Tech., Inst., Cognition and Learning, Vol. 4, pp. 31-63
Reprints available directly from the publisher
Photocopying permitted by license only

Addressing Cognitive Differences and Gender During Problem Solving

Ivon Arroyo, Beverly P. Woolf[1]*, Carole R. Beal[2]*

[1] *Computer Science Department, University of Massachusetts, Amherst, MA 01003 U.S.A.*
[2] *Information Sciences Institute, University of Southern California*

This research evaluated the impact of supplementing user models with additional data about cognitive features of the student. Supplemental data included individual differences variables such as: developmental stage of the learner (Piagetian), spatial ability, math-facts-retrieval and gender. These differences were applied along with multimedia and customization in two intelligent tutoring systems, one for arithmetic and one for geometry. The research resulted in the general conclusion that enhancing user models with detailed information about cognitive ability led to improved response to instruction. This is especially important to consider for domains for which there are well-established group differences, such as gender differences in mathematics.

Keywords: Cognitive development, individualization of instruction, student modeling, tutoring software, mathematics education, spatial ability, mathematics facts retrieval.

1 CUSTOMIZATION AND MULTIMEDIA IMPROVE LEARNING

Both customized teaching and multimedia have been shown to be effective for learning (Lepper et al., 1993; Tversky et al., 2002). Research points to the central role of one-to-one individualized instruction by a peer,

*Corresponding author: E-mail: ivon@cs.umass.edu; http://www.cs.umass.edu/~ivon

parent, teacher, or other more experienced mentor and demonstrates that students learn better when teaching is customized to their learning needs (Greenfield & Lave, 1982; Lepper et al., 1993). Research suggests that novices construct deep knowledge about a domain through interaction with a more knowledgeable expert and one-to-one tutoring provides better learning results than lectures by one-two sigma (Brown et al., 1998; Ericsson et al., 1993; Graesser et al., 1995; Bloom, 1984).

Multimedia produces higher learning results when present in educational software (Mayer, 2001). Graphics and animations can promote inference and discovery by making visible an underlying structure or process; they can represent information beyond text, reduce the need for words, reduce the burden on memory and processing by off-loading some learning; and they can be engaging, attractive and motivating (Perez & White, 1985; Rieber, 1991a; Sirikasem & Shebilske, 1991; Levie & Lintz, 1982; Tversky et al., 2002; Dwyer, 1978; Larkin & Simon, 1987; Mayer 1989). Multimedia is commonly found in commercial educational software (Beal et al., 2002). Yet, intelligent agents used to individualize teaching have not generally taken advantage of the instructional possibilities of dynamic production of multimedia techniques, and multimedia tutors (those using digital sound, animations and interactive graphics) do not typically customize their material for individual students.

Current intelligent tutors model a student's knowledge and have provided effective help (Lester et al., 1997; Mayo & Mitrovic, 2001). However, only preliminary attempts have been made to incorporate knowledge of individual differences (e.g., cognitive skills, gender) and to use this information to customize instruction within a tutor (Shute, 1995; Arroyo et al., 2000, 2004a). The benefits of customized tutoring are especially clear for students with relatively poor skills and for under-represented students in some fields, e.g., minorities and women in science courses. Classroom research on student motivation strongly demonstrates that students with the weakest skills, i.e., those who need help the most, are the least likely to request help from the teacher or classmates (Aleven et al., 2003; Karabenick, 1998). Yet there are some indications that students are more willing to seek help from software than from a human instructor. In the private context of the tutor environment, students with weak skills benefit the most, seem comfortable requesting hints, make more use of help and instruction and demonstrate improved performance (Arroyo, 2004b). This is the reverse of the usual findings in the regular classroom, in which higher achieving students are most likely to request help from teachers and classmates.

The research described here models student group characteristics (e.g., cognitive skills profile, gender) to further optimize the advantages of computer tutors. Explanations and hints are tailored to the needs and skills of individual students, mimicking the approach of human tutors.

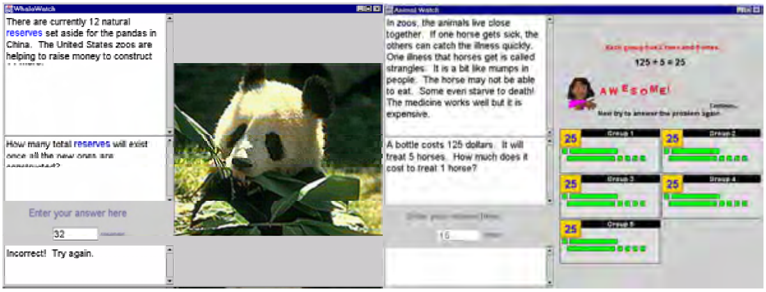
Two forms of adaptation of explanations and hints are used: *micro*adaption to select the content (e.g., problems) to assess or instruct the student and *macro*adaption to select the best type of feedback/assistance for the learner (Shute et al., 2005; Arroyo et al., 2000). Microadaptation, or real-time selection of content in response to a learner's inferred knowledge and skill state, is domain-dependent, e.g., simple skill problems are presented before difficult problems. Decisions about content are based on student performance and subsequent inferences of students' knowledge and skills compared with the level they should achieve when instruction is complete. However all students are placed in fixed classification cells based on pretests or behavior with the tutor, e.g., they know or do not know a skill. If a student incorrectly solves a problem, a tutor might i) present material relating to the *same* concept or skill, ii) administer a slightly *easier* task, to see the extent of the problem, or iii) present additional practice or *remedial* instruction. Macroadaptation or the customization of teaching material based on stable individual learner qualities, e.g., cognitive or perceptual abilities, is dependent on individual student features, not domain topics. The tutor behaves differently with each student based on individual learning style or need. In contrast with microadaptation, it is domain-independent and relates to decisions about the format and/or sequence of the content presented to the learner (Shute 1993).

The research described here developed a cognitive model in connection with *macroadaption*. The cognitive model, a specialized user model that records and infers cognitive learning characteristics, e.g., visual skills, problem solving ability and memory retrieval speed, is based on cognitive information that is usually collected before instruction begins and used to make informed decisions regarding the type of content or instruction best suited for the individual. Material beyond cognitive characteristics, such as perceptual, personality or learning styles may also be collected. The cognitive model might also record observable student activities (number of problems solved correctly, time to respond) and infer unobservable features (cognitive, perceptual, and learning style).

This article describes the integration of both customized questions and multimedia into two intelligent tutors, in which enhanced user models select from a range of problems and help. The goal has been to i) articulate the link between cognitive skills and individual learning, and ii) individualize teaching based cognitive and student models. The results of this research suggest that enhancing user models with detailed information about user cognitive characteristics leads to improved responses in instruction and optimized learning outcomes and that enhanced user models can further optimize the advantages of customized instruction. Section 2 describes the use of cognitive models and macroadaption in AnimalWatch, an intelligent tutor for arithmetic skills. Section 3 describes cognitive models, customization and animated help within Wayang Outpost, a geometry tutor. Section 4 discusses the results of these two studies and future work to customize intelligent tutors.

2 ANIMALWATCH: AN ARITHMETIC TUTOR

We developed and evaluated AnimalWatch, an intelligent tutor for arithmetic designed to encourage positive attitudes towards mathematics among students at the end of elementary school. The software followed an immediate help provision model: when students made an error, the system responded with help, including a text message, a simple hint, or a more extended explanation. The immediate provision of help supported students' beliefs that they could succeed in solving difficult problems through effort and support, and forestalled the self-derogatory attributions reported by female students that they lack the ability to solve difficult problems. When a student made a mistake, as shown in the simple addition problem in Figure 1, the tutor may present either a simple text response (left), or a manipulative hint, shown for a division problem (125/5), in which the student worked with five groups of cuisenaire rods of 25 units each (right), or a traditional numeric procedure (Figure 2, right). We investigated the impact of different types of help in relation to student cognitive developmental level (concrete vs. formal operational level) as diagnosed with cognitive pretests. The tutor contained approximately 1000 mathematics word problem templates and associated instruction for topics from simple addition through fractions and mixed numbers.[1] Problems were adaptively selected depending on the student's cognitive mastery of different skills and taking into account problem difficulty factors.



There are currently 12 natural reserves set aside for the pandas in China. The United States zoos are helping to raise money to construct

How many total reserves will exist once all the new ones are constructed?

32

Incorrect! Try again.

In zoos, the animals live close together. If one horse gets sick, the others can catch the illness quickly. One illness that horses get is called strangles. It is a bit like mumps in people. The horse may not be able to eat. Some even starve to death! The medicine works well but it is expensive.

A bottle costs 125 dollars. It will treat 5 horses. How much does it cost to treat 1 horse?

