Wayang Outpost: Intelligent Tutoring for High Stakes Achievement Tests

Ivon Arroyo, Carole Beal, Tom Murray, Rena Walles, Beverly Woolf
University of Massachusetts Amherst

Abstract. We describe Wayang Outpost, a web-based ITS for the Math section of the Scholastic Aptitude Test (SAT). Wayang Outpost has several distinctive features: a large breadth of problems with multimedia animations and sound in the help, problems embedded in narrative and fantasy contexts, alternative teaching strategies for students of different mental rotation abilities and memory retrieval speed. We present evaluations of the tutor that lead to conclusions of effectiveness, which depend on the interaction of teaching strategies and cognitive abilities. A student model built from student interactions with the tutor is described.

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 Massachusetts 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. Although much ITS research focuses on military, industry, and other non-academic training situations, evaluation studies of several ITS for school mathematics also show benefits to student users in school settings. Specifically, studies of the Algebra Tutor (Koedinger, 1997) for algebra and the AnimalWatch tutor (UMass-Amherst) for arithmetic indicate that student users successfully master specific skills and that their attitudes towards math become more positive as a result of working with the software (Arroyo, 2003; Beal&Arroyo, 2002).

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 problems, using information about each student's cognitive skills to customize instruction and improve performance on high stakes assessments, Figure 1.

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 dynamic multimedia techniques such as sound and animation in the way that Wayang Outpost does. Such techniques are common in commercial software (Beal et al., 2002), and have produced higher learning when present in educational software (Mayer, 2001). Second, 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 and SAT-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). Finally, Wayang Outpost incorporates the interactions of previous users with the system to create a data-centric student model, an approach that a few researchers have used, and that has proved beneficial in terms of predictability and enhanced learning (Beck et al., 2000; Mayo & Mitrovic, 2002).

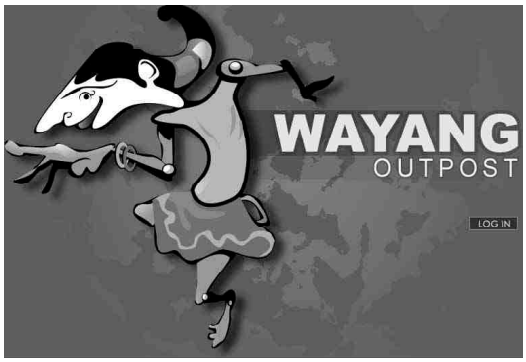

Figure 1. Wayang Outpost showing a shadow puppet.

2. System Description

Wayang Outpost is a web-based intelligent tutoring system created with National Science Foundation support. 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, as well as the need to work quickly due to the time constraints imposed by the testing situation. Wayang Outpost provides instruction via a web site, ensuring easy access to students either at

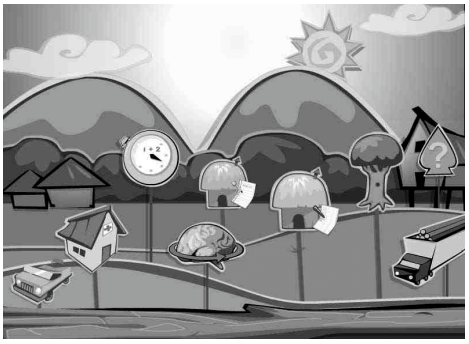
Figure 2. The Wayang Village showing buttons for different student modules, including the SAT hut, adventures, spatial skills test and the math fact retrieval test.

home or from any school connected to the Internet. 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, Figures 1-2. The system is available at http://ccbit.cs.umass.edu/wayang.

