Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism

Ivon Arroyo[1,2], Joseph E. Beck[1,2], Beverly Park Woolf[1,2], Carole R. Beal[3], and Klaus Schultz[2]

[1] Department of Computer Science, University of Massachusetts, Amherst, MA 01003, U.S.A.
[2] School of Education, University of Massachusetts, Amherst
[3] Department of Psychology, University of Massachusetts, Amherst
{Ivon, Beck, Bev}@cs.umass.edu, cbeal@psych.umass.edu

Abstract. We have built empirical models of elementary-school students' behavior from analyzing student interaction with a mathematics tutor with the objective of building teaching policies for individually different students. This model incorporates external information about the student, namely cognitive development and gender. It also incorporates hint features, namely the degree of interactivity and symbolism of each hint given. We found that boys benefit better from non-interactive and low-intrusive hints, while girls benefit better from highly interactive hints. We found that low symbolic hints are more effective for low cognitive ability students than highly symbolic ones, and the opposite happens for high cognitive ability students.

1 Introduction

Plenty of research effort has been devoted to finding optimal teaching strategies for *all* students while making tutoring decisions in ITS. For example, the PACT algebra tutor has been evaluated with two alternative teaching strategies. In an experimental version of PACT students had to explain their reasoning in addition to entering solutions to problems, while students in a control version just entered a numeric answer. The former version of PACT was found to be more effective than the latter one [1]. As can be seen, research on teaching strategies has been aimed at finding effective teaching strategies for all students. However, there is evidence that some specific teaching strategies are only effective for specific groups of students. For example, [9] concluded that their ISIS inquiry-based science tutor was most effective for high aptitude students, and less effective for low aptitude students.

We want to go beyond these single-teaching-strategy findings by looking at strategies that are effective for *individually different students*. [8] proposes a multiple-method approach to individualization, which involves the design of alternate treatments that engage different groups of students through alternative

educational stimula. This is called *macroadaptation*, as opposed the usual *microadaptation* that consists of a generic and fine-grained kind of adaptivity that depends on the student's progression in the tutor [12]. The traditional microadaptation in an ITS generally consists of a higher estimation of the student's proficiency as the person shows mastery of the topics being tutored.

There has been some work on macroadaptation in Tutoring Systems. In [11] a battery of IQ questions was submitted at the beginning of the SMART tutoring session and four different empirical student models were derived which depended on these IQ scores. These student models provided a high predictive value in determining students' state of knowledge. However, there were no *qualitatively* different treatments for these groups of students. We want to extend this work in four aspects:

1. We want to macroadapt our system to new populations of users. Our population of students is young children instead of adults;
2. Shute's pre-tests were pencil and paper while our pre-tests are computer-based, shown at the beginning of the first session, so that the data are ready for the ITS to be used in tutoring decisions;
3. We are looking at other individual differences instead of IQ. We follow on the cognitive abilities differences by giving cognitive development tests which we consider relevant for students of this age, and also extend it to incorporate gender differences in learning;
4. The alternative treatments we propose are qualitatively different. We have built hints that differ in two dimensions: formalism of the feedback hints (low symbolism vs. high symbolism), and interactivity of hints (high interactivity vs. low interactivity).

In this paper, we want to transmit two main ideas. The first is that macroadapting a tutoring system to individually different students increases the effectiveness of the tutoring system. The second is to show that the specific partitioning of students and hints that we have chosen is a valid and important one.

2 Methodology

We chose to work in the context of a mathematics ITS for elementary-school children, which has proven to be an effective tutoring system. The methodology that we use consists of classifying hints and students along two dimensions, to then analyze the effectiveness of types of hints against groups of students. This section describes the ITS, the classification of hints and students and how we measured hint effectiveness.

2.1 The domain

Animalwatch is an Intelligent Tutoring System that teaches arithmetic to elementary school students. Animalwatch integrates mathematics with the biological sciences. Specifically, math problems are designed to motivate students to use

mathematics in the context of practical problem solving, embedded in an narrative related to endangered species. Animalwatch teaches fractions and whole numbers at a 4th-6th grade level. It provides mathematics instruction for each student based on a dynamically updated probabilistic student model. Problems are dynamically generated based on inferences about the student's knowledge, progressing from simple one-digit whole-number addition problems to problems that involve fractions with different denominators.

2.2 Hints and student categorizations

When a student has trouble solving a problem, Animalwatch initiates a tutoring interaction that helps the student work through the problem. We have built multiple hints to aid on each topic, that were classified along two dimensions. Those dimensions are the degree of hint symbolism and interactivity, which are discussed in section 3.