15

125 ÷ 5 = 25

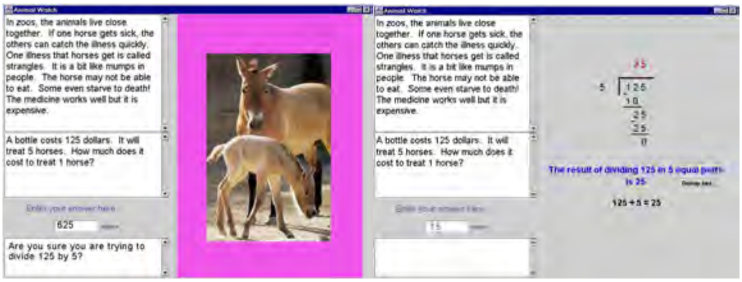
AWESOME!

Now try to answer the problem again

FIGURE 1
AnimalWatch, an Arithmetic Tutor with simple hints (left) and manipulative hints (right).

AnimalWatch word problems focused on endangered species, and as students worked, they learned about a species' history, habitat, environmental threats, and so on (see Figures 1 and 2). Several endangered species were included, e.g., right whale, giant panda and Taki horse. The domain model was arranged as a topic network where nodes represented skills to be taught, such as "least common multiple" or "two column subtraction." Links between nodes represent a prerequisite relationship. For instance, the ability to "add" is a prerequisite to learning how to "multiply" . Other links between nodes represent subskills, e.g. the topic "adding fractions" has the subskills of finding a least common denominator (LCM), converting the fractions to an equivalent form with a new numerator, adding numerators, simplifying the result, and making the result proper. AnimalWatch included a student model that represented the user's knowledge of the target domain (students aged 9-12, Grades 4-6) and selected problems for each student from a database of word problem templates instantiated with appropriate operands, depending on the student's current proficiency level. The student model was maintained in a Bayesian overlay network, and continually updated the tutor's estimate of each student's ability and understanding of the mathematics domain, following a cognitive mastery approach (Corbett, Anderson & O'Brien, 1995) and tailoring the difficulty of problems depending on estimates of mastery of those skills.

As with other intelligent tutoring systems, student mistakes were corrected through individualized multimedia help. Hints were provided immediately when the student entered an incorrect answer. In Figure 2, a student's incorrect answer to a division problem first elicited a simple text



In zoos, the animals live close together. If one horse gets sick, the others can catch the illness quickly. One illness that horses get is called strangles. It is a bit like mumps in people. The horse may not be able to eat. Some even starve to death! The medicine works well but it is expensive.

A bottle costs 125 dollars. It will treat 5 horses. How much does it cost to treat 1 horse?

Are you sure you are trying to divide 125 by 5?

In zoos, the animals live close together. If one horse gets sick, the others can catch the illness quickly. One illness that horses get is called strangles. It is a bit like mumps in people. The horse may not be able to eat. Some even starve to death! The medicine works well but it is expensive.

A bottle costs 125 dollars. It will treat 5 horses. How much does it cost to treat 1 horse?

The result of dividing 125 in 5 equal parts is 25

125 ÷ 5 = 25

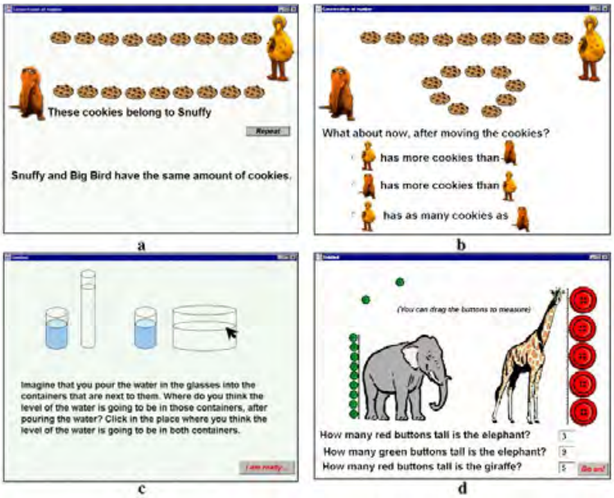
FIGURE 2
Textual and symbolic hints for a division problem.

message "Are you sure you are trying to divide 126 by 5?" (left). Later, the tutor provided richer explanations and multimedia if the first response was not effective, e.g., an interactive symbolic hint in Figure 2 (right) demonstrated a traditional procedure to solve long division operations. This symbolic representation was based on formal algebra rules in contrast to the manipulative in Figure 1 (right) that supported students to manipulate and interact with objects representing numbers.

However, in contrast to traditional intelligent tutoring systems, AnimalWatch included a range of help types and explanations that reflected the cognitive diversity of the target user group: pre-adolescent from 10 to 12 years old. As noted above, students at the transition to middle school show a wide variation in their ability to reason in abstract terms; in fact, many students in Grades 5 and 6 still rely on highly concrete solution approaches, such as counting on fingers (or using other forms of visual counters). Thus manipulative hints were included, e.g., Cuisinaire rods to divide 125 by 5 in Figure 1 (right). The next three sections describe cognitive differences in preadolescents and how they are recorded by AnimalWatch (Section 2.1), the AnimalWatch experimental method (Section 2.2) and results of the AnimalWatch experiment (Section 2.3).

2.1 Cognitive Differences in Preadolescents

Jean Piaget argued that one of the primary characteristics of human cognitive development is a progression from reasoning about one's immediate experience to the ability to reason about abstractions (Piaget, 1953, 1964; Voyat, 1982; Ginsburg et. al, 1998). In AnimalWatch, we analyzed preadolescents' cognitive development using an on-line battery of interactive multimedia tasks based on those developed by Piaget and tested with hundreds of children (Arroyo et al., 1999). Screenshots of several tasks are shown in Figure 3, specifically, number conservation (a and b), substance conservation(c) and measurement and proportionality (d). The tasks that assessed students' proficiency with concrete reasoning included:



These cookies belong to Snuffy

Snuffy and Big Bird have the same amount of cookies.

What about now, after moving the cookies?

○ has more cookies than

○ has more cookies than

○ has as many cookies as

Imagine that you pour the water in the glasses into the containers that are next to them. Where do you think the level of the water is going to be in those containers, after pouring the water? Click in the place where you think the level of the water is going to be in both containers.

(You can drag the buttons to measure)

How many red buttons tall is the elephant? 3

How many green buttons tall is the elephant? 9

How many red buttons tall is the giraffe? 5

FIGURE 3
The Piagetian Pretest in AnimalWatch.

- *Number conservation:* Students observed two identical sets of cookies (each set consisted of nine cookies horizontally aligned) (see Figure 3a). When the elements of one set were moved to form a small circle (see Figure 3b) students were asked whether the number of cookies was altered.
- *Substance conservation:* Students were presented with two identical vessels with the same amount of liquid (see Figure 3c). Each container had another empty one next to it: one was narrow and the other one was wide. Students were asked to determine where the level of water would be in the two empty vessels if the liquid in the two full vessels was poured into them.

- *Area conservation:* Students compared two areas of the same size but different shape.
- *Seriation*: Students ordered a group of pencils from the shortest to the longest one.
- *Class inclusion*: Students determined whether there were more dogs or more animals in a set with different kinds of animals, in which the largest subset was dogs.
- *Functionality*: Students invented an algorithm to solve a problem of ordering pencils by length when they could only see two of them at a time.
- *Reversibility:* Students were shown an animation of three colored balls entering a can from one end, one after the other. They were asked to determine the order in which the elements would come out of the same end of the can.

Three more tasks determined whether the student was at the more developmentally-advanced formal operational or abstract reasoning stage:

- *Establishment of hypotheses, control of variables in experimental design, drawing of conclusions:* Students used a simulation plant growth experiment to identify the conditions of temperature and illumination that would produce optimal plant growth.
- *Proportionality*: Students were shown two animals of different height and given two different measurement system units (large red buttons and small green buttons) (see Figure 3d). They were asked to measure the first animal with the first measurement unit and then again with the second measurement unit. Then they were asked to infer the height of the second animal with the first measurement system. The buttons could be dragged across the screen to measure the animals.
- *Combinatorial analysis*: Students were asked to generate combinations of four switches to open a safe.

2.2 AnimalWatch Experimental Method

The AnimalWatch experiment investigated whether students' responses to different forms of instruction would vary with their cognitive development as measured by the Piagetian test. We explored possible interactions of student cognitive development, gender and learning. After measuring each student's cognitive development with the Piagetian test described above, students were classified as being in the *concrete* or *early-*

formal operational stage of cognitive development, depending on the number and level of abstraction of the tasks they accomplished.

The tutor was evaluated in realistic classroom settings, with 154 participating students, who attended sixth grade classes in an urban school in Western Massachusetts. These students were characterized by ethnic and economic diversity, including White, African-American, and Puerto-Rican-Latino students. Students ranged from 10 to 12 years old and there were approximately equal numbers of girls and boys. Sixth grade students (instead of lower grade students) were part of the study because the curriculum in the more urban district lagged behind that of other schools and many of the students needed remedial help.

AnimalWatch automatically recorded student interactions with the tutor, including problems and type of hint selected for each student and the responses given by the student. The tutor's adaptive problem selection mechanism ensured that the problem challenge level for each student was fairly consistent: for each student, problems were difficult enough to support a small number of mistakes per problem, the assumption being that students must make errors and correct them in order to learn (Arroyo et al., 2003).