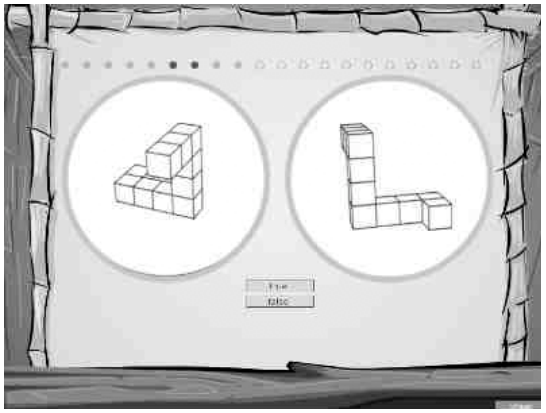

Figure 3. Assessing the student's spatial abilities through a standard mental rotation test.

Each math problem (currently, a battery of SAT-Math problems provided by the College Board) is presented as a Flash movie, Figure 4, including an animated character based on the traditional Indonesian art form of shadow puppetry, Figure 1 (Wayang means shadow puppet). If the student answers incorrectly, or requests help, the teacher character provides step-by-step instruction and guidance in the form of Flash animations with audio. For example, on a geometry problem, the student might see an angle with a known value rotate and move over to the corresponding angle with an unknown value on a parallel line, thus emphasizing the principle of correspondence. 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 mathematics ITS which relied heavily on screen-based text.

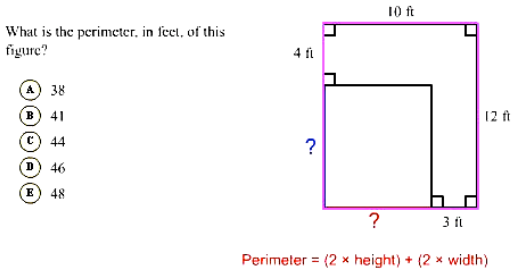

Figure 4. A visual hint provided during a geometry problem. The visual hint suggests to mentally invert the cut-out portion to reveal the intact rectangle, which makes it apparent that the missing lengths are already known. Two forms of help are available: 1) a fairly traditional analytic approach, such as setting up an equation; and 2) a visual component showing that flipping part of a figure would provide key information to make the solution simpler.

3. 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, Figure 3, as

indicated by performance on an standard assessment of mental rotation skill (Casey et al., 1997; Vandenberg et al., 1978). Both tests have captured gender differences in the past. Differences in the most effective pedagogical approaches are explored in the following sections.

4. Help in Wayang Outpost

Each geometry problem in Wayang is linked to two alternative types of hints: one hint provides a computational and numeric approach and the second provides spatial transformations and visual estimations, generally encompassing a spatial "trick" that makes the problem much simpler to solve. An example is shown in Figure 4. 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 women often avoid 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.

5. 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, Figure 5, takes the student across the rainforest to rescue orangutans trapped in a fire. Within the fantasy adventure, Figure 9, students are provided with text message hints and shown SAT problems with multimedia help, Figure 4, which are similar to the problem being solved within the adventure. T he Lori Perkins character 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.

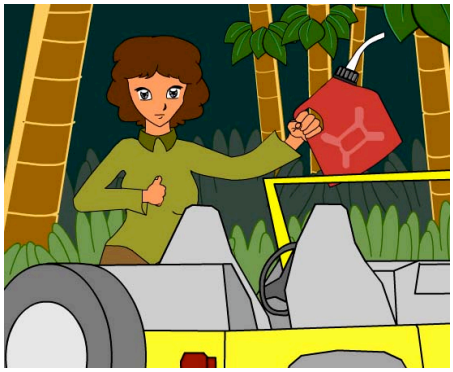

Figure 5. Anne Russon leads the student to save oraguntans. "We need to find the shortest route and estimate how much extra gas we should take"

6. Architecture.

Wayang Outpost provides instruction through the WWW, ensuring easy access to students either at home or from any school connected to the Internet. As the student works through a problem, performance data (e.g., latency, answer choice, hints requested) are stored in a centralized database. This data constitutes what we called the "episodic" student model, i.e. raw data about every student interactions with the system. From this data, the system makes inferences on an ongoing basis to select problems at the appropriate level of challenge, and chooses hints that will be helpful for the student, as described in the student modeling section. Communication between Flash and the Java-based tutor is via XML calls, Figure 6.

