Analyzing students' response to help provision in an elementary mathematics Intelligent Tutoring System

IVON ARROYO[1,2], JOSEPH E. BECK[1,2], CAROLE R. BEAL[3], RACHEL WING[3], BEVERLY P. WOOLF [1,2]
IVON@CS.UMASS.EDU (http://www.cs.umass.edu/~ivon), BECK@CS.UMASS.EDU, CBEAL@PSYCH.UMASS.EDU, REWING@PSYCH.UMASS.EDU, BEV@CS.UMASS.EDU
[1]Department of Computer Science
[2]School of Education
[3]Department of Psychology
University of Massachusetts, Amherst, MA 01003

1. INTRODUCTION

There has been a vast amount of research to identify the relationship between different students' learning traits and different educational treatments (e.g. [Jonassen, 93]). In particular, different help provision mechanisms can work differently for students who are individually different. For example, [Luckin, 99] noted how students of different abilities in the same classroom benefit differently from alternative help mechanisms, some of them more child-controlled, other ones more controlled by the system.

Given that we know that individual differences can play an important role in learning, and given that our goal is to build an Intelligent Tutoring System (ITS) that are effective in its tutoring, we hypothesized that cognitive development would be an important individual difference to take into account while providing help. We diagnosed cognitive development according to Piagetian psychology [Piaget, 53] in our population of 8-11 year olds. It is known that students at this age range can handle many of what Piaget called concrete operations (the ability to understand the mathematical nature of concrete objects). Some students are transitioning to a stage where they can handle formal operations (the ability to manipulate abstract ideas). Even though students tend to develop specific cognitive abilities at an average age, not everyone does it at the same age. This makes cognitive development an interesting variable to focus on while looking for individual differences among student-users, and its relationship to tutoring. We diagnosed cognitive development with a computer-based test [Arroyo, 99], and analyzed how students of different cognitive development responded to the abstractness/concreteness of hints.

An immediate goal of our currently funded project is to enhance girls' confidence in their ability to learn mathematics [Beal, 00]. Our proposal is to intervene before students enter high school (8-11 years old, end of elementary and beginning of middle school), through the use of an artificially intelligent multimedia computer tutor for mathematics: AnimalWatch. Because we were interested in gender issues with respect to our tutoring system, we explored how girls and boys reacted to the existing help in the tutoring system. The results we found are interesting and suggest that both cognitive development and gender are important individual difference to take into account when designing a help system.

2. ANIMALWATCH, A MATHEMATICS TUTORING SYSTEM FOR CHILDREN

Animalwatch tutors arithmetic with word problems about endangered species. Thus, it integrates mathematics, narrative and biology. It takes the student through a series of word problems, which dynamically chooses from a large database of word problem templates, which are instantiated with appropriate operands, depending on the student's current proficiency. Animalwatch maintains a probabilistic overlay *student model* and makes inferences about the student's knowledge as he solves problems. Animalwatch increases the difficulty of the problems it gives depending on the student's progress [Beck, 97], going from simple whole number addition problems to others that involve fractions with different denominators. After the student-user logs in, he chooses one out of three stories about different endangered species (also known as *adventures*), and then goes through a sequence of *episodes* in the story. Each problem the student is given has an appropriate mathematical difficulty given the system's estimation of the student's mathematical knowledge. When the student enters an incorrect answer, AnimalWatch gives hints, increasing the level of information provided in the hint as the student continues to make mistakes. In its traditional mode, Animalwatch presents short messages with little information first and more interactive and structured hints later if the former ones could not make the student enter the correct answer. These last hints are more structured, stepped and interactive, guiding the student to get to the correct solution with the aid of multimedia.

3. TRIALS IN SCHOOLS AND ANALYSIS OF RESULTS

We performed three trials in the last four years, all of them trying to get feedback to improve the system and analyzing the benefits and flaws of the help we were providing. The following is a summary of the results we found in each of these studies with respect to the different genders and different cognitive abilities of students.

3.1 Spring 1998: Students with different cognitive abilities and their response to hints

In Spring 1998, we ran a study in a rural area of Massachusetts with 50 5th grade students (9-10 year olds), using the ITS for 3 hours. We wanted to find if the remediation we provided was good, and in what degree it accounted for increases in self-confidence and positive attitudes towards math. We tested two versions of Animalwatch: 24 students used Animalwatch its traditional mode (with multimedia interactive hints which taught the student how to get the answer) and 26 sudents used a version with only short messages as hints (control version). The short messages encouraged the student to re-check their work, and after a couple of failures, told the correct answer to the student. The number of problems that they solved correctly with no help weighed by the difficulty of the problems they solved was our outcome measure of performance in the tutoring system.