We have chosen to categorize students along two dimensions: gender and cognitive development. Both these categorizations and the reasons for selecting them are discussed in section 4. Gender is easy to diagnose, but cognitive development is not. We built a computer-based Piagetian test to obtain estimates of cognitive development. A detailed description of this test is given in [2].

2.3 Measuring hint effectiveness: The experiment and data processing

Within an Animalwatch tutoring session, each student goes through a succession of problems. Whenever a wrong answer is entered, a hint is shown. Hints progressively increase the amount of information given. The first hints provide little information, but if the student keeps entering wrong answers, Animalwatch gives hints that will ultimately guide the student through the whole problem-solving process. Hints are given randomly along the two dimensions discussed before, i.e. regardless of the interactivity or symbolism level. Thus, if there were four hints to be picked that could be categorized as: highly interactive-highly symbolic, highly interactive-low symbolic, low interactive-highly symbolic, low interactive-low symbolic, Animalwatch picks randomly one of these four hints.

We analyze how the number of mistakes the student has made changes from problem to problem after seeing a particular kind of hint. Suppose some student gets a subtraction problem. The student answers incorrectly and after a sequence of unsuccessful hints and re-trials is finally given a hint of type Z. Immediately after that, the student enters the correct answer having made a total of X mistakes in this problem. The student then gets a new problem on subtraction of whole numbers. In this new problem, the student makes a total of Y mistakes. We take the difference X-Y as a measure of the effectiveness of hints of type Z, which represents how the number of mistakes is reduced after seeing a hint of type Z (see figure 1).

We look for main and interaction effects for gender, cognitive development, hint interactivity and hint symbolism in predicting hint effectiveness via an

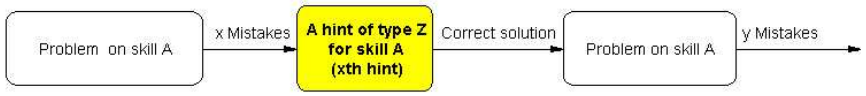

Fig. 1. A case with X-Y difference of mistakes, which is our measure of hint effectiveness

ANOVA. However, it is important to note that other variables could affect our measure of hint effectiveness. For example, a hint would have different effects depending on the difficulty of the before-hint and after-hint problem. In addition, our measure could vary depending on the proficiency of the student at the skill when she saw the hint. The hint could also produce different effects depending on the amount of information that it provides (which also implies that it would be selected at a very specific moment with respect to other hints). We thus perform an analysis the variance with these last variables as covariates, in order to account for their effect.

3 Categorization of hints

Hints were classified along two dimensions: their degree of interactivity and their degree of symbolism. This section discusses this partitioning.

Hint interactivity. We categorize each hint as being highly interactive or low interactive. A synonym of highly interactive hints is "learning-by-doing" hints and a synonym for low interactive hints is "learning-by-being-told" hints (see table 1). Both high and low interactive hints provide plenty of information. However, while a highly interactive hint asks the student for numerous and various kinds of input at each step (dragging and dropping, large amounts of textbox input, etc.), the low interactive hints' interaction involves at most pressing a button to step through an animation, or entering one single number into a text box, or just reading a message. Low interactive hints are also less intrusive and faster to go through, as they require less input from the student. Figure 2 gives examples of high and low interactive hints.

Hint symbolism. The second dimension is the level of numerical symbolism that each of the hints has. We found a way to explain each problem-solving process with two alternative hints: a highly numeric (highly symbolic) one and a concrete (low symbolic) one. Concrete hints involve the use of base-10 blocks for whole number problems and bars that can be partitioned for fraction problems, while highly symbolic hints involve a more abstract procedure that involves direct operations over numerals. We consider operations over numerals as being of a higher level of abstraction because each numeral represents one or more of the concrete objects that are manipulated in the concrete hints. Symbolic hints do not make a connection with real life objects, while concrete hints do. Symbolic hints provide students with powerful tools to reach solutions that can