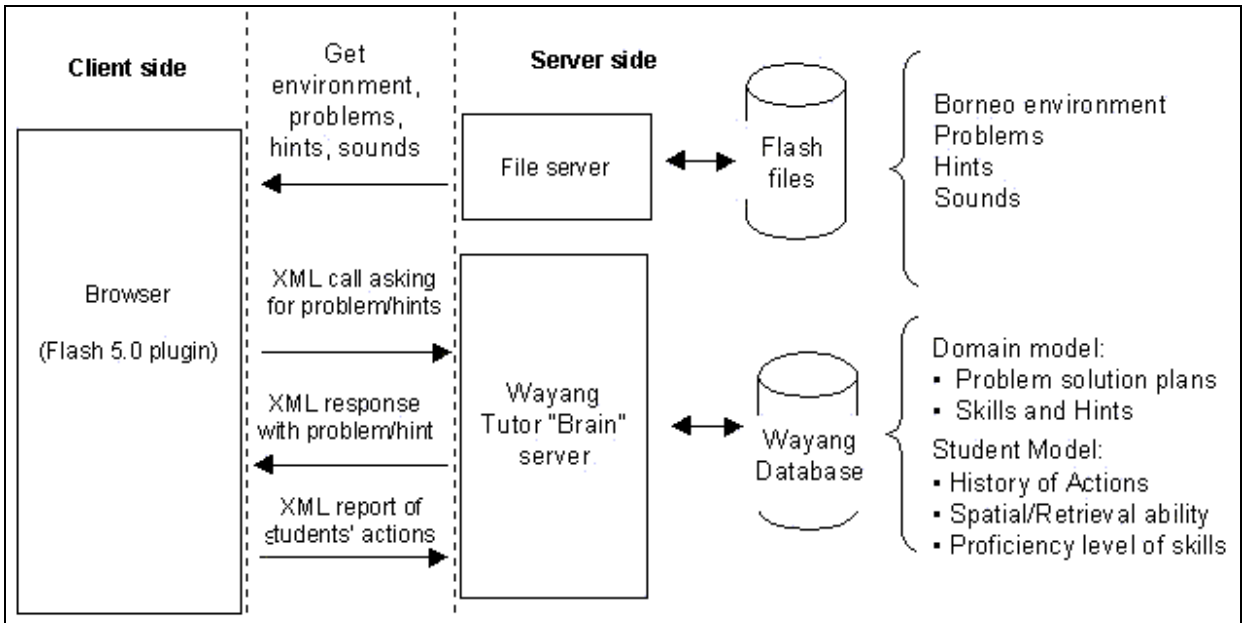

Figure 6. The architecture of Wayang Outpost

7. Student Model

The main purpose of the student model is to guide problem and hint selection. Problem selection should be geared to problems that 1) utilize skills that are not likely to be known, and problems where 2) students are likely to make some mistakes, or ask for a certain amount of hints (Murray & Arroyo, 2002). Unlike other tutors, Wayang Outpost doesn't trace each step of the students' solution, because this would be too expensive to implement for so many different problems. However, the tutor observes the hints requested by the student to reach the solution. Because each hint that Wayang provides (to aid in a step of the solution) also summarizes previous steps, the tutor may skip hints and save time when it realizes that the student doesn't need them. But what are useful hints to show? We believe that useful hints are those which will make the student reach a correct answer right after they are shown these hints, as the probability of the student needing help with that last step was high.

One of our goals is to provide improved tutoring and more effective learning by building a student model that can predict students' behavior after seeing a hint (e.g., Will the student provide a correct response or ask for another hint? What and how many hints will be requested for a problem? Does the student already know a skill or has a good chance of learning the skill after seeing specific problems or hints?).

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.