We compared the behavior of low and high cognitively developed students, as measured with our test (a cut point of 5.5 from a score in a [0..10] range). We found that low and high cognitive development students did similarly in the experimental version while the low ability students did a lot worse in the control version (see figure 1). This means that the help we were providing was particularly good for those who needed it more. However, high cognitive development students did similarly in both versions, thus implying that help didn't make a difference for them. There is actually a trade-off between how much time students spend seeing hints and how much they progress by solving new problems. If hinting is excessive or inappropriate, this can prevent students from going on to solve new problems and prevents them to progress in the curriculum.

We wondered how much better it would be to produce hints that are appropriate for high vs. low development students. If we could build such hints, we would be able to adapt help provision and make the ITS select one hint or another depending on each student's cognitive development.

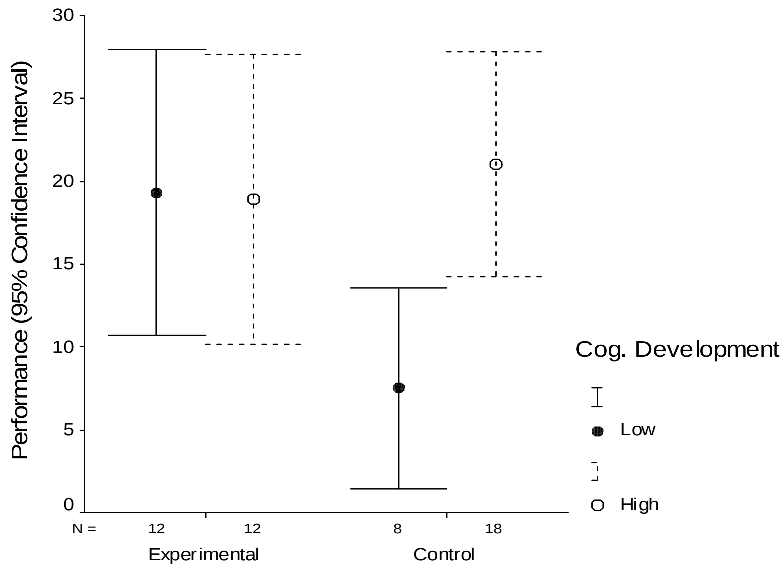

Figure 1. Different cognitive development students' performance in two versions of the ITS

3.2 Animalwatch with concrete vs. symbolic hints

We designed a new set of hints (actually a whole new Animalwatch after changing programming languages). On top of the short messages, we specifically designed two alternative types of highly informative and interactive hints to tutor whole numbers and fractions: highly numeric (highly symbolic) hints and concrete (low symbolic) hints. 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 procedures that are more abstract, involving direct operations over numerals. Physical objects are used in the concrete and conceptual explanations, which involve high manipulation (dragging, dropping, partitioning by clicking, etc.) and are highly visual. While these concrete procedures seem simpler and more intuitive, the manipulation becomes harder as the operands involved get larger.

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. They are highly symbolic and more abstract, focussing on numeric procedures. These explanations may be faster for the student to go through than an equivalent concrete one, though it is harder to make sense of how each step relates to a real life situation (i.e. adding multi-digit numbers in columns, and writing the carried number on top of the column to the left). Symbolic hints provide students with powerful procedural algorithms that can be easily generalized to problems with big numbers (see figure 2).