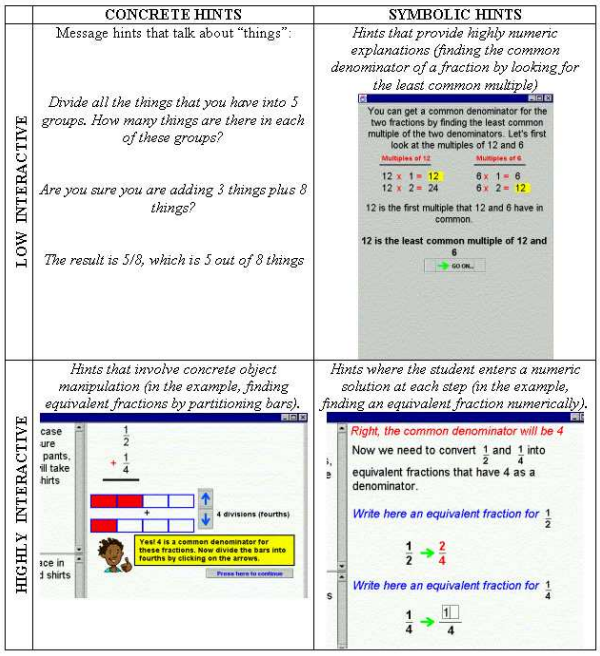

Fig. 2. classification of hints into four categories

be generalized easily to problems with big numbers. Low symbolic hints will be referred as "concrete hints", while we will refer to highly symbolic hints as just "symbolic hints" (see figure 2).

4 Categorization of students

Students were classified along two dimensions: their gender and their level of cognitive development. This section discusses this partitioning.

4.1 Cognitive development

It is known that 5th grade students are at an age of transition from handling concrete to formal operations [10]. It is known that although students develop specific cognitive abilities at an average age, not all students do it exactly at the same time [5]. All this made us believe that we are very likely to be tutoring students that can handle different levels of abstraction in our math domain although they have the same age. We hypothesized that the concreteness or abstractness of the ITS's help system could make a difference in how much they understood and learned. We built a computer-based cognitive development pre-test to diagnose students cognitive ability [2]. This test evaluates students mastery of

concrete operations (conservation, serializing, reversing, etc.) and formal operations (proportions, experiment design, combinatorial analysis) through a battery of computer-based Piagetian tasks [13]. In various previous experiments we found that this measure was a good predictor of mathematics ability ($R=0.513$, $p<0.000$).

4.2 Categorizing across gender

Extensive research makes us believe that socio-cultural factors can contribute to gender differences in learning when considering students of the age of our population of students. Much research has shown that at the beginning in early adolescence, gender differences exist in math self concept and math utility [6]. Some studies indicate that girls experiences in the classroom contribute to their lower interest and confidence in math learning by the middle school period [4]. Moreover, there is starting to be evidence that girls and boys have different approaches to problem solving [7]. In one of our Animalwatch trials in 1998, we found that girls who were given a version of the tutor with highly interactive and information rich hints performed better than in a version which had only short messages as hints. Meanwhile, the complete opposite happened for boys (they preferred and did better in a version that only gave short messages with scarce information as hints) [3].

5 Experiment and results

In spring of 1999, we experimented with Animalwatch and 60 fifth grade students from a rural area. Students were exposed to 3 one-hour sessions of Animalwatch in the following way: At the beginning of the first session, students went through the cognitive pre-test, and then they started using Animalwatch.

Luckily, girls and boys did not differ significantly in our measure of cognitive development (two-tailed t-test, $p<0.2$), so we could compare girls and boys without their cognitive ability being an intervening factor.

We gathered 5272 cases like the ones discussed in section 2.3. Each case represented a problem that a student couldn't solve correctly from the start, which was finally solved immediately after seeing the hint whose effectiveness we are measuring.

As discussed in section 2.3, we analyzed the effectiveness of different hints for different students by using an ANOVA with four variables as covariates: difficulty of the before-hint problem, difficulty of the after-hint problem, proficiency of the student when she saw the hint, and amount of information/order of selection of the hint. These four variables predicted 64% of the variance of our hint effectiveness measurement. The variable "amount-of-information/order-of-hint-selection" was the best predictor ($F(1,5270)=3352$, $p<0.000$), followed by the difficulty of the first problem ($F(1,5270)=27.66$, $p<0.000$) and the difficulty of the second problem ($F(1,5270)=22.68$, $p<0.000$), and last by the proficiency of the student ($F(1,5270)=15.41$, $p<0.000$).

5.1 Gender main effects and interactions

We performed a 2-way ANOVA for interactivity and gender in predicting difference of mistakes with the four covariates discussed above, and found the following results. First, interactivity by itself did not make a difference in predicting our hint effectiveness measure ($p<0.9$). However, gender did by itself. It seems that girls in general tended to improve more than boys did ($F(1,5270)=21.81$, $p<0.000$). We also found an interaction between hint interactivity and gender in predicting hint effectiveness ($F(1,5270)=17.57$, $p<0.000$). Figure 3.a shows the predictions of our model after accounting for the four variables discussed in previous sections. It seems that boys tended to do a lot worse with highly interactive hints, while girls did better with highly interactive hints. We thus found consistent evidence with our 1998 study: girls again seemed to benefit more from highly interactive hints, while boys didn't, this time regardless of the amount of information provided.