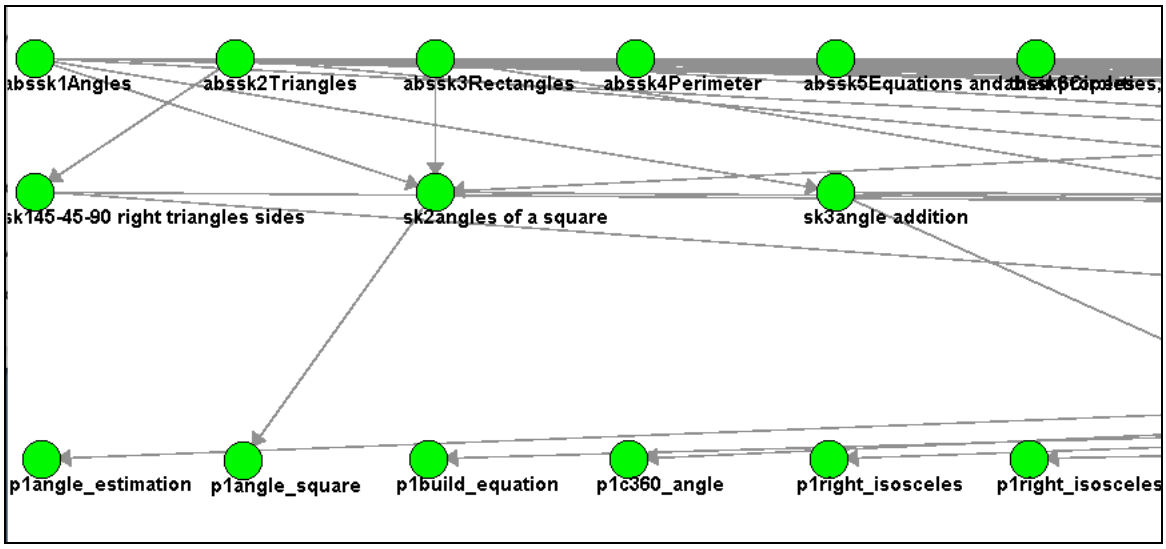

Figure 7. A data-centric Bayesian Belief Network student model

We utilize an approach to student modeling that combines human expertise and a data-centric approach. A 'data-centric' approach with Bayesian Networks implies that the structure and conditional probabilities of the BBN are collected from actual students' interactions with the system (Mayo & Mitrovic, 2002). In other words, prior data from previous users of the system was collected off-line to develop a probability model that a particular on-line student has a certain set of geometry skills. Human knowledge engineering helped determine the skills involved in the problem solutions, and to identify hints to help students at different steps of the solution related to specific skills (such as finding the area of a circle). The data-centric approach implies the collection of conditional probability tables (CPTs) from prior students' use of a system that made random decisions about problem and hint selection. This approach is different from Mayo's in that 1) we model skills instead of constraints; 2) the BBN has multiple layers (hidden mastery nodes); and 3) we model help seeking to determine how seeing a hint is related to knowing a skill. This approach is similar to the student model developed in

ANDES (Conati et al., 2002) in that there are nodes to represent skills and hints. However, Conati's model did not learn CPTs from student interactions with the system.

The Bayesian Network has three layers, Figure 7. The highest two layers model the probability that the student *already knows a skill, is learning a skill* or *is not learning a skill*. The top layer indicates abstract skills (e.g., knows about angles, triangles, rectangles, etc. The middle layer indicates specialized skills (e.g., knows about 145-45-90 right triangle, angles of a square). The lowest layer models students' response to hints. Figure 7 shows leaf nodes for problem 1, and indicates that several hints can be shown to the student. The leaf nodes indicate available hints, such as p1angle_square hint (bottom level, left side, Figure 7) and takes on 3 values (instead of T and F): answered correctly immediately after seeing the hint, not answered correctly after seeing the hint; and not requested the hint at all (60 students as indicated in the top left cell of Table 1,). This hint node is linked to a more generic node in layer 2 that models the probability that the student knows the angles of a square. Skill nodes can take 3 possible values (instead of T and F): the student *already knew the skill*, *learned the skill*, or *didn't learn skill*. Note that many hint nodes are linked to each skill (all hints that aid the learning of that skill).

