

Tutoring for SAT-Math with Wayang Outpost

Ivon Arroyo, Rena Wallace, Carole R. Beal, Beverly P. Woolf

University of Massachusetts, Amherst

Abstract. We describe our on-going work in the creation of a web-based Intelligent Tutoring System (ITS) for the mathematics section of the Scholastic Aptitude Test (SAT). Wayang Outpost focuses on geometry problems, and uses web-based multimedia to communicate concepts to the student. Decisions about problem and help selection made on a remote web server, which stores data on student-system interactions and will eventually reason about students' cognitive abilities. Wayang has been designed with alternative teaching strategies. Difficulties in developing ITS for high-stakes achievement tests are analyzed.

Introduction

High stakes achievement tests have become increasingly relevant in the past years in the United States, as a student's performance on them can have a significant impact on students' access to future educational opportunities (such as admission to universities, scholarships and, in some states, graduation from high school). At the same time, concern is growing that the use of such tests simply exacerbates existing group differences, and puts female students and those from traditionally underrepresented minority groups at a disadvantage. In particular, a steady 50-point gender difference has been maintained over the last 30 years (Langenfeld, 1997). Because of the importance of the score in these tests, new approaches are required to help all students perform to the best of their ability on high stakes tests. It is suspected that spatial ability and math fact retrieval are important determinants of the score in these standardized tests. Some studies found that when mental rotation ability was statistically accounted for, the gender difference in SAT-Math disappeared (Casey, 1995). In other studies, math fact retrieval was found to be an important cause of gender differences (Royer, 1999). This paper describes our ongoing work in creating "Wayang Outpost", an Intelligent Tutoring System to prepare students for the mathematics section of the Scholastic Aptitude Test (SAT). The intention is that Wayang does not only tutor every student effectively by tailoring the difficulty of problems and help thanks to a student model and intelligent pedagogical decisions, but also addresses factors that make female students score lower in these tests. By taking into account the cognitive abilities of each student, we expect Wayang Outpost to prepare both genders for SAT-Math. Wayang currently focuses on geometry problems, the major source of gender differences in the past. We intend to demo the Wayang system during the workshop, and present some of our preliminary results from an evaluation with 60 students in two different schools, and our ideas for future work on it.

1. System description

Wayang Outpost is a web-based tutoring system that teaches students to solve geometry SAT problems. Wayang Outpost uses multimedia to help students in their problem solving process, directing their attention, animating parts of the solution, and emphasizing concepts with sound. Multimedia is also used to motivate students, by adding a video-game style to the tutor. Situated in the context of a FlashTM animated research station in the Borneo rainforest, students address environmental issues around the orangutan endangered-species.

A pedagogical agent, an animated Indonesian shadow puppet, presents SAT problems. When the student requests help, the puppet provides step-by-step instruction in the form of animations, displaying different emotions (sad, happy or confused), see figure 1.