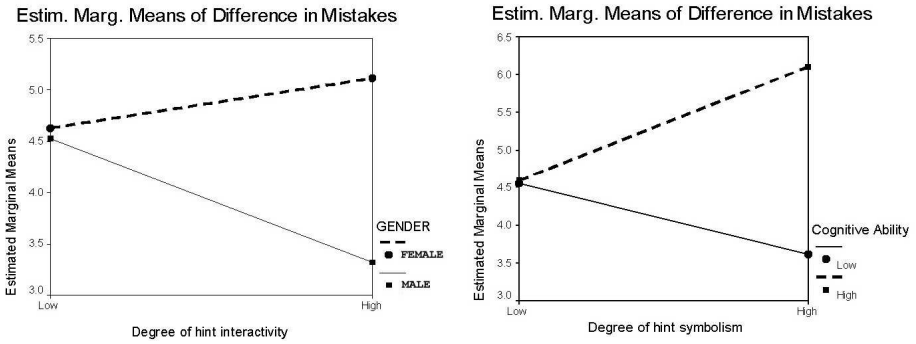

Fig. 3. a) Effects for gender and degree of hint interactivity (left) b) Effects for cognitive development and hint symbolism (right)

5.2 Cognitive development effects and interactions

We analyzed the relationship between symbolism and our cognitive development estimates. We performed a 2-way ANOVA for cognitive ability and hint symbolism in predicting the difference of mistakes with the four covariates. We found no main effect for hint symbolism alone. That is, the degree of hint symbolism was not a significant predictor of our hint effectiveness measure ($p<0.2$). However, cognitive ability was ($F(1,5270)=37.35$, $p<0.000$). This did not surprise us, as we thought it was obvious that higher cognitive ability students could in general benefit more from any kind of help that we gave them. We also found an interaction effect between symbolism and cognitive ability ($F(1,5270)=35.48$, $p<0.000$). Figure 3.b shows how both students with low and high cognitive ability do as well with low symbolic "concrete" hints, but high cognitive ability students do

better with highly symbolic hints while low cognitive ability students do worse with highly symbolic hints.

5.3 Three way interaction effects

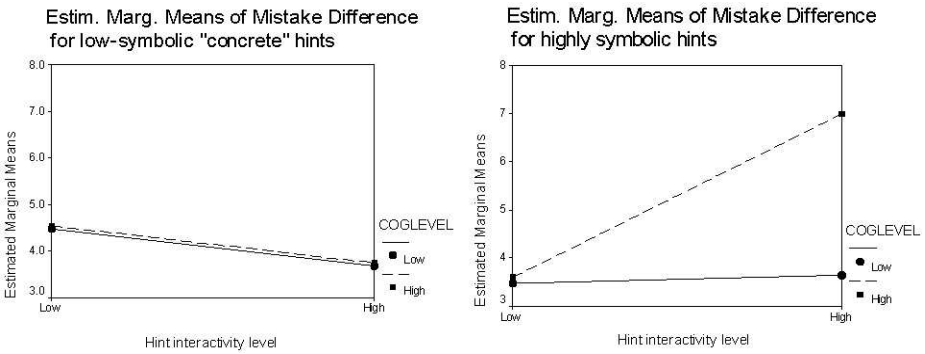

Fig. 4. 3-way interaction effects for interactivity, gender and cognitive ability

Analyzing three-way interaction effects helped us predict 10% more of the remaining unpredicted variance. We found that students of high cognitive ability get excellent results when the high symbolism of a hint is aided by a high interactivity than when it is not. Meanwhile, students of low cognitive ability cannot get to those levels of effectiveness when they are given the same interactive and symbolic condition ($F(2,5269)=5.59$, $p<0.018$, see figure 4).

We also found 3-way interaction effects when taking into account both gender and cognitive ability with respect to hint symbolism.