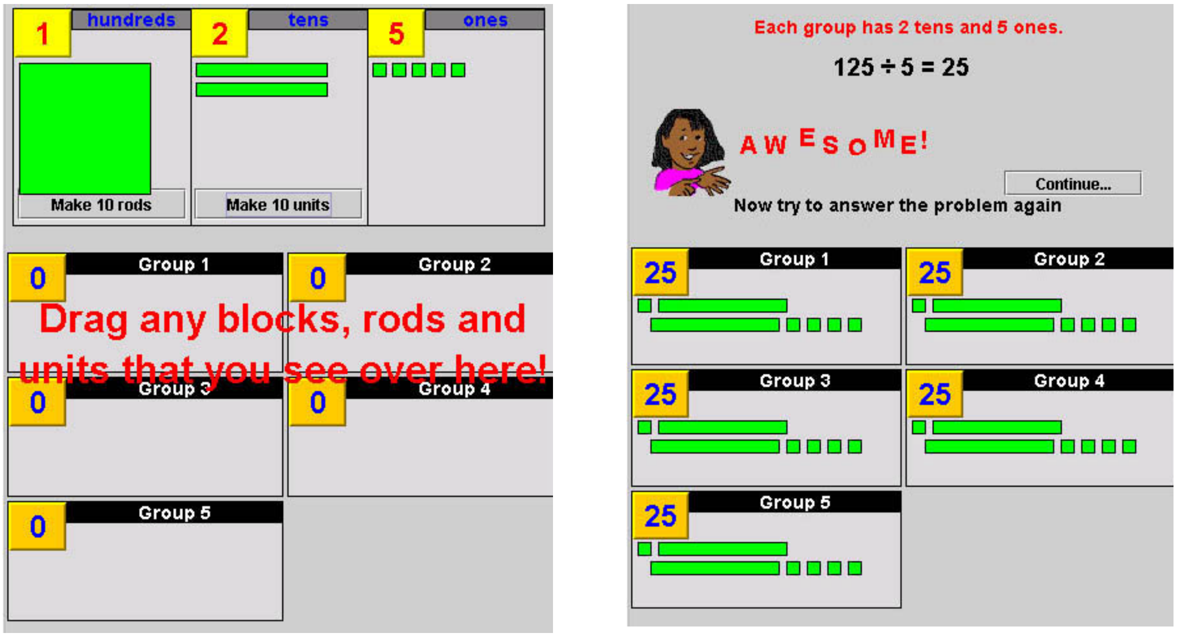

Figure 2a. A concrete hint for dividing 125 by 5

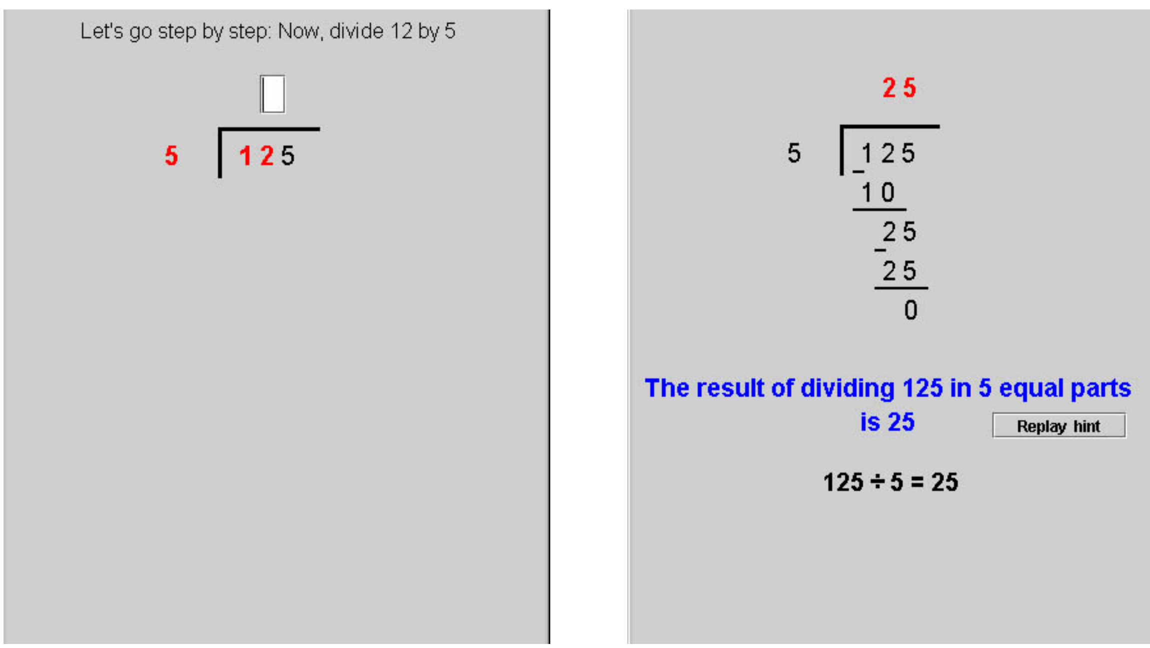

Figure 2b. A symbolic hint for dividing 125 by 5

3.3 Spring 1999: Per hint analysis of symbolism * cognitive level

In spring 1999, we tested the new version of Animalwatch with the new hints. This was not a controlled study like the one done in spring 1998. Instead, all students got all hints, randomly chosen depending on their level of concreteness-symbolism. We then analyzed for the effects for each hint by looking at how the number of mistakes changed from the problem where the hint was seen to the mistakes in a similar problem right after they saw the hint. For more information on how this measure was computed, see [Arroyo, 00].