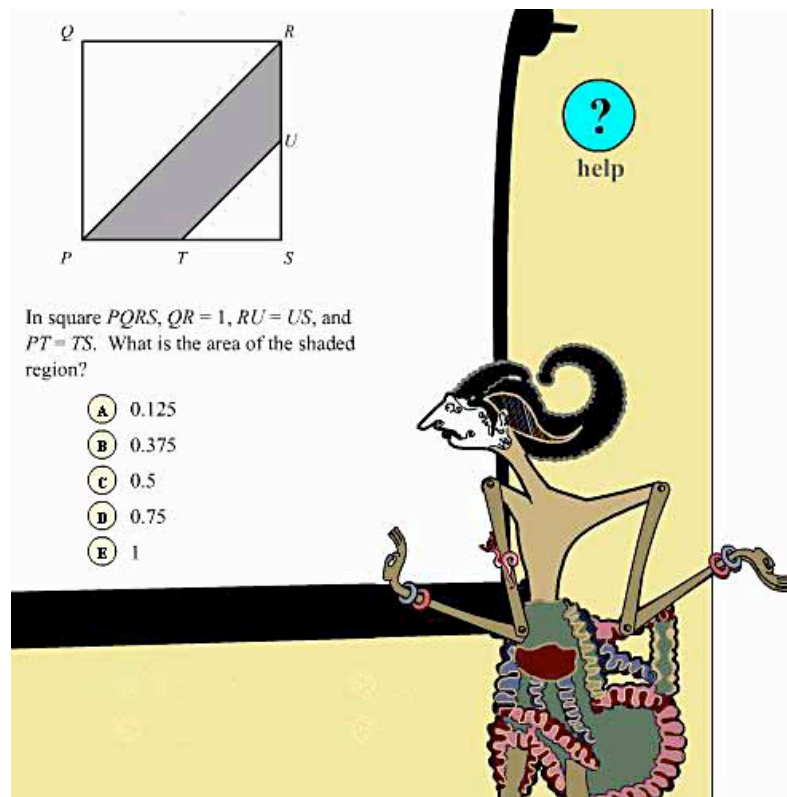


Figure 1. The Wayang Outpost tutor

Our past research has suggested that student cognitive abilities and learning styles are important at predicting performance and hint success in a tutoring system (Arroyo, 2003; Arroyo, 2000). In this case, we are currently analyzing the relevance of spatial ability as measured with the Purdue mental rotation test (Vandenberg, 1978) and, in the future, math fact retrieval speed (Royer, 1999) on the prediction of problem solving success and help success. Performance in the tutor will be used to identify the most critical cognitive skills predicting successful solutions, which may differ by gender.