The p1_angle_square hint node has a conditional probability table (CPT) attached to it that was obtained from prior student logs shown in Table 1. The top table indicates prior student observations or the number of cases when past students behaved in a certain way based on their knowledge of a skill. The first cell in the prior student observation table indicates the number of students who *already knew a skill* and did not request the hint p1_angle_square for problem 1 (60 students). A student *already knows a skill* when he/she scores above a certain percentage in the first problems in the session that use that skill (or in a computer-based pre-test). The student *learned the skill* when there is a significant improvement from the beginning to the end of the session in problems involving the skill. A student *did not learn the skill* when the student had a low score for the skill at the beginning of the session and there is not a significant improvement in the number of correctly solved problems involving the skill from the beginning to the end of the session. The conditional probability table is built out of these observations (e.g. there were 60 students who already knew the skill and did not request the p1_angle_square hint).

Number of cases in which the student	Student *already knew skill*	Student *learned skill*	Student *did not learn skill*
No request of hint	60	9	30
Correct after hint	12	65	5
Not correct after hint	8	26	15

Probability of…	**Student *already Knows skill***	**Student *learns skill***	**Student *does not learn skill***
No request of hint	0.75	0.09	0.6
Correct after hint	0.15	0.65	0.1
Not correct after hint	0.10	0.26	0.3

Table 1. Conditional probability attached to the p1angle_square hint node, including prior observations (top) and conditional probability (bottom).

The conditional probability of a student *knowing*, *learning* or *not learning* a particular skill is indicated in the lower table. To calculate the conditional probability of a student *knowing*, *learning* or *not learning* that skill, for all possible values of the current hint node (for each column) the number of cases is added, and each cell in the column is divided by this sum. The overall effect is that if the student entered a correct answer after seeing the p1_angle_square hint, then that node is 'observed' with the 'Correct-after-hint' value, and that affects the help seeking predictions for other hints that tackle the same skill, as they are indirectly connected through the same skill node.

The highest layer of the model serves to generalize across skills, based on the idea that if a student knows how to calculate the perimeter of a triangle, then it is likely that they know how to calculate the perimeter of a square (both perimeter of a triangle and perimeter of a square are linked to the abstract skill called 'perimeter'). Note that the skill 'sk2angles-of-a-square' is linked to the 'abstract level' skill called '**abssk1Angles'**, which is a macro-level skill that is also linked to other skills such as 'sum of internal angles of a triangle', or 'supplementary angles'. When using the student model in real time, the prior probabilities of abstract skills will be initiated based on computer-based pretest scores.

We are still investigating the accuracy of this model at predicting students' knowledge using cross-validation techniques (obtaining the CPTs from 80% of the data, then testing with the remaining 20%). Then, we will compare the system's predictions to what the remaining 20% of the students actually did (e.g., did they request any hints? Which hints did they see and how does it compare to what the model predicted? How many hints did the student ask for compared to what the model estimates he/she is most likely to ask? What is the advantage of having the 'abstract skill' layer vs. not having it?). The effectiveness of a system that makes decisions based on this student model will be tested in Massachusetts schools at the end of Spring 2004. Results on the predictability of the model and evaluation with students will be presented at the ITS conference.

8. Evaluation studies

Figure 8. Students using Wayang Outpost at school. Deerfield, MA.

Prior research indicates that students did learn with the Wayang Tutor, improving an average of 19% from pre to post-test after 1.5 hours of problem solving, just by having the system make random selection of problems and hints (Arroyo et al., 2004). One important goal of the evaluation presented in this section was to test 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. Thus, one goal was to decide whether hint selection should be adjusted to the cognitive skills of each student. As described in the previous sections, two help strategies may be provided by the tutor, emphasizing either spatial transformations 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 retrieval speed of mathematics facts from memory, or is it better for students with low speed of math fact retrieval? Given a student with a specific cognitive profile, what type of help should be selected for him/her?