We analyzed the relationship between symbolism and cognitive development by performing a per-hint analysis. We performed a 2-way ANOVA for cognitive development and hint symbolism in predicting hint effectiveness. We gathered 5272 cases for our hint effectiveness measure, and built a very accurate model (65% of the variance) in predicting hint effectiveness, which depended on the current proficiency of the student, problem difficulty, amount of information provided in the hint, gender, cognitive development, hint interactivity level and hint symbolism level.

An interesting outcome is that we found a significant interaction effect between hint symbolism and cognitive development ($F(1,5270)=35.48$, $p<0.000$). Figure 5 isolates the interaction effect. The figure shows how high cognitive development students do better with highly symbolic hints, and low cognitive development students do worse with such hints. Mistake reduction is higher for high ability students when seeing a highly symbolic hint that when seeing a low symbolic (concrete) one. Meanwhile, low cognitive development students reduce their mistakes more when seeing a concrete hint that when seeing a highly symbolic one.

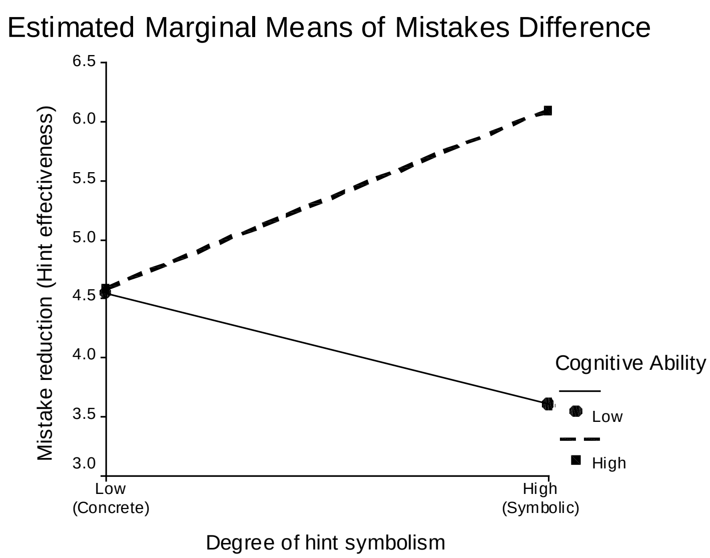

Figure 3. Significant interaction effect between cognitive development and hint symbolism

3.4 Fall 2000: Controlled study measuring symbolism*cognitive level

In fall 2000, we ran a controlled study with a small number of subjects. We tested two versions of the system: one with concrete hints (however, it would give symbolic hints when numbers got too large as the number of concrete objects wouldn't fit in the screen) and another one which only had higly symbolic hints. We will call these two versions "low symbolic" and "highly symbolic". Cognitive development was again diagnosed for the student-users before using the system with the Piaget pre-test. We performed a pre-test and post-test of emotional attitudes towards mathematics and towards one's ability to do mathematics [Eccles, 93]. We didn't perform a mathematics pre and post-test, as we thought that performance within Animalwatch would give us a good idea of how well students did. We will, though, in our next study, as we realized is important to analyze gains independently from the tutoring system. The performance measures we computed for each student are the following:

1. the last proficiency recorded by the system averaged across all mastered topics (finalpro),
2. the mean of all the proficiency values recorded in the student model while the student used the system (avgprof),
3. The number of topics that the student mastered (nmastop),
4. The average number of problems that it took the student to master the topic, for all those topics that the student mastered (hmtilmas),
5. The pre to post-test improvement in self-appreciation of one's ability to do math (scincr),
6. The pre to post-test improvement in the perception of math utility (utilincr),
7. The pre to post-test improvement in math liking (likeincr).

Table 1. Mean performance scores and attitudes' change scores. Best scores are highlighted

Coglevel	Symbolism	N	Finalpro	Avgprof	Nmastop	Hmtilmas	Scnfincr	Utilincr	Likeincr
Low	Low symb	2	**.83**	**.71**	**6.0**	**4.50**	**.08**	**.33**	**.33**
Low	High symb	5	.72	.61	3.6	5.93	-.20	-.11	.02
High	Low symb	6	.90	.75	5.67	4.52	**.09**	.12	.21
High	High symb	6	**.92**	**.82**	**7.0**	**4.43**	.00	**.14**	**.34**

We compared these measures across the four groups. Table 1 outlines the results. The version with low symbolic (concrete) hints outperforms the version with highly symbolic hints for low cognitive development students. Meanwhile, the version with highly symbolic hints outperforms the version with low symbolic (concrete) hints for high cognitive development students, except for changes in self confidence (zero increase instead of 0.9 increase in the concrete version). Because of the size of the groups, this is a merely informative study. However, results are consistent with the prior study, so it is some type of validation to the prior study. We think that giving hints that are more concrete to low cognitive development students and highly symbolic hints to high cognitive development students would increase the overall performance. This would imply a coarse grained kind of adaptation that could aid the traditional microadaptation that is done for each individual student depending on his performance within the tutoring session. We will run a large study in the near future with 200 students to confirm these results at a larger scale.