When students entered an incorrect answer, the tutoring system responded with a hint, example, or other form of help. Initially, a simple text message (e.g., "Try again", "Are you sure you are adding 35 and 75?") was delivered. In order to meet the individual needs of students at different stages of cognitive development, a variety of hint types were provided (see Figures 1 and 2) until the student was eventually shown the correct answer along with the solution steps required.

Two versions of the tutor were available based on two different help treatments; one version presented concrete hints, e.g., manipulative and graphics, and the second provided formal hints e.g., traditional numeric procedures. Each student was randomly assigned to one of two versions of the tutor; different versions varied only in the kind of help provided. Alternative problems and hints were provided and then correlated with the cognitive development and gender of the student. Students were classified into *concrete* (N = 75) and *formal-numeric* (N = 79) categories based on their Piaget test scores. Students completed a survey of attitudes towards mathematics (Eccles et al., 1993) along with the Piagetian task battery and then used one of the two versions of AnimalWatch for about 3 hours over the course of several class sessions. Finally they completed a posttest of mathematics attitudes. We did not assign mathematics pre- and posttests for this tutor. However, we did use a learning metric (Section 2.3.2) to calculate the mistake reduction or the decline in errors observed on several problems of the same topic and difficulty level.

2.3 Results of the AnimalWatch Experiment

The AnimalWatch experiment was designed to determine whether enhancing user models with detailed information about a student's cognitive ability would lead to improved response to instruction, especially in the cases which led to well-established group differences, such as gender differences in mathematics. Another goal was to determine whether enhanced user models would further optimize other advantages of customized instruction. The results are described in terms of the impact of cognitive development on student performance (Section 2.3.1), the interaction of student learning (mistake-reduction) and cognitive level (Section 2.3.2) and observed gender effects in AnimalWatch (Section 2.3.2)

2.3.1 Cognitive Development and Student Performance

We investigated students' performance on AnimalWatch as a function of their level of cognitive development. The distribution of correct responses to the Piaget test was close to normal (min=1, max=10, S.M.=5.9, S.D.=1.77). In general, students tended to enter more correct responses at the concrete tasks (sample mean for correctness of concrete tasks = 0.62, S.M. for correct formal tasks = 0.2; paired samples t-test: $p<0.001$). As expected, students who accomplished only a few correct Piagetian tasks (considered concrete-operational students) succeeded only at concrete operational mathematical tasks, failing at formal-operational tasks. Students at the high end of the Piagetian tests, those who accomplished more than the average number of correct responses to the Piagetian tasks (considered early formal to formal operational students) succeeded at most of the concrete tasks and demonstrated at least some ability to solve problems requiring hypothetical thought, combinational and/or proportional reasoning. The Piagetian measure was a good predictor of mathematics ability as measured with a pencil and paper pretest in previous years (Pearson Correlation R=0.51, $p<0.001$), and there was a significant correlation for the re-testing of cognitive ability after using the tutoring system (Pearson Correlation R=0.7, $p<0.001$) indicating that the test is reliable. In addition, students diagnosed at the higher formal-operational level made fewer errors in the arithmetic problems than students at the concrete-operational stage. For instance, students considered formal-operational made 0.86 mistakes on average in the multiplication of 2-3 digits problems, while students considered concrete-operational made 1.56 mistakes on average in the same kind of problems. This can also be noted in the graphs for the subtraction topic shown in Figure 4 (students who are concrete-operational started off making more mistakes than formal-operational students).

These last results provide evidence of the instrument's validity and reliability. Because both cognitive tests and mathematics have to do with the manipulation of abstractions (logico-mathematical thinking), it is reasonable that both the mathematics test and Piagetian test outcomes are correlated. Because concrete operations should be easier than formal tasks, it is reasonable that students accomplish more concrete tasks than those that require formal and systematic thinking. Because re-taking the test gives similar outcomes, we conclude that the instrument is reliable.

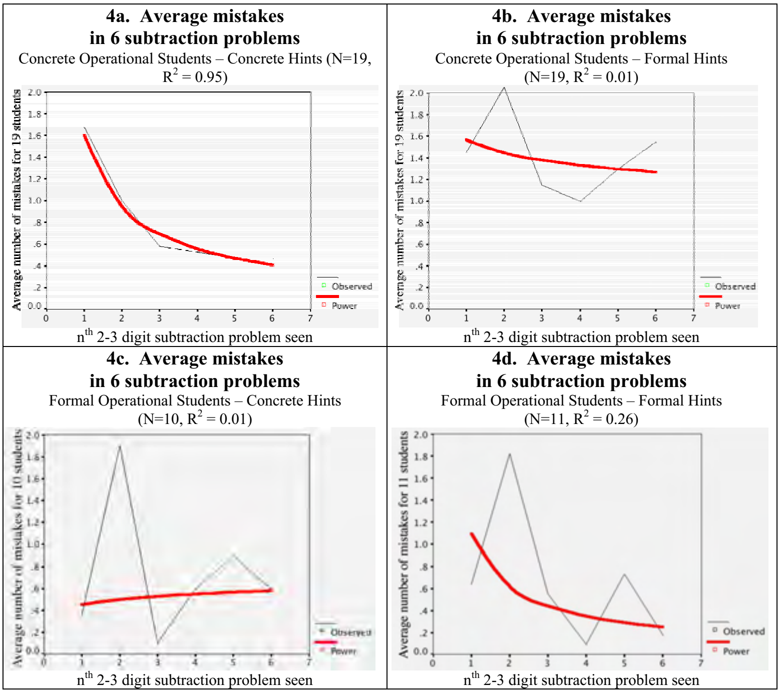
2.3.2 Interaction of Mistake-Reduction and Cognitive Level

It is reasonable to expect that the number of a student's mistakes will decrease with time for problems that involve *similar difficulty*, and whose solution requires the application of similar skills (Corbett, Anderson & O'Brien, 1995). We analyzed students' learning by taking into account their improvement in subsequent similar problems for the same topic and difficulty level, e.g., 'easy addition' or 'difficult fraction subtraction.' This analysis focused on the decline in errors that resulted from a student viewing help and learning from it (learning curves). Mitrovic and colleagues (2002) used a similar measure to analyze how the number of incorrect uses of a skill decreased as a function of the nth opportunity to use that skill. They showed how this mistake-change behaved as a power curve, reflecting a decrease in mistakes.

We built on this approach and considered learning curves, for groups that varied in their individual level of cognitive development (concrete/formal) and the type of hint (concrete/formal). These variables were measured against the number of errors made by students who saw six problems of a specific topic and level of difficulty, e.g., difficult subtraction. As shown in Figure 4, the mistake-reduction rate for subtraction problems involving 2-3 digit operands was analyzed. The *x*-axis represents the 1st to 6th problem attempted by the student, and the *y*-axis represents the average number of mistakes made in that problem across all students. We selected a subset of students for this analysis: students who made mistakes in the first or the second problem seen of this kind (students who seemed to have some trouble with subtraction), so that we could analyze how the hints might have helped them to improve.

When comparing the learning curves for a group of effective learners to another group of less effective learners, effective learners should: 1) decrease their mistakes faster (indicated by the overall slope of the learning curve), 2) display less randomness in their mistake behavior (as indicated by the fit of the learning curve) and 3) reach lower minimums (the learning curve for the effective learners should be more prone to reach mistakes closer to zero).

Figure 4 shows the regressed power learning curves for the mistake data of the four different groups (2 cognitive levels and 2 help versions) for subtraction problems. The four graphs compare the behavior of students at different cognitive levels who have been exposed to different help treatments. Thin lines connect observed data points, and thick lines represent best-fitting power curves that were fit to the data with regression. Figure 4a shows the learning curve for students classified as concrete learners using concrete help.



4a. Average mistakes in 6 subtraction problems
Concrete Operational Students – Concrete Hints (N=19, $R^2 = 0.95$)

4b. Average mistakes in 6 subtraction problems
Concrete Operational Students – Formal Hints (N=19, $R^2 = 0.01$)

4c. Average mistakes in 6 subtraction problems
Formal Operational Students – Concrete Hints (N=10, $R^2 = 0.01$)

4d. Average mistakes in 6 subtraction problems
Formal Operational Students – Formal Hints (N=11, $R^2 = 0.26$)

Average number of mistakes for 19 students
Average number of mistakes for 19 students
Average number of mistakes for 10 students
Average number of mistakes for 11 students
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0.0
0
1
2
3
4
5
6
7
Observed
Power
n^th 2-3 digit subtraction problem seen

FIGURE 4
Power curves for students learning of whole number subtraction.

Students at the lower cognitive development level (see Figures 4a and 4b) seem to be differentially helped by the type of hints provided. That is, students who receive concrete hints (see Figure 4a) show smoother, tighter curves with more pronounced slopes (reaching lower minimums) compared with students receiving formal hints, shown in Figure 4b. Thus, learning of subtraction problems seems more pronounced for concrete students using concrete help than for concrete students using formal help.