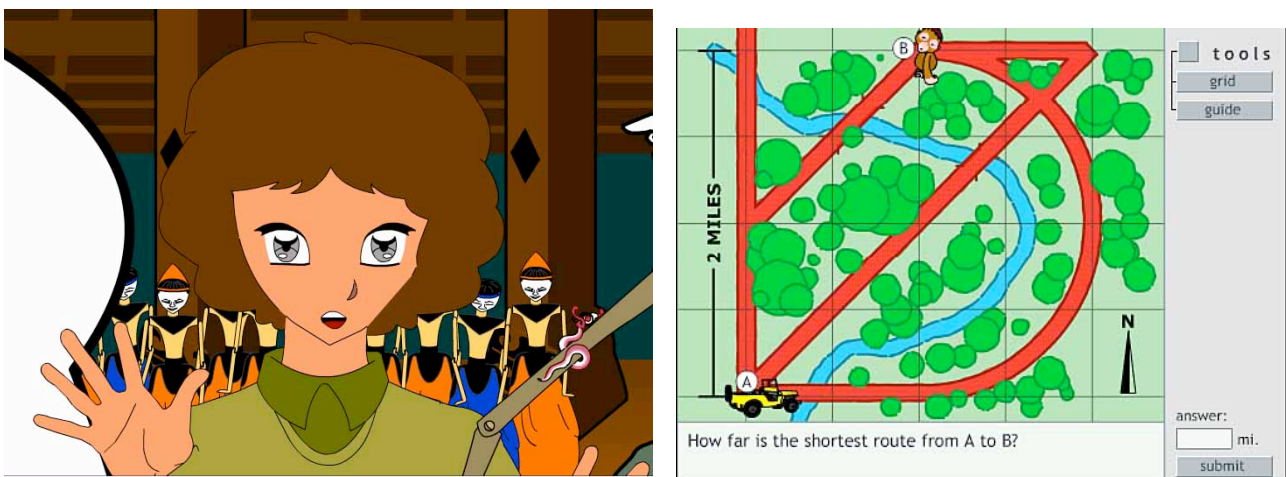


Figure 2. "Rescuing trapped orangutans" adventure

Wayang contains “short-term transfer” problems --variations of the operands of an SAT problem and minor changes in figures-- to assess learning. These problems are designed with the idea of assessing student improvement within the student session. Wayang also incorporates “long-term transfer” problems in the form of animated adventures, which present real-world math problems following a story line. These problems do not have SAT format, as questions are not multiple-choice and are more sophisticated than SAT problems. Adventures are based on current cartoon styles and are rich in movement, sounds and character emotions. In addition, adventures incorporate the presence of characters inspired on actual scientists working with the orangutan species. Figure 2 shows the character of Anne Russon within an adventure, a professor from York University who specializes in the study of these primates. Performance on adventure problems becomes a measure of transfer of math skills from the multiple-choice tutor to a different context where the same mathematics skills are needed to solve problematic situations, such as determining if it is safe to jump down a cliff, estimating how much gas is needed to get back to the camp, etc. Adventure problems do not provide help, as their major goal was to have a posttest measure of transfer (that student would be really eager to take). However, we are considering the possibility of adding help to it, and analyzing differences in the amount of help asked for from pretest adventure to posttest adventure.

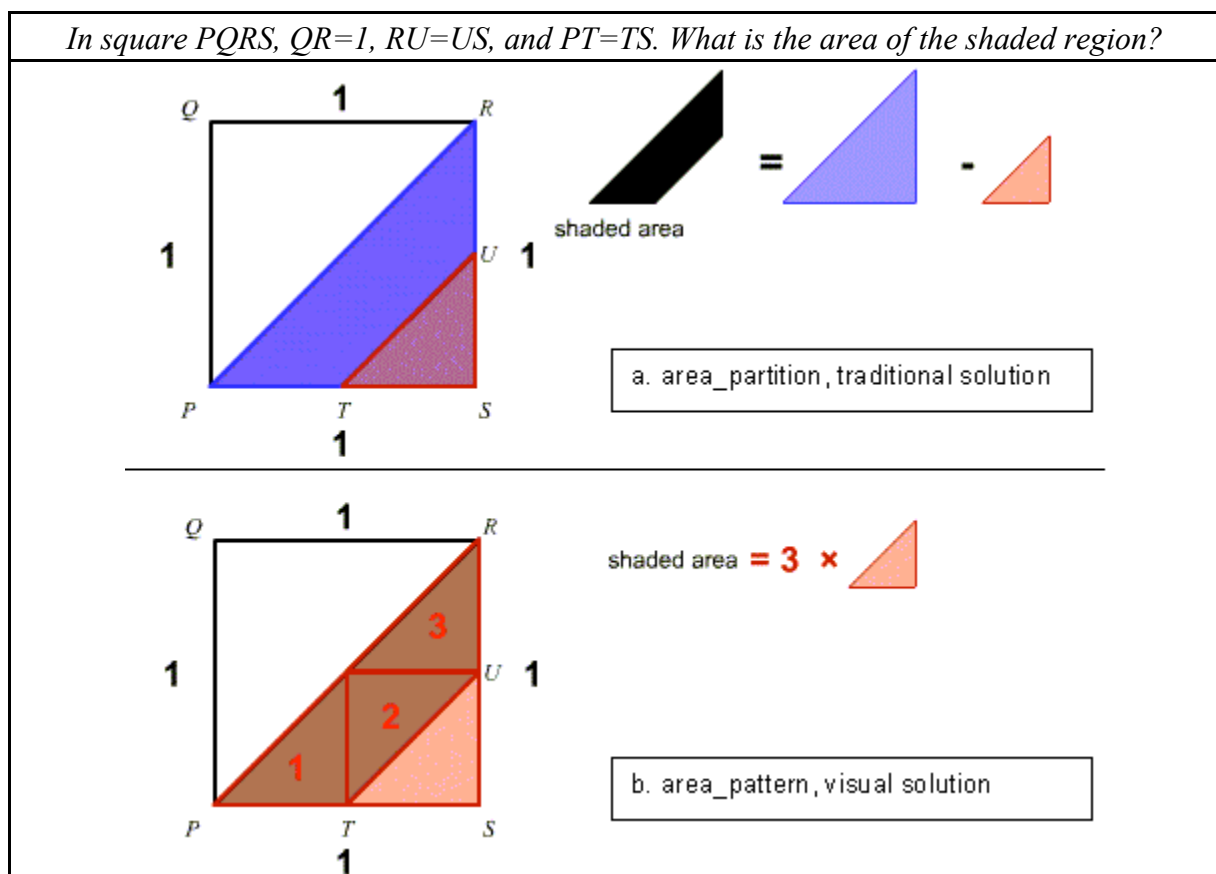


Figure 3. Two alternative kinds of hints to an SAT problem

After analyzing computer science graduate and undergraduate students successfully solve these problems, we detected two major approaches: one relying on estimations related to spatial abilities (e.g. mental rotations), and a second one relying on fact retrieval and algebra (e.g. heavy symbolization). We created two kinds of hints, one based on an analytical