4. BOYS' AND GIRLS' RESPONSE TO HELP PROVISION IN ANIMALWATCH

It has been found before that the different genders have different attitudes towards problem solving. Because of the nature of our project, we are interested in finding if there is some specific kind of pedagogy that works for girls vs. boys. Is it worth it to consider gender while teaching students in ILEs? This section summarizes the results we found in several trials in schools.

4.1 Spring 1998: Boys' and girls' response to different intensity of feedback

In the spring 1998 study described in section 3.1, we expected that help would be better for all students in terms of the generation of positive mathematics attitudes. However, we found instead that girls and boys had different pre to post-test emotional attitude changes depending on the help mechanism the ITS provided, even though there were no significant performance differences among the two genders. The experimental hint mechanism was more intense, more interactive, and more informative. The experimental version gave more hints to the students when they made mistakes than the control version. Students were randomly assigned to the experimental and the control versions. It is understandable that the apparent overall advantage of the experimental over the control version shouldn't be too big: if help wasn't extremely helpful or given when it was really needed, this should delay students from trying to solve other problems correctly, thus affecting our performance measure.

While analyzing performance in the two versions of the system, we found that boys did somewhat worse in the experimental version, though not significantly worse. Meanwhile, girls did somewhat better in the experimental version, though not significantly better. However, what really surprised us was the emotional impact that the two versions of the system had over students. With respect to changes in self-confidence, we found that boys benefited more from the control (low-hinting) version, while girls' self-confidence was almost not changed by any of the two versions of the system. Moreover, boys' self-confidence was harmed by the hint intense version. Boys also seemed to be harmed in measures of math liking and math utility in the experimental version. Meanwhile, this did not happen for girls in the experimental version. Girls' self-confidence didn't improve, but it was initially very high, so there could be a ceiling effect. We decided to keep these results in mind, but not draw any conclusions. The fact that boys seemed to have done better with less help astonished us.

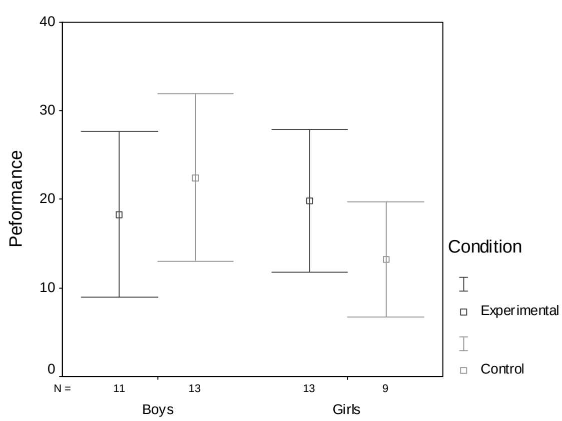

Figure 4: Differences in performance

Table 2. Performance across groups

Gender	Condition	N	Performance
Girls	Control	10	13.10
Girls	Experimental	13	19.85
Boys	Control	16	19.25
Boys	Experimental	11	18.27

Table 3. Changes in self-confidence (student's perception of their own math ability)

Gender	Condition	N	Before	After
Girls	Experimental	13	5.4	5.41
Girls	Control	10	4.83	4.88
Boys	Experimental	11	5.29	4.98
Boys	Control	16	5.26	5.62

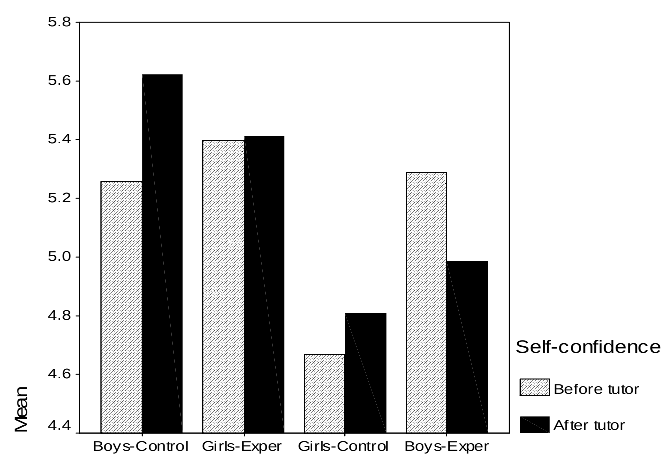

Figure 5. Changes in self confidence across groups

4.2 Spring 1999: Boys' and girls' response to hint interactivity

In spring 1999, we tested the new version of Animalwatch with a new set of hints. As mentioned in section 3.3, these hints were very different from each other. In order to describe them and look