There are fewer cases of formal operational students (see Figures 4c and 4d) because fewer formal operational students had trouble with the first 2-3 digit subtraction problems and thus were not selected for this analysis (N=10 in Figure 4c and N=11 in Figure 4d). We think this is the reason why more randomness is observed for formal than for concrete operational students (see Figure 4a and 4b).

Formal operational students seeing formal hints have a learning curve with better fit (see Figure 4d) than formal operational students seeing concrete hints (see Figure 4c). In other words, formal hints seem more effective than concrete hints for formal operational students. Similar differences were observed for students' performance at subtraction and multiplication problems.

2.3.3 Gender Effects in AnimalWatch

Past results with AnimalWatch indicated that females in particular found highly structured and interactive help helpful, impacting their attitudes towards mathematics (Arroyo et al., 2001). One possible explanation for this is the time spent by females on the hints. Female students spent about 25% more time on hints than did male students (independent samples t-test, $p<0.001$).

We evaluated the null hypothesis that all learning curves should be the same across gender, help types, and cognitive development. We computed a new dependent variable: the average mistake decrease from the first to the sixth problem, for each student's mistake history. This time we considered all students for the analysis, regardless of whether they had made mistakes in the first or second problem of a particular topic and difficulty level (otherwise, the number of cases per group would be too small). In the case of 2-3 digit subtraction problems, we found a significant interaction effect of gender, AnimalWatch help version, and student cognitive development ($N=102$, $F=6.48$, $p<0.001$) in predicting average mistake change. Bonferroni confidence intervals (critical values: $df_{error} = t_{0.01,\ 101} = 2.62$) revealed that for male students with concrete reasoning ($N=15$), the worst rate of improvement was observed when receiving formal-numeric help. In fact, the error rates of these concrete boys were higher than for female students with similar concrete operational thought. In contrast, male students capable of abstract reasoning made good progress with formal-numeric help, decreasing their mistakes significantly more with formal-numeric hints than with concrete help. However, significant differences between formal and concrete-operation females receiving different kinds of help were not observed. One possibility is that females' compensated with further effort because they spent more time on hints, thus making all interactive hints effective. Another possibility is that the number of cases per cell is simply too small for such a large number of partitions (2 genders x 2 cognitive levels x 2 help types). In any case, we found an interaction effect for gender and response to different hint types and learning outcome, indicating that students' responses to different forms of instruction and hint type seemed stronger for male students than for female students, who spent significantly more time on hints. A similar effect was seen for the case of multiplication problems.

3 WAYANG OUTPOST: A GEOMETRY TUTOR

Wayang Outpost, our second tutor, was a Web-based geometry tutor[2] that used real-time multimedia to train students for the mathematics section of the standardized Scholastic Aptitude Test (SAT). These high stakes achievement tests that cover mathematics, verbal and analytic skills have become increasingly important in the U.S.A. and can have a significant impact on students' access to future educational opportunities such as admission to universities and scholarships. New effective teaching approaches are needed to help all students perform to the best of their ability on these tests.

Advances in multimedia have created unprecedented opportunities for a new generation of knowledge-based learning environments, ranging from simple graphics and animation to interactive lifelike pedagogical agents in rich, self-explaining 3D worlds. A vast literature exists on the effect of graphics on learning (e.g., Tversky et al., 2002; Betrancourt & Tversky, 2000; Ferguson & Hegarty, 1995; Hegarty, et al., 1999, 2002; Payne et al., 1992; Rieber, 1991b). Sometimes multimedia environments support next-generation natural language dialogue (Aleven et al., 2004; Graesser et al.,

[2] *Wayang Outpost* may be accessed from http://wayang.cs.umass.edu/.

2003; Lane & VanLehn, in press; Rose, et al. 2001). Animated pedagogical agents have been used to observe students' progress, provide visually contextualized problem-solving advice and play a powerful motivational role (Lester et al., 1997).

Research results about the impact of graphics on learning are mixed. Proponents of using graphics for learning suggest many advantages, including engagement, motivation and attraction (Levie & Lintz, 1982; Rieber, 1991a,b; Sirikasem & Shebilsky, 1991). Visual characteristics of the multimedia can help articulate the spatial elements and relations in the domain (Larkin & Simon, 1987; Tversky, 1995, 2001). However, some graphics fail to have a positive impact on learning, in part because the graphics might be difficult to perceive and understand (Lowe 1999; Slocum et al., 2000; Tversky et al., 2002). In some cases the structure and content of the graphic does not correspond to the desired structure and content of the situation. In the discipline of geometry, an appropriate correspondence between the graphics and the discipline is quite natural, since the problem solving elements (e.g., angles and lines) are strongly linked to similar elements in the graphic.

Mathematics education is one discipline that has benefited from the use of multimedia. Johari (2003) examined the effect of two inductive multimedia programs, one using a coordinate graph and its language, on university students' ability to conceptualize variables and create equations from word problems. Results suggested that inductive multimedia facilitates the incorporation of instructional strategies such as inquiry learning from data, tutorials and schema. The results are consistent with propositions recognizing the conceptual richness of visuals, specifically coordinate graphs, in mathematics education. Researchers using multimedia in mathematics education are cautioned to focus the multimedia on teaching inductive problem-solving strategies or scientific heuristics (for example, by working backwards, working inductively, or applying algebraic thinking to data) (Martinello & Cook, 1999) and to include the language of mathematics and mathematics visuals (graphs) (Kaput, 1992).

The next two sections describe the methods used to develop the Wayang Outpost experiment (Section 3.1) and results of that experiment (Section 3.2).

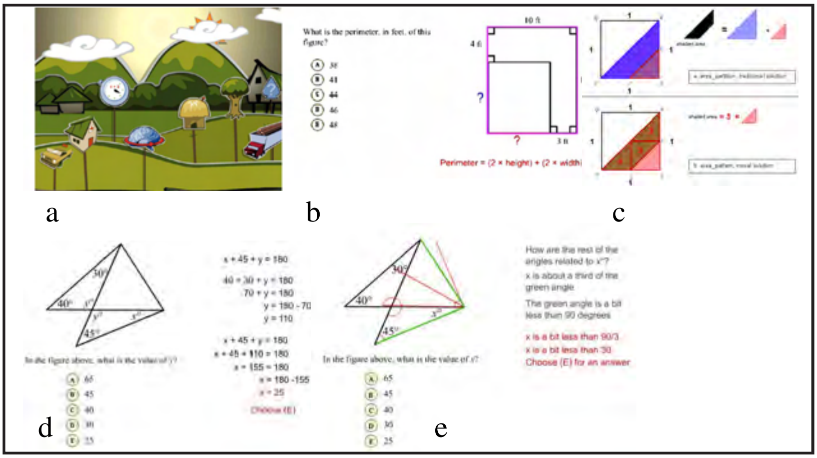
3.1 Methods in Wayang Outpost

Wayang Outpost used digital animation and sound along with information about a student's cognitive skills to customize instruction and improve learning. Wayang Outpost presented an animated classroom situated in a fictitious research station in Borneo, which provided rich real-world content for geometric problems involving endangered species material (see Figure 5).

Before working with the geometry tutor, students' spatial ability, math-fact retrieval skills and geometry knowledge were measured. Then students worked with the tutor to complete SAT problems and engaged in fantasy adventures (see Figure 6). Wayang Outpost provided problems typical of those on high stakes achievement tests that 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.

When students first entered the site they were presented with a variety of activities, as shown in Figure 5a. An orangutan research laboratory in Borneo served as the major environment and the village provided top-level navigation for modules, such as pretests, SAT problems and adventures. Two forms of help were available for each problem, including fairly traditional analytic hints (such as setting up equations) along with visual hints and animated lines that support rapid estimation and novel thinking about geometry.

SAT-mathematics problems were presented with animated characters based on the traditional Indonesian art form of shadow puppetry (Wayang; hence the name of the system). If the student answered incorrectly, or requested help, multimedia explanations provided step-by step instruction


What is the perimeter, in feet, of this figure?

Ⓐ 38
Ⓑ 41
Ⓒ 44
Ⓓ 46
Ⓔ 48

Perimeter = (2 × height) + (2 × width)

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

Ⓐ 65
Ⓑ 45
Ⓒ 40
Ⓓ 30
Ⓔ 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?