8.1 Experiment design

Over a hundred students in rural area schools in Massachusetts were randomly assigned to two different versions of the system: one providing spatial help, the other providing computational help. Students took a computer-based Purdue mental rotation test (Vandenberg et al., 1978), Figure 3, 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).

Learning was measured with a *'Learning Factor'* that described how students decrease their need for help in subsequent problems, on average. Ideally, students would need less help as time goes by, after accounting for the difficulty of each problem seen. This learning measure describes the percentage of help requests the student made, minus what is expected for the problem. For instance, if a large group of students on average tended to ask for 2 hints in a specific problem before answering it correctly, and the current student requested 3 hints, then, the student requested 50% more hints than expected (0.5). If the student instead requested 2 hints, then he/she is having an average behavior (0% more hints than expected). If the student requested no hints at all, he/she is requesting 100% less hints than expected (-1.0). The average difference in help requests between the (i+1)th and the ith problem in all the tutoring session becomes a measure of how students need for help fades way before choosing a correct answer. This is an observable measure of learning, which should be higher when students learn more.

Students used Wayang Outpost for about 2 hours. They were given a survey after using the tutor, to evaluate their perceptions of help and their willingness to use the system again. After eliminating incomplete data (e.g. students who took the tests but didn't use the tutor, or students who used the tutor but took only one of the tests) we had 95 students with full data. 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 11-12 students.

After using the tutor and getting feedback, students used the *adventures* of the system for about an hour. After that, students were given a survey asking for feedback about the adventures and evaluating their willingness to use the system again.

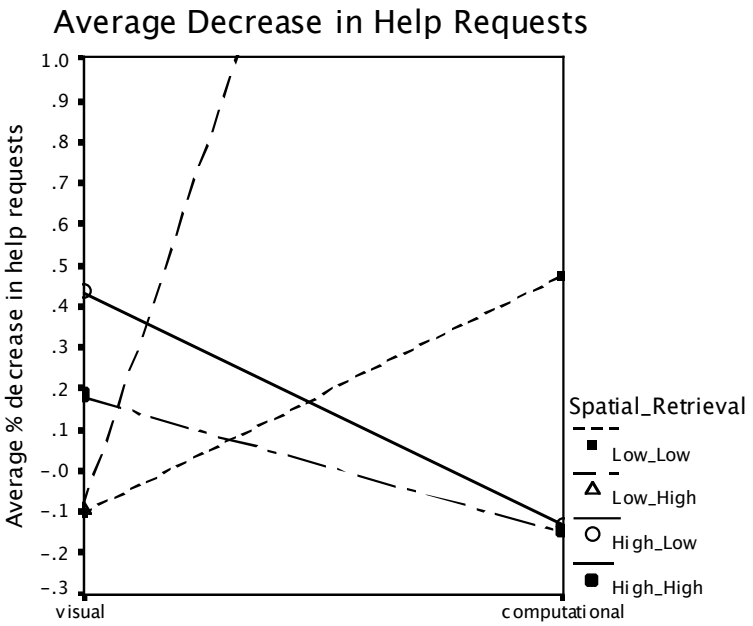

Figure 9. Interaction effect of the learning factor between student groups with different spatial and math fact retrieval skills. Students with low-spatial and high-retrieval profiles learn more with computational help, and students with high-spatial and low-retrieval profiles, learn more with spatial transformations.

8.2 Results

Cognitive abilities & teaching strategies. We found significant gender differences in spatial ability, specifically a significant difference in the number of correct responses (t-test, t=2, p=0.05), females having significantly less correct answers than males. At the same time, females spent more time in each test item, though not significantly more. We did not find differences for the math fact retrieval test though. We found significant interaction effects between the learning factors described in previous sections, based on an ANOVA, Figure 9. We found an interaction effect between cognitive profile and hint type ($F=2.97$, $p=0.08$; $R^2 = 0.68$). The means suggest that hints capitalize on students' cognitive profile; when a student has a low-spatial and high-retrieval profile, learning is higher with computational help, and when the student has a high-spatial and low-retrieval profile, hints that rely on spatial transformations produce the most learning.