approach, the second based on a visual estimations approach. We think these strategies will affect students of varying spatial abilities differently. Figure 3 shows two alternative hints in the solution plan of figure 3, namely *area_pattern* and *area_partition*. Figure 3.a shows *area_partition*, a traditional way to solve this problem. Figure 3.b shows a spatial hint, which identifies a pattern of three smaller triangles in the shaded area, by moving and flipping the small triangle in the bottom corner. Such spatial “tricks” allow the problem to become simpler, or have shorter solutions, or make accurate estimations. We think that the success of spatial hints may be related to spatial ability, and eventually to gender. If such is the case, we may consider training low spatial ability students in spatial abilities such as flipping figures, estimating lengths, mental rotations. We may also consider training students in the fast memory retrieval of math facts.

The main “intelligent” pedagogical decisions in Wayang Outpost will have to do with problem and help selection. We aim for a small number of mistakes per problem and certain amount of time per problem (speed is an important matter in these tests), and avoiding that the student “skips” the problem. Each problem has associated solution plans, with hints and skills attached to each step in the solution. The tutor will determine which steps in the solution the student has most likely failed at, targeting the help to faulty skills. It will choose a visual or analytical hint depending on estimates of hint effectiveness. Evaluation studies of Wayang Outpost have recently been conducted, where mistake reduction in subsequent problems and pre to post-test improvements are measured to assess help effectiveness and problem difficulty.

3. Differences with other ITS. Challenges.

Several issues make ITS for high-stakes achievement test tutors difficult to implement. The first challenge is *what to tutor*. Most ITS designers know what topics they tutor, while the major burden is to decide how to teach them. When dealing with high-stakes achievement tests, the content is known but the skills to tutor are hidden. Thus, deep knowledge engineering is needed, which implies unfolding alternative approaches to problems, and identifying skills at various levels (mathematics, test-taking and problem solving skills). The second challenge is that, unlike most mathematics educational software, many of the skills involved seem to be independent from each other, with hard to estimate difficulty levels. In addition, high-stakes tests have multiple-choice questions, which imply a high guessing factor. This adds uncertainty to the beliefs of students’ knowledge: even if a problem is correctly answered, there is always 20% chance that the student has randomly chosen an answer. Last, there is a large breadth of skills while the skill frequency per problem is low, and it is rare that two problems involve exactly the same skills. For instance, problem 1 applies skills A, B and C and problem 2 applies skills C and D). This poses a challenge to the estimation of the difficulty of problems, and thus to the decisions of problem selection.

We believe that a data-driven approach is the best way to generate a domain, student and pedagogical models for high-stake-test intelligent tutors. In particular, the high degree of uncertainty makes Bayesian Belief Networks a convenient technique to tackle these problems, in which nodes could range from observable problems and hints to skills at different levels of abstraction. This would allow that after a student solves a problem involving both skills A and C, information is gained about the likelihood of success of a

problem involving skills A and D, even if the problems don't involve exactly the same skills. In addition, such data-oriented approach would allow learning about the usefulness of hints. Given some student knowledge state and student cognitive abilities, we will build models to predict whether not lead the student to enter a correct solution immediately after it was shown. Following the philosophy of Mayo and Mitrovic (2001) we expect to learn conditional probabilities of success at different problems and conditional probabilities of skill mastery, from actual student interactions with the system. There are difficulties related to using BBNs, mainly that they can take exponential time at propagation of evidence. We are thus considering the utilization of approximate inference in the case that time is not convenient for exact inference updates.