Ⓐ 65
Ⓑ 45
Ⓒ 40
Ⓓ 30
Ⓔ 25

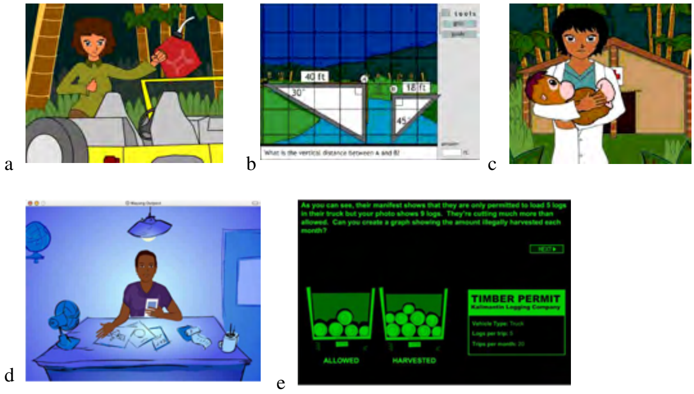
How are the rest of the angles related to x°?
x is about a third of the green angle
The green angle is a bit less than 90 degrees
x is a bit less than 90/3
x is a bit less than 30
Choose (E) for an answer

FIGURE 5
Wayang Outpost provides customized hints in geometry.

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. Explanations and hints 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 figure, and talking, rather than a heavy reliance on screen-based text.

On another problem, a visual hint suggested that the student mentally invert the cut-out portion of a drawing to reveal an intact rectangle, suggesting that the missing lengths are already known (see Figure 5b). The student also calculated the area of a dark central rhomboid (see Figure 5c). The traditional way to solve such a problem is to find the area of the larger triangle and subtract the area of the smaller one (see Figure 5c top). Visual hints instead identified a pattern of three smaller triangles in the shaded area (see Figure 5c bottom) by moving and flipping the triangle at the bottom. Thus, only one triangle area needed to be computed and multiplied by three. Multimedia hints explained this through moving triangles. In the last example, a fairly traditional analytic approach described how to solve for the value of the angle x, by setting up equations (see Figures 5d and e). Visual help provided animated figures for the same problems and provided lines that suggested that a student mentally translate angles to determine the missing values.

Fantasy Adventure Problems. The ultimate goal of the Wayang Outpost was to enhance students' conceptual understanding of mathematical concepts and increase their ability to draw on their own skills to solve novel problems. Therefore, Wayang Outpost also incorporated fantasy adventures that used real-world content and challenging problems requiring multiple steps and skills to solve (see Figure 6). Performance on this component became a measure of transfer of mathematics skills from SAT problems to real-world contexts where the same mathematics skills were needed to solve problems, such as to calculate the amount of roofing material needed to rebuild an orangutan nursery destroyed by a fire (see Figure 6c) and to compute whether a jeep could safely ride across a broken bridge, given that it could withstand a fall of a certain height (see Figure 6b). In other adventures, a scientist asked students to save three orangutans after a fire broke out in the forest (see Figure 6a) and another asked students to monitor over-logging in the forest, specifically to check whether harvested timber was consistent with the allowable amounts (see Figure 6d and e). If a



What is the vertical distance between A and B?

As you can see, their manifest shows that they are only permitted to load 5 logs in their truck but your photo shows 9 logs. They're cutting much more than allowed. Can you create a graph showing the amount illegally harvested each month?

Figure 6. Fantasy component in *Wayang Outpost*.

FIGURE 6
Fantasy component in Wayang Outpost.

student had difficulty in a fantasy problem, a simple geometry problem, not embedded in the real-world context, was presented to remind the student of basic geometry principles.

Animated characters, based on real female scientists, led each virtual adventure. For example, the character based on Anne Russon of York University, an expert on the orangutan, invited students to rescue the orangutans trapped in a fire. A second character based on Lori Perkins, an expert on orangutan conservation and research, led an investigation of illegal logging involving the over-harvesting of rainforest teakwood, flooding and loss of orangutan habitat. In this adventure, students calculated discrepancies between the observed and permitted areas of harvest (see Figure 6e); orangutan habitat area lost to flooding; perimeter distances required to detour around flooded areas; and the amount of travel required to reach areas with cell phone access. Based on survey data, evidence showed that fantasy problems were motivating and engaging, especially for women.

The next two sections describe the background in cognitive studies that led to the WayangTutor experiment (Section 3.1.1) and the student participants and variables evaluated (Section 3.1.2).

3.1.1 Cognitive Skills and SAT-mathematics Performance

The major hypothesis evaluated in the Wayang experiment was that choice of hint type (analytic or visual) would help students develop new strategies for geometry problem solving that would be more effective for their learning style. For example, students who scored high on spatial ability assessment might learn more when receiving a high proportion of hints emphasizing mental rotation and estimation, approaches that are generally more effective in a timed testing situation. Students with poor spatial ability may not understand such hints and might learn more when receiving fewer of them.

Earlier research with AnimalWatch suggested that the assessment of cognitive skills was relevant to the selection of teaching strategies and external representations to improve learning results. For instance, hints that used concrete materials in their explanations yielded higher learning for students at early cognitive development stages, than those that explained the solution with numerical procedures (Arroyo et al., 2000). Similarly, Wayang Outpost functioned as a research test-bed to investigate the interaction of gender and cognitive skills in geometry problem solving and the selection of the best pedagogical approach.

Other research has shown that spatial cognition, mathematics fact proficiency and gender are strong predictors of performance on mathematics achievement tests such as the SAT-mathematics exam (Casey et al., 1995, 1997). Many studies have been performed on gender differences in mathematics, but the results are not consistent. However, small but statistically significant differences are the norm (see Feingold, 1988; Hyde et al., 1990). These between-gender differences are generally quite small compared to variability within each gender. Furthermore, these differences are becoming smaller over time (Linn & Hyde, 1989).

Math-fact retrieval (MFR) is a measure of a student's proficiency with mathematics facts, the probability that a student can rapidly retrieve an answer to a simple mathematics operation from memory (Royer et al., 1999). Proficiency with basic mathematics facts is also known to correlate with gender, although the relations are complex: among the population of high-achieving students, males tend to retrieve basic mathematics facts from memory more rapidly and accurately than their female peers, whereas the reverse is true for males with poor academic skills (Royer et al., 1999).

The Wayang experiment was designed to uncover: i) the relation between geometry SAT-mathematics performance and both spatial ability and MFR, ii) gender differences on both tasks, and iii) whether MFR or spatial ability interacted with the type of hint (analytic/visual) to predict test scores.

3.1.2 Student Participants and Variables

The geometry tutor was evaluated with 149 high school students, ages

13-15 in rural and urban area schools in Massachusetts (Walles, 2005). In each study, students were randomly assigned to two different versions of the system: one version presented spatial help and the other presented computational help. A third (control) group took the pre- and posttest but did not use the tutor. Instead, they were engaged in regular mathematics classroom activities. Students took a computer-based mental rotation test (Vandenberg & Kuse, 1978) and a computer-based test that assessed a student's speed and accuracy in determining whether simple mathematics facts (such as 5x4=20) were true or false (Royer et al., 1999).

The SAT geometry tests provided turned out to be quite difficult for students; they did not do particularly well on the pretests. On average, students correctly answered less than 25% of these items (chance level is 20%, as there were 5 choices). Students' spontaneous comments also reinforced the impression that the domain was very difficult, as did their active use of multimedia help while working with the tutor. On average, students entered one incorrect answer before requesting help. This suggests that, for the most part, they were not randomly clicking answers until they stumbled on the correct response, but were actively trying to learn problem-solving strategies.

Students were tested before they used the geometry tutor. On-line assessment tools based on a standard instrument to measure mental rotation skills (Casey et al., 1997; Vandenberg & Kuse, 1978) measured spatial ability. A separate on-line assessment of the student's proficiency with mathematics facts was made indicating the degree of fluency (accuracy and speed) of arithmetic computation (Royer et al., 1999).

The second form of help, based on spatial transformations and visual estimations, was often rapid and involved fewer computations than the traditional analytic strategy. Although estimation methods may be somewhat imprecise, in the context of multiple-choice assessments such as the SAT-mathematics exam, visual-estimation strategies can lead to a plausible correct answer in a short time and thus reduce the possibility that simple computation errors will result in the wrong answer choice. Thus the tutor taught students to do estimation in order to eliminate options that are obviously incorrect.

3.2 Results of the Wayang Outpost Experiment

Results measured the impact of individual difference variables (spatial ability, math-fact retrieval and gender) on student learning. Learning was measured by the differences in scores between the pre- and posttests.

Student response on these tests fell into one of four categories: correct, incorrect, skipped, and not seen. The latter category indicated that the student did not reach a problem due to lack of time. In addition to pre- and posttest scores, we considered the impact of the tutor on students' willingness to attempt unfamiliar and challenging mathematics problems. The next two sections describe the impact on student learning (Section 3.2.1) and the perceived gender effects (Section 3.2.2).