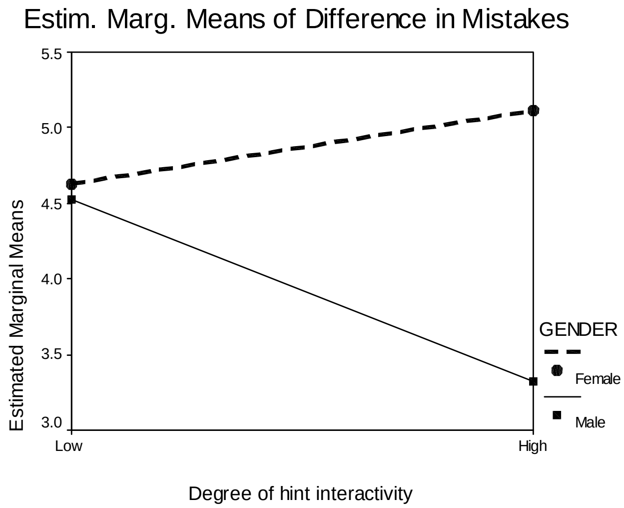

Figure 6. Interaction effect between gender and hint interactivity

for what features in the hints were better for whom, we classified the hints along many dimensions: amount of text, amount of interactivity, concreteness, amount of information, etc. In an exploratory kind of analysis, we found that hint interactivity was an important factor that influenced effectiveness for the different genders.

The spring 1999 study was not controlled, as all hints were given to all students regardless of their interactivity level. Highly interactive hints also generally provided more information about how to reach the solution compared to other hints. Less interactive hints consisted of short messages or some kind of hint where they only needed to click on a button in order to get through it. We analyzed for the effects for each hint by looking at how the number of mistakes changed from the problem where the hint was seen to the number of mistakes in a problem after they saw the hint. For more information on how this measure was computed, see [Arroyo, 00].

We found that girls benefited better from highly interactive hints, and that boys benefited better from low interactive hints ($F(1,5270)=17.57$, $p<0.000$). From this study, we could build an accurate model about how students behaved within this version of the tutoring system. Figure 6 isolates that interaction effect after accounting for other intervening variables.

4.2 Fall 2000: Boys' and girls' response to hints

As mentioned before in section 3.4, we performed a new study in fall 2000, which controlled for the level of symbolism of hints. The prior studies made us think that the time the student was required to put while receiving feedback could have something to do with how help was effective for boys and girls. A third version of the system was used in the experiment, which gave *both* highly symbolic and low symbolic hints: whenever the first hint chosen had failed[1] the alternative hint was given to the student (the one with the other symbolism level). Thus, this version of the

[1] The hint is considered to fail because the student entered an incorrect answer even after seeing the hint.

system gave more hints to the student before giving out the right answer. We can compare the performance of the group of students who used the "high hinting time" version against the students who used the low symbolic version and the high symbolic version grouped together, which had "lower hinting time", as they selected either a low or high symbolic hint, but never both. We found that there was no consistent benefit for girls' in either the low or high hinting versions, but there was an advantage for boys in the low hinting version (see table 4). The increase in self confidence for boys was the only one which was slightly better in the high hinting time version, but still the low hinting time version is positive.

Table 4. Mean performance scores and attitudes' change scores. Best scores are highlighted

Gender	Help system	N	Finalpro	Avgprof	Nmastop	Hmtilmas	Scnfincr	Utilincr	Linkincr
Boy	Low hinting time	7	**.99**	**.90**	**6.43**	**5.29**	.10	**.10**	**.48**
	High hinting time	14	.88	.72	5.86	6.64	**.12**	.00	.05
Girl	Low hinting time	9	**.78**	.68	5.44	**6.0**	**.11**	**.11**	**.04**
	High hinting time	16	.75	**.71**	**5.88**	**6.0**	-.02	-.06	-.10

We were curious to see how much time boys and girls spent in any hint. While comparing the average spent by the children in any one hint, we found that boys actually spend significantly less time in a hint, daring to risk a new answer before girls do (figure 7). The fact that boys are not eager to spend as much time as girls in hints supports the result that boys are less willing to receive long periods of feedback than girls, something which is even noticeable when observing students interact with the ITS.