We also made statistical analyses on another dependent variable: students' perception of the helpfulness of the system. This was part of a survey given to students after they finished using the system, answering the question *"What did you think about the help in the system?"*. Possible answers to the question were: Wasn't good or clear at all (1), In general wasn't clear or helpful (2), helpful sometimes (3), Helpful most of the time (4), It was great (5). An ANOVA yielded no significant differences for the 8 groups; however, the means follow the exact same direction as the learning measure.

Given the consistency of both students' perceptions of help and of our measure of learning with the different kinds of help, our conclusion is that the system should provide the hints that capitalize on the students' cognitive strengths, if there is one cognitive ability that is clearly better than the other one. If both abilities are low, computational help was best. If both abilities are high, spatial hints should be provided, as they yield higher learning. A new version of the system will be built that macro-adapts hint selection to math fact retrieval and spatial abilities. It will be tested against another version of the system that does not adapt its hint selection depending on cognitive abilities

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

Table 2. Learning as an average percentage reduction of the need for help

	Low spatial		High Spatial	
	Spatial help	Comput. Help	Spatial help	Comput. Help
Low math retrieval	3.2	**4**	**3.9**	3.76
High math retrieval	3.5	**3.8**	**4**	3.6

Table 3. Students' perception of the helpfulness of the system (1 to 5 scale)

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, Figures 5 and 11. 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): Would not use it again (1), Use it one more session (2), Come back a couple more times (3), Use it many more times (4), As many times as possible (5). We later compared girls' and boys' motivation to use the fantasy adventures.

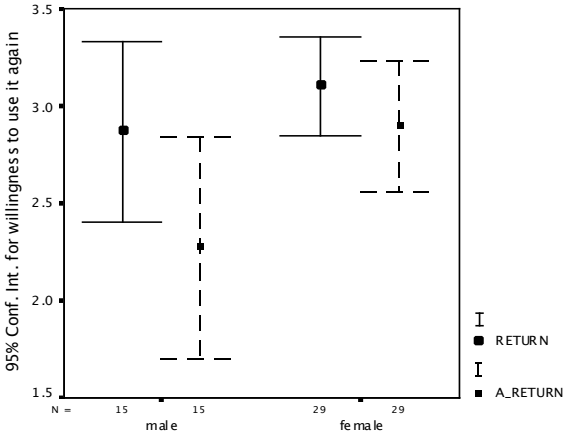

Figure 10. Willingness to return to the plain SAT tutor and willingness to return to the fantasy embedded tutor.

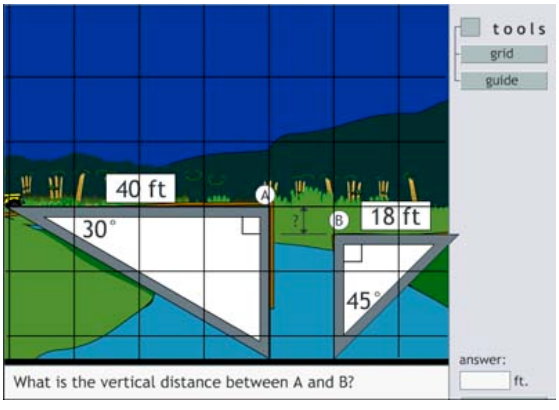

Figure 11. A problem embedded into a mathematics adventure: Will the jeep make it if we try to jump from A to B? The jeep can jump no more than 3 feet.