3.2.1 Impact on Student Learning

Students did learn from the Wayang Outpost tutor, despite the difficulty of the domain. The pre- and posttests included some algebra problems and some skills not tutored by Wayang. Students who used the tutor improved significantly on tutored (geometry) skills from pre- to posttest: the number of correct responses to geometry questions was *M = .35 and M = 3.34 (out of 15)*, respectively, and did not improve significantly on non-tutored skills (algebra). This finding helped to eliminate the possibility that observed benefits associated with a tutor might reflect the general novelty of working with a computer. Students in the control group did not improve significatly from pre- to posttest (they barely improved at all). Thus, the improvement associated with the tutored students cannot be attributed to increased familiarity with the test format, or learning from taking the pretest. The tutor intervention group did not significantly improve on the non-tutored algebra problems, and the control group did not improve from pre- to posttest on either question type. Paired sample comparisons revealed a significant increase in the percentage of correct questions from pre- to posttest, $t(134) = -4.91$, $p < .001$, ($M = 23.16$ and $M = 29.73$, respectively), a significant decrease in the percentage of skipped questions, $t(134) = 3.98$, $p < .001$, ($M = 12.10$ and $M = 4.15$), and no significant change in the percentage of incorrect questions ($M = 64.74$ and $M = 66.12$).

In order to gauge the Wayang tutor's effectiveness in tutoring students of varying ability, students were divided into groups based on their pretest scores using a median split technique, e.g., (low scorer/high scorer). Paired sample comparisons revealed that students who performed poorly on the pretest showed significant improvement from pre to post- test, while those who received higher scores on the pretest showed no significant change from pre to posttest. The system appears to be most effective for those who start off at a disadvantage. These students also made more use of the help features in the tutor. In the private context of the tutor environment, students with weaker skills seem comfortable requesting help and instruction, and their performance improved accordingly.

In the traditional classroom of one teacher and thirty students, higher achieving students are most likely to request help from teachers and classmates and frequent interventions have the greatest benefit for those students who were already doing well to begin with ('the rich get richer' effect). Students with the weakest skills, i.e., those who need help the most, are the least likely to request help from the teacher or classmates (Aleven et al., 2003; Karabenick, 1998). A well-designed intervention in a traditional classroom can raise the floor effect for the lower aptitude students also.

Evidence from the data suggests that students found the visual help more useful or engaging. Students in the visual condition (N=75, *M* = 11.10) viewed more mode-specific hints than their counterparts in the analytic condition (*M* = 7.11). There was a main effect of multimedia help type with respect to the number of problems where all help was seen before a correct answer was given. Specifically, those in the visual condition (*N=75, M* = 13.90) had a higher number of this type of problem than those in the analytic condition (*N=74, M* = 10.75). There was also a consistent trend whereby those in the visual condition viewed more help than those in the analytic condition. Because students had to request help step-by-step, this finding suggests that the visual help may be more intriguing than the more traditional analytic help. Perhaps students were more attracted to the visual help since its strategies are not often presented in the classroom (Aleven, 2001). An alternative explanation for the fact that students in the visual condition spent more time with hints is that visual hints were less precise, i.e., not mathematically precise. Since visual explanations were based on simulation and estimation, student were not familiar with this line of reasoning and thus needed to spend more time with them and to see more.

A repeated-measures ANOVA revealed a significant overall difference in percentage of questions answered correctly from pre- to posttest, F(1,101)=20.20, p<.001. Students showed a 27% increase over their pretest score at posttest time (pretest M=22.60, SD=13.40; posttest (M=28.62, SD=12.53) almost a third better than what they achieved in the pretest; this is also a 6% improvement of the test in a couple of hours.

Additional analyses of specific classrooms revealed increases ranging from seven to 300 percent of the pretest score at posttest time. A second repeated-measures ANOVA revealed a significant change in the percentage of questions skipped with no answer, F(1,84)=15.62, p<.001. Students left significantly fewer questions blank in the posttest (M=5.09, SD=13.45) than in the pretest (M=15.10, SD=24.70). Thus, the Wayang tutor was effective in its goal of improving student performance on SAT-mathematics geometry problems. Students who used Wayang significantly increased in their percent correct and decreased in their percent of skipped questions from pretest to posttest, quite encouraging for the shortest period of time (2-3 hours) students were exposed to it.

3.2.2 Gender Effects in Wayang Outpost

There appeared to be a trend by which males (total hints requested, *M* = 85.17) asked for more help than their female counterparts (*M* = 63.12). Royer et al. (1999) noted that male students show a bimodal distribution: the best students are more often male, but so are the worst performing students. Researchers have noted that some students used help features to search for the correct answer (referred to as gaming the system) and it appears that males were more likely to do this than females.

Males and females did not differ on the online measure of mental rotation. In the past males have generally been found to outperform their female peers on measures of math-fact retrieval speed (Royer et al., 1999), but surprisingly, females significantly outperformed their male counterparts on the measure of math-fact retrieval accuracy. Controlling for all other variables, math-fact retrieval score and pretest score were both significant predictors of posttest performance. Specifically, higher math-fact retrieval scores and pretest scores were individually associated with higher posttest scores. Students who were fast and accurate on the math-fact retrieval task performed better on the posttest than those students who were slow and inaccurate, affecting their actual learning. This suggests that training cognitive skills such as accuracy and speed of retrieval of math-facts (such as multiplication tables, simple addition, etc.) affects not only overall performance in a test, but also actual learning with the tutoring system.

Students attempted more items on the posttest, suggesting that they learned enough to at least try to tackle the challenging items, but were not yet necessarily prepared enough to answer them correctly. Clearly, it will be important to provide students with enough tutoring to ensure that their greater confidence is matched by the required skills.

There were no significant interactions with either gender (p=.46) in predicting overall pre to posttest improvement. Despite the fact that females did not improve more than males, their perception of the system was more positive than for males, and their willingness to use it again was significantly more positive. Based on survey results, females were especially motivated to use the fantasy component, which involved female “role models,” and thought highly of the system.

Results showed that adapting the provided hints to students' basic cognitive skills (matching teaching strategies to students' cognitive strengths) can yield higher learning results (Arroyo et al., 2004a). Table 1 shows the result of a partial correlation between two different indicators of learning (students' perception of learning and pre- to posttest improvement in short-term transfer items, a subset of problems in the test that were similar to problems they saw in the tutor) and students' mathematics fact retrieval (MFR) score and spatial ability score, for the two versions of the system (analytic hints and visual hints). The correlation is partial because pretest score and amount of help requested were accounted for, as both should affect learning. For the case of mathematics fact retrieval, a positive significant correlation was observed between the two learning variables on the left and mathematics fact retrieval *only when students received analytic hints.* This implies that mathematics fact retrieval is a cognitive ability that impacts how much a student learns when they are taught with computational or analytic explanations. Meanwhile, there was a significant negative correlation between student's perception of their own learning and the spatial hints. This is very interesting, as it suggests that students of weak mathematics fact retrieval ability perceive they learn with visual hints, at least more than their counterparts of higher MFR ability. It suggests that students of low MFR report they understand better explanations that rely on teaching approaches that avoid their cognitive weakness (students of low MFR may do better if taught to estimate an answer visually instead of carrying out thorough computation).

Another analysis consisted of an ANOVA that 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 of low math-fact retrieval did not learn much when exposed to computational hints, while they did improve when exposed to highly visual hints (Arroyo et al., 2004a). A similar ANOVA performed only on the male population gave no significant interaction effects between hint type and math-fact

TABLE 1
Partial correlation coefficients of learning variables against students' mathematics fact retrieval and spatial ability (** p<0.001, * p<0.05).

	MFR Score		Spatial ability	
	Analytic Hints N=46	Visual Hints N=55	Analytic Hints N=46	Visual Hints N=55
Learned perception (survey)	.33**	-.25*	.11	.21
Posttest – pretest score	.34**	-.05	.00	-.19

retrieval. These results suggest that in particular girls of low mathematics fact retrieval should be provided explanations that don't capitalize on the speed and accuracy of such basic mathematics facts. Another approach for improving student learning through individualized tutor response is to train students' performance on basic cognitive abilities. We have built systems, for example, to train students on spatial abilities (e.g., mental rotations with 3-dimensional cubes) or math-fact retrieval (e.g., exercises to help students memorize arithmetic facts) (Woolf et al., 2003; Royer et al., 1999). These training activities will be evaluated to see whether they improve performance on geometry problem solving in the future.

4 CONCLUSIONS

This research demonstrated that individual cognitive abilities and gender lead to different interactions with an intelligent tutor. This work significantly extends the traditional ITS learner model to include additional information about cognitive variables, e.g., analyses of spatial ability and math-fact retrieval skills, known to predict mathematics achievement. The results of experiments with two intelligent tutors, AnimalWatch and Wayang Outpost, converge to support the conclusion that enhancing user models with detailed information about user cognitive characteristics can lead to improved instructional response. Incorporating information about young students' cognitive development characteristics allows for more sensitive user models, which in turn can be used to provide qualitatively different kinds of help to students of different cognitive abilities, thus improving learning. This is especially important to consider for domains in which there are well-established group differences, such as gender in the case of mathematics, and individual characteristics that are known to predict learning outcomes.