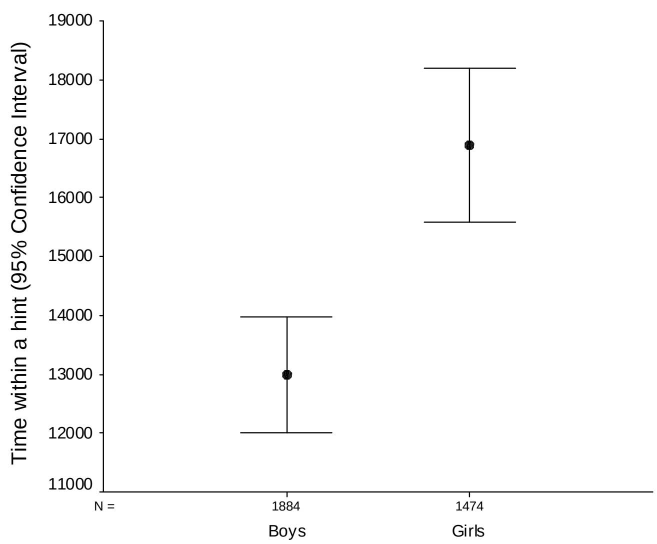

Figure 7: Average time spent by boys and girls within a hint, in milliseconds

5. INTEGRATING GENDER AND COGNITIVE EFFECTS

Given the results of the prior sections, what type of help should we give to a student with a certain gender and a specific cognitive development? We know that giving both hints one after the other is not good for boys at all, and not clearly better for girls. Then, probably we should pick just one, either concrete or symbolic, depending on student's cognitive development, and give the answer if the student fails again. While comparing gender and the high-low symbolism of hints, we were not expecting a gender difference, given that interactivity seemed to be an issue for the different genders but not symbolism. Still, we compared them, and obtained the results in table 5.

Table 5. Mean performance scores and attitudes' change scores. Best scores are highlighted

Gender	Help system	N	finalpro	Avgprof	Nmastop	Hmtilmas	Scincr	UtIncr	Likeinc
Boys	Concrete	3	.83	.78	4.67	4.46	**.20**	.07	.33
Boys	Symbolic	4	**.93**	**.84**	**7.75**	**4.31**	.09	**.27**	**.55**
Girls	Concrete	5	**.91**	**.73**	**6.4**	**4.53**	**.03**	**.23**	**.19**
Girls	Symbolic	5	.73	.61	4.14	5.37	-.18	-.13	-.08

After seeing the results in table 5, we thought that probably these girls would have lower cognitive development than boys. Actually, there were no boys with low cognitive development at all. Because of the group size differences and the differences in cognitive abilities (no low cognitive development boys), we cannot say anything about low cognitive development boys compared to low cognitive development girls. However, it is worth it to see the differences among high cognitive girls and high cognitive boys:

Table 5. Mean performance scores and attitudes' changes for high cognitive development students

Gender	Help system	N	finalpro	Avgprof	Nmastop	Hmtilmas	Scincr	UtIncr	Likeinc
Boys-High	Concrete	3	.83	.78	4.67	4.46	**.20**	.07	.33
	Symbolic	4	**.95**	**.86**	**7.75**	**4.35**	.09	**.27**	**.55**
Girls-High	Concrete	3	**.95**	**.73**	**6.67**	**4.55**	**.00**	**.17**	**.00**
	Symbolic	2	.83	.71	5.5	4.64	-.27	-.27	-.27

As expected, high cognitive development boys do better in the version with symbolic hints. However, high cognitive development girls do better in the low symbolic version, something we didn't expect to find. Of course girls of low cognitive development benefit better from the hints that are more concrete AND more interactive (not shown in the last table). However it is not obvious at all that high cognitive development girls seem to have done better with the concrete hints. For these girls, it seems that the high structuredness and interactivity of the concrete hints was more important than the abstraction of the symbolic hints that was appropriate for their cognitive development. Even worse, the symbolic version seemed to be harmful for girls in the emotional dimension (negative values in table 5). We should keep in mind that these two datapoints are not enough to draw any conclusions. It will be interesting to find in the near future how low cognitive development boys behave, as there is a trade-off between the concreteness of a hint (which is appropriate for the low cognitive development aspect) and the lower interactivity of the symbolic hints (which seem to be better for boys according to our results). We will be running a large study with 200 students in the near future to analyze these gender differences further.