4. Evaluation studies

Evaluation studies have been carried with 63 high school students from two different high schools in Massachusetts, during March and April 2003. Students ranged from 16 to 18 years old. The intention was to try out the first version of the system with actual students, analyzing its effectiveness, and gathering data about the difficulty of problems and the usefulness of hints. Students took the Purdue mental rotation test, and then a 45-minute pretest with a battery of SAT geometry problems that were different from the ones in the tutor, but tackled similar topics. Students were then assigned to one of two versions of the system: one that provided visual estimation hints or one that provided analytical hints. Both versions provided randomly selected problems, as we hope to learn information about the difficulty of problems from this data. Students used the Wayang tutor for about 1 hour. Students then took a 45-minute posttest, similar to the pretest but with different problems. We guaranteed the posttest was of the same difficulty than the pretest by giving test A to half of the class as a pretest and test B to the other half of the class; we then switched tests for posttest time. We had students see the adventure with long-term transfer problems. We asked students for feedback on the tutor and on the adventure. We are currently in the process of analyzing this data. However, we will present some descriptive statistics and preliminary results in this section.

As can be seen in table 1, students significantly improved the percentage of correct answers from pre to post-test. They also reduced the amount of questions they skipped in the test. The amount of incorrect answers stayed the same. Students learned with the system even if they were exposed to it for just one hour. This evaluation shows that the help we are providing is effective, and sets a baseline for higher improvements in the future due to adaptive problem and help selection. We are starting to analyze the relevance of gender and spatial ability in relation to the learning rates produced by specific hints. Future versions of Wayang Outpost will incorporate a student model and pedagogical decisions that are based on this past experiments with high school students.

In general, students enjoyed working with the system, and said they preferred it to a standard math class. To the open question of what they thought of the animations as a way of learning many students responded something along the lines of: "that's the way I like to learn, by being shown and told how to solve each step of the problem". We are still analyzing these open questions. Students enjoyed the cartoons and the narrative of the

adventure, even though they found the long-term transfer problems challenging, as the system provided no help for those problems.

Descriptive Statistics

| | N | Minimum | Maximum | Mean | Std. Deviation | Paired samples t-test |
|----------------------|----|---------|---------|-------|----------------|-----------------------|
| % Correct pretest | 68 | 9.52 | 76.19 | 37.98 | 12.97 | p<0.010 |
| % Correct posttest | 67 | 9.52 | 85.00 | 42.69 | 16.34 | |
| % Skipped pretest | 68 | .00 | 55.00 | 15.33 | 17.85 | p<0.085 |
| % Skipped posttest | 67 | .00 | 66.66 | 11.00 | 17.77 | |
| % Incorrect pretest | 68 | 4.76 | 85.71 | 46.75 | 19.51 | p<0.644 |
| % Incorrect posttest | 67 | 9.52 | 80.95 | 46.29 | 17.35 | |
| Pretest score* | 68 | -9.0 | 43.0 | 13.23 | 10.6 | p<0.017 |
| Posttest score* | 67 | -11.4 | 48.8 | 16.52 | 12.33 | |
| Valid N (listwise) | 63 | | | | | |

* this score takes into account correct, incorrect and skipped problems

Table 1. Differences in performance from pretest to posttest

4. Summary

We have described Wayang Outpost, a tutoring system for the mathematics section of the SAT achievement test. Building ITS for high-stakes tests poses new challenges, such as determining problem difficulty, high guessing factors and a large breadth of skills with low occurrence per problem. We believe a data-centric approach such as the learning of Bayesian networks from actual student interactions with the system is appropriate to tackle this high level of uncertainty. Wayang Outpost will be nationally and internationally disseminated at the end of this research and development process. The web-based architecture will make this last stage easy to accomplish.

References

Arroyo, I.; Beal, C.; Woolf, B; Murray, T. (2003) Further results on gender and cognitive differences in help effectiveness. Proceedings of the 11th International Conference on Artificial Intelligence in Education.

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 5th International Conference on Intelligent Tutoring Systems.

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

Casey, M; 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.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational measurement*, 16, 20-26.

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

Mayo M., Mitrovic A. (2001) Optimising ITS behaviour using Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education* 12:124-153.