We found that performance on both spatial ability and math-fact retrieval tasks predicted high school students' scores on a mock SAT-mathematics test; gender differences were observed on both tasks; and that MFR was a stronger predictor than spatial ability of test scores. Students with high and low scores on a mock SAT-mathematics pretest were equally likely to request multimedia help. In past work, students with weak skills may not have requested help because the nature of the help available in the software was not effective for them, e.g., long hint sequences might lead to help avoidance in students with relatively low prior knowledge (Aleven &

Koedinger, 2002). Overall, as evidenced by various help seeking data, students in the visual condition of Wayang Outpost were more engaged by help than those in the analytic condition. The results described in this article suggest that if multimedia help resources are better tailored to students' individual cognitive characteristics, help seeking rates might rise among those students who need it most, i.e., lower achieving students.

This research does not challenge the importance of modeling student knowledge but rather extends it and argues that enhanced user modeling, when accompanied by a wider variety of pedagogical approaches, can further extend the advantages of customized instruction. This approach does require that a large range of qualitatively different problems and hints be available, possibly involving different external representations (e.g. concrete blocks or abstract symbols, visual estimations/transformations or analytic strategies to solve a problem). The pedagogical agent, which determines which problems or hints to present, must have the instructional resources to respond to students with diverse learning needs. For example, when a young student whose reasoning is highly concrete makes a problem solving error, the system must be able to select a highly structured and concrete explanation that includes a virtual manipulative; presenting an explanation that shows the solution with numeric procedures will lead to reduced learning. Concrete base-10 blocks are widley accepted and used in the early grades; however, the novelty relies on the software providing qualitatively different explanations to two students who coexiast within one same classroom or computer-lab.

In Wayang Outpost, the teaching of geometry was enhanced through multimedia and customization of problems based on knowledge of cognitive skills and gender. The tutor was very beneficial for students in general, showing high improvements from pre- to posttest. In addition, recent results indicate that students have significantly higher passing rates in standardized tests than control groups who do not use the tutoring system. Students who performed poorly on the pretest benefited most from the tutor, showing a significant improvement from pre- to posttest. These students made more use of the help features in the tutor. In the private context of the tutor environment, students with weaker skills seem comfortable requesting help and instruction, and their performance improved accordingly.

Girls were especially motivated to use the fantasy component and some studies showed that adapting geometry hints to students' basic cognitive skills yields higher learning results. However, the two studies did not support a unified conclusion about gender effects. An analysis of learning indicators showed that cognitive development and gender do impact learning outcome and many individual gender differences were found in the two studies. Gender differences indicated that sometimes the link between cognitive abilities and type of teaching was stronger for one gender than for the other one. We conclude that it is premature to advocate specific teaching strategies merely based on gender.

These findings have important implications for the design and implementation of future intelligent tutoring systems. For example, the goal of seeding user models with relevant information about a learner's cognitive characteristics strongly impacts the range of instructional resources that must be designed and implemented for an intelligent tutor; if users are not uniform, then the help provided cannot be 'one-size-fits-all.' This approach is consistent with studies of expert human tutors who alternate among different strategies to find explanations that are most likely to be helpful for a particular student (Evans & Michaels, 2006). The results of this research do not imply that the tutor should present a single representational form for hints for a single student. Actually, research shows that solving problems with multiple representations can enhance learning (Ainsworth et al., 1998) and in reality, it is desirable to have students progress to more abstract reasoning. These research results do imply that students with low cognitive skills should be first taught with concrete representations within Interactive Learning Environments such as AnimalWatch. The case of whether further along the way, instruction should be slowly "translated" or "faded" into a more abstract representation should be further investigated. Such approaches seemed effective in one-to-one human tutoring (Resnick, 1982). However, it is nearly impossible to individualize how fast to "fade" between representations within a traditional classroom depending on each student's cognitive development. Now we have the evidence and the possibility to customize such strategies within computer-based tutors.

Several modern approaches are being implemented to improve the customization of tutors. For example, log data and machine learning techniques are being used to learn which teaching strategies have worked in the past and are most effective for individual students (Beck & Woolf, 2000; Woolf et al., 2004; Johns et al., 2005). Currently intelligent tutors use fixed algorithms or rules to classify current students into categories, based on pretests and behavior. Then they customize problems and hints, effectively treating all students equivalently as long as they fall into the same group. Such tutors do not improve over time and do not 'learn' from past experiences with students; rather they treat all students similarly years after the tutor was developed. Newer approaches use the log data of hundreds of previous students to model student performance and discover student skills. Bayesian Networks are often used to learn hidden variables, e.g., skills and motivation, and then to develop a policy to select problems or hints, based on past student records and the current student's individual differences.

In another approach, models of student motivation, attention and engagement are being used to understand and influence the choice of teaching strategy. Such models are developed in a number of ways. For example, on-line tracking of gaze has been used to identify cases in which a student is or is not focused on a problem (Johnson & Beal, 2004). In another approach, inferences are made about a student's engagement, motivation and learning based on observable behavior, e.g., number of hints seen, time spent on hints (Arroyo et al., 2004b). Both approaches may help future learning and may inform us in the design of adaptive tutors to teach students with special needs, including students with learning disabilities, e.g., attention deficit and hyperactivity disorder, or physical limitations, e.g., deafness or immobility. As more students work with adaptive tutors, large and complex data records are being mined to identify student characteristics and refine pedagogical decisions.

The research described in this paper contributes toward both the design and the evaluation of enhanced user models as well as to the effort to develop tutoring systems that improve over time and provide optimized customized instruction. This work advances the state-of-the-art in both knowledge-based learning environments and multimodal interfaces. By achieving significant gains in learning effectiveness, the combination of these technologies can bring about fundamental improvements in learning for students in the classroom as well as for those working independently or collaborating over the Internet.

Acknowledgements

We gratefully acknowledge support of this work from the National Science Foundation through four awards: 1) "Gaining Confidence in Mathematics: Instructional Technology for Girls," Carole Beal, PI, Experimental Projects for Women and Girls Award # HRD-9714757, 1/1/98-12/31/2000; 2) "MPWG: Gaining Confidence in Mathematics: Intelligent Tutors for Girls and Women," Carole Beal, PI, Beverly Park Woolf, co-PI, Experimental Projects for Women and Girls Award, #HRD-9555737, 7/1/96 -6/30/97, 3) "AnimalWorld: Enhancing High School

Women's Mathematical Competence", HRD/EHR #012080, C. Beal, P.I. and B. Woolf & M. Royer, co-PIs; 4) "Learning to Teach: The Next Generation of Intelligent Tutor Systems," REC-0411776, B. Woolf, PI with A. Barto, S. Mahadevan, D. Fisher & I. Arroyo, Co-PIs. 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

Ainsworth, S., Wood, D.J. & O'Malley, C. (1998). There is more than one way to solve a problem: Evaluating a learning environment that supports the development of children's multiplication skills. *Learning and Instruction*, **8**(2), 141-157.

Aleven, V. (2001). Helping students to become better help seekers: Towards supporting metacognition in a cognitive tutor. Paper presented during a workshop organized by the NSF/DFG-sponsored German-USA Early Career Research Exchange Program: Research on Learning Technologies and Technology-Supported Education, Tübingen, Germany.

Aleven, V. & Koedinger, K. R. (2000). The need for tutorial dialog to support self-explanation. In C. P. Rose & R. Freedman (Eds.), *Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium.* Technical Report FS-00-01. Menlo Park, CA: AAAI Press, pp. 65-73.

Aleven, V. & Koedinger, K. R. (2002). An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science,* **26**(2), 147-179.

Aleven, V., Stahl, E., Schworm, S., Fischer, F. & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research,* **73**(3), 277-320.

Aleven, V., Ogden, A., Popescu, O., Torrey, C. & Koedinger, K. (2004). Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In J. Lester, R. M. Vicari & F. Paraguaca (Eds.), *Intelligent Tutoring Systems: 7th International Conference, ITS 2004*. Berlin: Springer, pp. 443-454.

Arroyo, I., Beck, J., Schultz, K. & Woolf, B. (1999). Piagetian psychology in intelligent tutoring systems. *Proceedings of the 9th International Conference on Artificial Intelligence in Education,* pp. 600-602.

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*, Montreal.

Arroyo, I., Beck, J. E., Beal, C. R., Wing, R. E. & Woolf, B. P. (2001). Analyzing students' response to help provision in an elementary mathematics Intelligent Tutoring System. Workshop on Help Provision and Help Seeking in Interactive Learning Environments, *10th International Conference on Artificial Intelligence in Education*. San Antonio, TX.

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*, pp. 368-370.

Arroyo, I., Beal, C. R., Murray, T., Walles, R. & Woolf, B. P. (2004a). Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests. *Intelligent Tutoring Systems, 7th International Conference, ITS 2004,* Maceiò, Alagoas, Brazil, Proceedings. Lecture Notes in Computer Science 3220. Springer, pp. 468-477.

Arroyo, I., Murray, T., Woolf, B. P. & Beal, C. R. (2004b). Inferring unobservable learning variables from students help seeking behavior. Intelligent Tutoring Systems, *7th International Conference, ITS 2004,* Maceiò, Alagoas, Brazil, Proceedings. Lecture Notes in Computer Science 3220 Springer, pp. 782-784.