Even though both the concrete and symbolic hints were previously considered as both "highly interactive and structured" (because of how they break the problem down into steps and interact with the student to get to the answer), concrete hints demand a lot more physical and attention effort from the user (mouse actions like dragging, dropping and clicking) while the symbolic version requires entering numbers and pressing return at each step. High cognitive girls may value this intense interactivity more than the concreteness-symbolism dimension.

6. LESSONS LEARNED AND FUTURE WORK

One important lesson to learn is how students' are sensitive to changes in the quality of the help provided. The symbolism in hints has proved to be an important determinant of performance for different groups of students, qualities of help which, to my understanding, had not been considered before for help systems in Interactive Learning Environments. Also, the time girls and boys are willing to spend in hinting is different. Boys spend less time in hints, and they seem to be bothered by intense hinting provision mechanisms. They seem to benefit more from explanations that are fast to check and go through, instead of interactive multimedia hints where they would have to devote lots of time and attention to interact with the hint. Giving that intense hinting mechanisms seem to be specially not good for boys, it would be interesting to see if they would do better with a mechanism of help seeking instead of help provision. Girls devote their time to go through any kind of explanation that we pose, and do better with those that are highly structured and interactive.

A second lesson has to do with how help effectiveness becomes reflected in actual student behavior. It is very usual to find measures of performance that depend on mistake reduction. However, we have shown that students are emotionally sensitive to different kinds of help they receive, and that sometimes those are not consistent with measures of performance (see table 5, self-confidence increase for boys). Analysing the emotional impact of help on top of the reduction of mistakes, and trying to maximise positive attitudes is an interesting contribution. After all, success in education has so much to do with the emotional relationship that people build with respect to the domain being taught. It would be interesting to study if there is any specific kind of help that benefits students who arrive at the tutoring session with different attitudes towards the topic being taught.

Many more groups of people should react differently to hints. For example, students with different attention abilities, students with various learning styles, etc. Also, there are many other features that can vary in hints: their spatial content, their reading level, the materials used in the explanations, their structure, etc. There are so many factors affecting the effectiveness of a help system, that the complexity of addressing more than a small amount of student descriptors and hint descriptors is overwhelming. As hint descriptors increase, analysing how each of those factors affect the effectiveness of hints becomes a very hard problem to deal with. The number of hints needed to test hypotheses increases exponentially with respect to the number of hint descriptors, and so does the number of human subjects needed to test for the relevance of these hint features. By the same token, as categorizations of students increase, the number of subjects needed to test hypotheses increases exponentially too.

The research presented in this paper was done with a macro-level of adaptivity in mind: because the student differences we considered are pretty permanent, we can implement an adaptive mechanism that would always give one type of hint to a student who falls into some stereotype. We have found students of different cognitive developments and different genders responding differently to the interactivity and symbolism of the help provided in the context of our Animalwatch Intelligent Tutoring System. We will be running a large study with 200 children in an urban area school where we will test hypotheses about what works for whom in the near future. After that, we will build a macro-adaptive mechanism in Animalwatch to give each group what is best for them.

REFERENCES

Arroyo, I.; Beck, J.; Schultz, K.; Woolf, B. (1999) Piagetian Psychology in Intelligent Tutoring Systems. In the *Proceedings of the Ninth 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 Fifth International Conference on Intelligent Tutoring Systems.*

Beck, J.; Stern, M.; Woolf, B.P. (1997) Using the Student Model to Control Problem Difficulty.In the Proceedings of Sixth International Conference on User Modeling. pp. 277-289.

Beck, J., Arroyo, I., Woolf, B. P., & Beal, C. R. (1999). Affecting self-confidence with an ITS. *Proceedings of the Ninth International Conference on Artificial Intelligence in Education,* pp. 611-613, Paris.

Beal, C.; Woolf, B.; Beck, J.; Arroyo, I.; Schultz, K.; Hart, D. (2000) Gaining Confidence in Mathematics: Instructional Technology for Girls. *Proceedings of International Conference on Mathematics/Science Education and Technology*. pp. 57-64.

Eccles, J.S., 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.

Jonassen, D.; Grabowski, B. (1993) Handbook of Individual Differences, learning and Instruction. Lawrence Erlbawm.

Luckin, R; du Boulay, B. (1999) Ecolab: the development and evaluation of a Vygotskian design framework. *International Journal of Artificial Intelligence in Education.* Volume 10. Number 2. pp. 198-220

Piaget, J. (1953, November) How Children Form Mathematical Concepts. *In Scientific American*.