.; Woolf, B.; Beal., C; Schultz, K. (2000) Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. Proceedings of the Fifth International Conference on Intelligent Tutoring Systems.

Arroyo, I. (2003). Quantitative evaluation of gender differences, cognitive development differences and software effectiveness for an elementary mathematics intelligent tutoring system. Doctoral dissertation. UMass Amherst.

Arroyo, I., Murray, T., Woolf, B.P., Beal, C.R. (2004) Inferring unobservable learning variables from students' help seeking behavior. This volume.

Casey, N.B.; Nuttall, R.; Pezaris, E.; Benbow, C. (1995). The influence of spatial ability on gender differences in math college entrance test scores across diverse samples. Developmental Psychology, 31, 697-705.

Royer, J.M., Tronsky, L.N., Chan, Y., Jackson, S.J., Merchant, H. (1999). Math fact retrieval as the cognitive mechanism underlying gender differences in math test performance. Contemporary Educational Psychology, 24.

Shute, V. (1995). SMART: Student Modeling Approach for Responsive Tutoring. In *User Modeling and User-Adapted Interaction*. 5:1-44.

Martin, K., Arroyo, I. (2004). AgentX: Using Reinforcement Learning to Improve the Effectiveness of Intelligent Tutoring Systems. This volume.

Vandenberg, G. S., & Kuse, R. A. (1978). Mental Rotations, A Group Test of Three-Dimensional Spatial Visualization. Perceptual and Motor Skills 47, 599-604

Wainer, H.; Steiberg, L. S. Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: a bidirectional validity study, Harvard Educational Review 62 no. 3 (1992), 323-336.