Baker, R.S., Roll, I., Corbett, A. T. & Koedinger, K. R. (2005). Do performance goals lead students to game the system? *Proceedings of the International Conference on Artificial Intelligence and Education (AIED 2005),* pp. 57-64.

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

Beck, J. & Woolf, B. (2000). High-level student modeling with machine learning. Proceedings of the 5th International *Conference on Intelligent Tutoring Systems*, Springer. pp. 584-593.

Betrancourt, M. & Tversky, B. (2000). Effects of computer animation on users' performance: A review. *Le travail humain*, **63**, 311-329

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4–16.

Brown, A. L., Ellery, S., and J.C. Campione (1998) Creating Zones of Proximal Development Electronically. In Thinking Practices in Mathematics and Science Learning. James Greeno and Shelly V. Goldman (Eds). Mahawah, NJ: Lawrence Erlbaun. 341-368.

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

Casey, N. B., 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.

Corbett, A.T., Anderson, J. R. & O'Brien, A.T. (1995). Student modeling in the ACT Programming Tutor. In P. Nichols, S. Chipman & B. Brennan (Eds.) *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum, pp. 19-41.

Dwyer, F. M. (1978). Strategies for Improving Visual Learning. State College, PA: Learning Services.

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development,* **64**, 830-847.

Ericsson, K. A., Krampe, R. T. & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review,* **100**, 363-406.

Evans, M. & Michaels, J. (2006). *One-on-one Tutoring by Humans and Machines*. Mahwah, NJ: Erlbaum

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, **43**, 95-103.

Ferguson, E. L. & Hegarty, M. (1995). Learning with real machines or diagrams: application of knowledge to real-world problems. *Cognition and Instruction*, **13**, 129-160.

Ginsburg, H., Klein, A. & Starkey, P. (1998). The development of children's mathematical thinking: connecting research with practice (Chapter 7). In the *Handbook of Child Psychology*. Fifth Edition. Volume 4: *Child psychology in practice*. Volume editors: Seigel & Renninger. John Wiley & Sons, Inc.

Graesser, A. C., Person, N. & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9*, 359-387.

Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A. & Person, N. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? In U. Hoppe & F. Verdejo & J. Kay (Eds.), *Proceedings of Artificial Intelligence in Education*. Amsterdam: IOS, pp. 47-54.

Greenfield, P. M. & Lave, J. (1982). Cognitive aspects of informal education. In D. Wagner & H. Stevenson (Eds.), *Cultural perspectives on child development.* San Francisco: Freeman, pp. 181-207.

Hegarty, M., Quilici, J., Narayanan, N. H., Holmquist, S. & Moreno, R. (1999). Multimedia instruction: lessons from evaluation of a theory-based design. *Journal of Educational Multimedia and Hypermedia,* **8**(2), 119-150.

Hegarty, M., Narayanan, N. H. & Freitas, P. (2002). Understanding machines from multimedia and hypermedia presentations. In J. Otero, J A. Leon & A. Graesser (Eds.), *The Psychology of Science Text Comprehension*. Hillsdale, NJ: Lawrence Erlbaum, pp. 357-384.

Hyde, J. S, Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Johari, A. (2003). Effects of inductive multimedia programs in mediating word problem translation. Journal of Instructional Psychology, 30(1), 47-68.

Johns, J., Jonsson, A., Mehranian, H., Arroyo, I., Woolf, B. P., Barto, A., Fisher, D. & Mahadevan. S. (2005). Evaluating the feasibility of learning student models from data. Proceedings of the Workshop on Educational Data Mining, 12th National Conference on Artificial Intelligence (AAAI-05).

Johnson, W. L. & Beal, C. R. (2004). "Achieving motivational and cognitive outcomes in mathematics using enhanced intelligent tutoring technology." National Science Foundation, Research on Learning and Education program (ROLE).

Kaput, J. (1992). Technology and mathematics. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics and Learning,* (pp. 515-556). New York: Macmillan.

Karabenick, S. A. (1998). Help seeking as a strategic resource. In S. A. Karabenick (Ed.), *Strategic Help Seeking: Implications for Learning and Teaching,* Mahwah, NJ: Erlbaum, pp. 1-11.

Lane, H. C. & VanLehn, K. (in press). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education*.

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science,* **11**, 65-99.

Lepper, M. R., Woolverton, M., Mumme, D. & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools*, Hillsdale NJ: Erlbaum, pp. 75-105.

Lester, J., Converse, S., Stone, B., Kahler, S. & Barlow, T. (1997). Animated pedagogical agents and problem-solving effectiveness: a large-scale empirical evaluation. *Proceedings of the 8th World Conference on Artificial Intelligence in Education*, pp. 23-30, Kobe, Japan.

Levie, W. H. & Lintz, R. (1982). Effects of text illustrations: a review of research. Educational *Communication and Technology*, **30**, 195-232.

Linn, M. C., & Hyde, J. S. (1989). Gender mathematics, and science. *Educational Researcher*, **18**, 17-27.

Lowe, R. (1999). Extracting information from an animation during complex visual processing. *European Journal of the Psychology of Education,* **14**, 225-244.

Martinello, M. & Cook, G. (1999). *Interdisciplinary Inquiry in Teaching and Learning (2E).* Upper Saddle River, NJ: Pearson.

Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology*, **81**, 240-46.

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

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

Mitrovic, A., Martin, B. & Mayo, M. (2002). Using evaluation to shape ITS design: results and experiences with SQL-Tutor. *User Modeling and User Adapted Interaction,* **12**, 243-279. Kluwer Academic Publishers.

Payne, S., Chesworth, L. & Hill, E. (1992). Animated demonstrations for exploratory learners, *Interacting with Computers*, **4,** 3-22.

Perez, E. C. & White, M. A. (1985). Student evaluation of motivational and learning attributes of microcomputer soft. Journal of Computer Based Instruction, 12(2), 39-43.

Piaget, J. (1953). How children form mathematical concepts. Scientific American, November.

Piaget, J. (1964). The Child's Conception of Number. Routledge & Kegan.

Resnick, L. (1982) Syntax and semantics in learning to subtract. In T.P. Carpenter, J.M Moser and T.A. Romberg, eds. Addition and Subtraction: a cognitive perspective. Erlbaum.

Rieber, L. P. (1991a). Animation, incidental learning, and continuing motivation. *Journal of Educational Psychology*, **83**, 318-328.

Rieber, L. P. (1991b). Effects of visual grouping strategies of computer-animated presentations on selective attention in science. Educational Technology, Research, and Development, 39, 5-15.

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K. & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. Proceedings of the International Conference of Artificial Intelligence in Education.

Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. G. & Marchant, H. G. (1999). Math fact retrieval as the cognitive mechanism underlying gender differences in math achievement test performance. *Contemporary Educational Psychology,* **24**, 181-266.

Shute, V. J. (1993). A comparison of learning environments: All that glitters… In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools* (pp. 47-74), Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Shute, V. J., Graf, E. A. & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. PytlikZillig, R. Bruning & M. Bodvarsson (Eds.), *Technology-Based Education: Bringing Researchers and Practitioners Together*, Greenwich, CT: Information Age Publishing, pp. 169-202.

Slocum, T. A., Yoder, S. C., Kessler, F. C. & Sluter, R. S. (2000). MapTime: software for exploring spatiotemporal data associated with point locations. *Cartographica*, **57**, 15-31.

Sirikasem, P. & Shebilske, W. L. (1991). The perception and metaperception of architectural designs communicated by video-computer imaging. *Psychological Research/Psychologische Forschung*, **53**, 113-126.

Tversky, B. (1995). Cognitive origins of graphic productions. In *Understanding Images: Finding Meaning in Digital Imagery*. New York: Springer-Verlag, pp. 29-53.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial Schemas and Abstract Thought*, Cambridge: MIT Press, pp. 79-111.

Tversky, B., Morrison, J. B. & Betrancourt, M. (2002). Animation: can it facilitate? *International Journal of Human-Computer Studies*, **57**(4), 247-262.

Vandenberg, G. S. & Kuse, R. A. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, **47**, 599-604.

Voyat, G. E. (1982). *Piaget Systematized*. Lawrence Erlbaum Associates.

Walles, R. (2005). Effects of Web-Based Tutoring Software on Math Test Performance: A Look at Gender, Math-Fact Retrieval Ability, Spatial Ability and Type of Help. Master's Thesis, University of Massachusetts, Department of Psychology.

Woolf, B. P., Romoser, M., Bergeron & Fisher, D.L. (2003). 3-Dimensional Visual Skills: Dynamic Adaptation to Cognitive Level, International Conference on Artificial Intelligence and Education, Sidney Australia, July 2003. pp 515-517.

Woolf, B., Barto, A., Mahadevan, S. & Fisher, D. (2004). "Learning to Teach: The Next Generation of Intelligent Tutor Systems." National Science Foundation, Research on Learning and Education program (ROLE).