We found a significant difference in boys' willingness to return to use the system for the fantasy component of the system (paired samples t-test, t=2.2, p=0.04), Figure 10. However, there was not a significant difference for girls' willingness to return to the system after seeing the adventures. This implies that girls enjoyed the adventures more than boys did, and a hypothesis for this is that girls may have felt more pleased with female characters.

9. Conclusions

We have described Wayang Outpost, a tutoring system for the mathematics section of the SAT (Scholastic Aptitude Test), used by students in Massachusetts high schools. Building student models for high-stakes achievement tests poses new challenges, such as determining problem difficulty, modeling high guessing factors and assessing a large breadth of student skills with low occurrence of each skill per problem. We approached the problem with a mix of human expertise (knowledge engineering) and data-centric techniques by learning the conditional probability tables of a Bayesian Network from prior student interactions with the system, which has nodes corresponding to hints (that correspond to steps in the solution) and skill nodes.

We have found that girls are specially motivated to use the *adventures* of the system, and that the tutor without intelligence was beneficial for students on average. We attribute this fact to the high quality of the help and the proper use of multimedia in the explanations (Mayer, 2001). Further evaluations summarized in this paper show how macro adapting the hints to students' cognitive skills can yield higher learning results. We concluded that students with low spatial ability should avoid visual help (instead should receive help that uses arithmetic, formulas and equations); students with high spatial ability should receive visual help (rely on spatial tricks and visual estimations of angles and lengths). Future work involves evaluating the impact of cognitive skills training on students' achievement with the tutor. We intend to investigate whether training spatial abilities (with a tutoring system that trains mental rotations with 3-dimensional cubes) or training math fact retrieval speed (with exercises that help a student memorize arithmetic facts) helps improve performance in the tutor and/or learning with different teaching strategies that capitalize on these basic cognitive skills.

Acknowledgements:
We gratefully acknowledge support for this work from the National Science Foundation, HRD/EHR #012080, Beal, Woolf, & Royer, "AnimalWorld: Enhancing High School Women's Mathematical Competence." 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.

Arroyo, I., Walles, R., Beal, C. R., Woolf, B. P. (2004). Effects of web-based tutoring software on students' math achievement . Accepted to the American Educational Research Association annual meeting. San Diego, CA.

Beal, C. R., Beck, J., Westbrook, D., Atkin, M., & Cohen, P. R. (2002, March). Intelligent modeling of the user in interactive entertainment. Paper presented at the AAAI Stanford Spring Symposium, Stanford CA.

Beal, C. R., Arroyo, I. (2002). The AnimalWatch project: Creating an intelligent computer mathematics tutor . In S. Calvert, A. Jordan, R. Cocking (Eds.), Children in the digital age. Greenwood.

Beck, J.; Woolf, B; Beal, C. (2000) ADVISOR: A machine learning architecture for intelligent tutor construction. In the Proceedings of the Seventeenth National Conference On Artificial Intelligence.

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.

Casey, N. B., Nuttall, R. L., & 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.

Conati C., Gertner A., VanLehn K., Druzdzel M. (1997). On-Line Student Modeling for Coached Problem Solving Using Bayesian Networks . In Jameson A., Paris C., Tasso C., (eds.) User Modeling; Proceedings of the sixth International Conference UM97. Springer.

Conati, C., Gertner, A.S., VanLehn, K.: Using Bayesian Networks to Manage Uncertainty in Student Modeling. User Modeling and User-Adapted Interaction 12 (4): 371-417 (2002)

Koedinger, K.R., Anderson, J.R., Hadley, W.H. and Mark, M.A. (1997). Intelligent tutoring goes to school in the big city , 8, 30-43.

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

Mayo, M., Mitrovic, A., Optimising ITS behaviour with Bayesian networks and decision theory , IJAIED, vol. 12(2), 2001, 124-153.

Murray, T., Arroyo, I. (2002). Towards Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems . Proceedings of the Sixth International Conference on Intelligent Tutoring Systems. Springer.

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, 181-266

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

Vandenberg, G. Steven, & Kuse, R. Allan. (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.