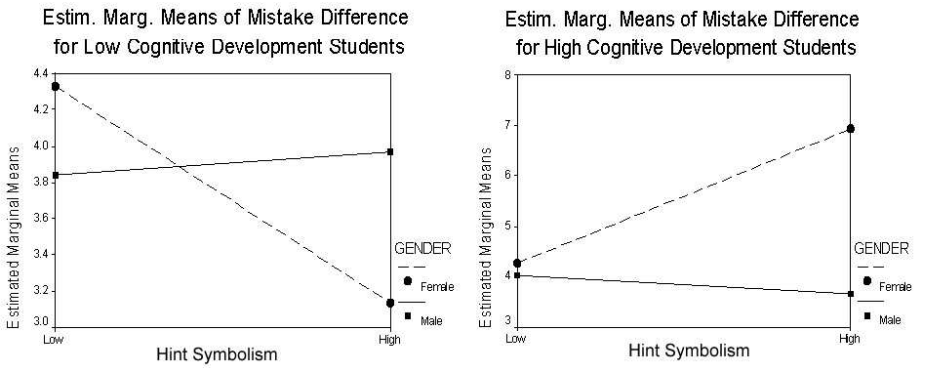

Fig. 5. 3-way interaction effects for symbolism, gender and cognitive ability

In these last graphs we can see how girls behave according to the behavior that we expected. Low cognitive ability girls do better with low symbolic hints and the opposite happens for high cognitive ability girls ($F(2,5269)=10$, $p<0.002$, see figure 5). Meanwhile, the degree of symbolism doesn't seem to affect boys at all compared to girls. One possible explanation is that maybe girls and boys have different goals while using the system. We are assuming that everyone is trying to minimize the number of errors when they get a problem. This way, if a hint were helpful, students would make fewer mistakes in the following similar problem. However, this doesn't necessarily happen if students don't have that objective in mind. Boys might be an instance of such a case. This is an issue that we want to investigate in future trials.

6 Summary and conclusions

In this paper, we have extended the idea of trying to find a correspondence between student types and teaching strategies, expressed as variations of hint features. In this particular case, we considered two specific individual differences and two hint classifications: gender and cognitive differences in the exposure to hints that differed in degrees of symbolism and interactivity.

We build on the philosophy that empirical data is a good source for finding teaching strategies. We introduced a concept of hint effectiveness measured by the decrease of mistakes from one problem to another one after seeing a specific hint type. We found interactions between cognitive development and symbolism that suggest low symbolic hints are more effective for low cognitive ability students than highly symbolic ones, and viceversa for high cognitive ability students. We found interactions between gender and hint interactivity which suggest boys do better with low interactive / low intrusive hints, and that the opposite happens for girls. We also found more sophisticated interactions that involved gender and cognitive abilities mixed together: Boys didnt seem to be very much affected by the degree of symbolism of a hint while girls did. We conclude that when these rules are applied to a tutoring system, its effectiveness should be higher.

We are aware that adaptation implies an exponentially bigger effort than non-adaptation. The more individual differences and hint features we take into account, the exponentially more variations of hints we may need to generate. This could make our tutoring system exponentially larger. However, we also think there are variables that are best differentiators of students' behavior and needs. Those are the variables that we are looking for, which represent important trends of behavior. Moreover, this type of research has two main consequences. It has the descriptive effect of understanding students' thinking better, and the prescriptive effect of deriving successful teaching methods [1].

[1] We acknowledge support for this work from the National Science Foundation, HRD-9714757. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agency.

References

1. Aleven, V. et al.: Tutoring answer explanation fosters learning with understanding. In Proceedings of the ninth international conference of Artificial Intelligence in Education (1999).
2. Arroyo, I.; Beck, J.; Schultz, K.; Woolf, B.: Piagetian Psychology in Intelligent Tutoring Systems. In the Proceedings of the Ninth International Conference on Artificial Intelligence in Education. (1999) pp. 600-602.
3. Beck, J., Arroyo, I., Woolf, B., Beal, C.: Affecting Self-confidence with an ITS. In the Proceedings of the Ninth International Conference on Artificial Intelligence in Education. (1999) pp. 611-613.
4. Beal, C.: Boys and girls: The development of gender roles. New York: McGraw Hill (1994).
5. Case, R.: The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge. Hillsdale NJ: Erlbaum. (1992)
6. Eccles, J.S., Wigfield, A., Harold, R.D., Blumenfeld, P.: Age and gender differences in children's self and task perceptions during elementary school. Child development, 64, (1993) 830-847.
7. Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., Levi, L. W.: A longitudinal study of gender differences in young children's mathematical thinking. Educational Researcher, 27, (1998, June-July) 6-11.
8. Jonassen, D.; Grabowski, B.: Handbook of Individual Differences, learning and Instruction. Lawrence Erlbawm (1993).
9. Meyer, T.N. et al.: A multi-year, large-scale field study of a learner controlled ITS. In Proceedings of the Ninth International Conference on Artificial Intelligence in Education (1999).
10. Piaget, J.: The Child's Conception of Number. Routledge and Kegan (1964).
11. Shute, V.: SMART: Student Modeling Approach for Responsive Tutoring. In User Modeling and User-Adapted Interaction (1995), 5, pp. 1-44.
12. Snow, R. E.: Individual differences and Instructional Theory. Educational Researcher (1977), 6, 11-15.
13. Voyat, G. E.: Piaget Systematized (1982).