The Impact of Decision Agency & Granularity on Aptitude Treatment Interaction in Tutoring

Guojing Zhou and Min Chi {gzhou3,mchi}@ncsu.edu
Department of Computer Science, North Carolina State University

Abstract

In this study, we explored the impact of the decision agency (Student vs. Tutor) and granularity (Problem vs. Step) across students with different levels of incoming competence (High vs. Low). Students were randomly assigned to four conditions and split into High and Low groups based on their pre-test scores. All students used the same Intelligent Tutoring System (ITS) called Pyrenees, followed the same general procedure, studied the same training materials, and worked through the same training problems. The only substantive differences among the four conditions were decision agency and granularity. That is: who decided to present an example or to solve a problem: the student or the ITS tutor; and was the decision made problem-by-problem or step-by-step? Our overall results showed that there were significant different impacts of the decision agency and granularity between High and Low students on learning performance. More specifically, for High students granularity was the more dominant factor in that step level decisions can be more effective than problem level decisions regardless of the decision agency; for Low students there was a significant interaction effect in that: Low students benefit significantly more when they were making problem-level decisions than making step-level decisions, but no significant difference was not found when the decisions were made by the tutor. Much to our surprise, both High and Low groups showed strong decision-making preference for problem solving over worked example at both problem and step levels.

Keywords: Aptitude Treatment Interaction, Pedagogical decisions, granularity, student-centered learning,

Introduction

Certain learners are less sensitive to learning environments and can always learn; while others are more sensitive to variations in learning environments and may fail to learn. In order to fully honor their promises, effective learning environments should exhibit an aptitude-treatment interaction (ATI), that is, its instruction should match to the aptitude of the learner (Cronbach & Snow, 1977). Intelligent Tutoring Systems (ITSs) are powerful educational technologies that support learning by providing step-by-step support and contextualized feedback adapted to individual learners and ITSs have demonstrated great success in many complex domains (Koedinger & et al., 1997; Vanlehn, 2006). In our work, we explored the possibility of improving the effectiveness of ITSs from two perspectives: decision agency and granularity. Here we split students into High and Low groups based on their incoming competence and investigated the impact of these two perspectives on ATI: how the decision agency and granularity would impact students' learning across the High and Low groups.

Decision Agency: ITSs are generally designed to support users' learning by providing instructions, scaffolded problem-solving practice and on demand help. Most of existing ITSs are *tutor-centered*. The tutor is responsible for selecting the next action to take at any given time. Each of these decisions affects student's successive actions and performance. In learning literature, the skills used to make such decisions are generally referred to as *pedagogical skills*. More formally, Chi et al. defined pedagogical skills are those "involve skillful execution of tactics, such as giving explanations and feedback or selecting the appropriate problems or questions to ask the students" (Chi, Siler, & Jeong, 2004). Most ITSs generally employ fixed pedagogical policies that do not adapt to users' needs. For example on most ITSs students are asked to *solve* a series of training problems while research showed that studying worked examples can be more effective than solving problems and the former generally takes much less time (McLaren & Isotani, 2011).

On the other hand, previous research showed that it is desirable for students to experience a sense of control over their own learning. For example, Cordova and Lepper (Cordova & Lepper, 1996) found that offering student choices over their learning could lead to significantly better learning outcome than those who were not offered. Letting students make decisions during the tutorial process should make them feel that they are actively directing their own learning process and not just passively following it. Furthermore, prior research suggested that offering student learning choices often exhibits an ATI effect: students with different levels of competence should be offered with different choices. For example, Young split learners into High vs. Low based on survey results and found that the performance difference between the High and Low learners was significantly greater under learner's control than under system control (Young, 1996). In this paper, we provided the students with different yet both reasonable choices and let them decide how they would like to study the materials and our goal is to investigate how these choices would impact their learning differently across High vs. Low students.

Granularity: Tutoring in domains such as math and science can be viewed as a two-loop procedure (Vanlehn, 2006). In the outer loop, the tutor makes tasks or problem-level decisions such as deciding what problem to solve next, while the inner loop controls step level decisions such as whether or not to give a hint. In educational literature, 'steps' often refer to the application of a major domain principle such as Newton's Third Law of Thermodynamics. Solving a complete problem generally involves applying many individual principles in a logical order. In theory, problem-level decisions are at a larger grain size and thus once students make one 'big' decision, they can focus on comprehending an example or solving a problem. However, such "big" decision

might not be very sensitive to students' specific moment-by-moment needs. For example, offering a complete worked example to students facing difficulty with a single principle may rob them of the chance to exercise other skills. When making step-level decisions, by contrast, students may be better able to tailor their decisions to their immediate needs and current knowledge level. However, making fine-grained decisions often requires more sophisticated decision-making skills. Prior research has shown interleaving worked examples with problem solving in both problem level and step level could result in improved learning performance comparing to doing problem solving only (Van Gog et al., 2011; Salden et al., 2010). However, it remained unclear how worked example and problem solving tasks should be provided to maximize the tutoring effectiveness. Therefore, in this paper, we are going to examine the impact of different decision granularity across learners with different incoming competence.

In this study, we strictly controlled the content to be equivalent for all participants by: 1) using an ITS which provides equal support for all learners; and 2) investigating on tutorial decisions that cover the same domain content at both problem and step levels, in this case Worked Examples (WE) versus Problem-Solving (PS). In WE, students were given a detailed example showing the expert solution for the problem or step. In PS, the students were tasked with solving the same problem or step using the ITS.

Previously we investigated the impact of granularity on the effectiveness of students' pedagogical decisions by comparing students' decisions against tutor's random yet reasonable decisions. Overall, our results showed that there was a significant interaction effect between decision agency (Tutor vs. Student) and granularity (Problem vs. Step) on learning. We found that step level decisions can be more effective than problem level decisions but the students were more likely to make effective pedagogical decisions at problem level than at step level (Zhou et al., 2016). In this paper, we further investigate the impact of decision agency and granularity across students with different levels of incoming competence. Following prior research, we divided students into High and Low groups based on their pretest scores and our primary research question is: would the impact of decision agency and granularity on learning differ between the High and Low students?

Background

WE/PS, vs. FWE: A number of researchers have examined the impacts of problem-level PS, problem-level WE, vs. *Faded Worked Example* (FWEs) (Renkl et al., 2002; Schwonke et al., 2009; Najar et al., 2014; Salden et al., 2010). FWEs interleave problem-solving steps with worked example steps within a single problem. Renkl et al. compared WE-PS pairs with FWE using a fixed fading policy (Renkl et al., 2002). In that study the number of example steps and problem-solving actions were strictly equal between the conditions. They found that FWEs with fixed fading policy significantly outperformed the WE-PS pairs, but no significant time-on-task differences were found. Schwonke et al. compared FWE with a fixed fading policy to tutored PS (Schwonke et al., 2009). Over the course of two studies, they found no significant differences between the two conditions in terms of their learning outcomes. However the FWE group spent significantly less time on task than the tutored PS group. Najar and colleagues compared FWE with an adaptive fading policy to WE-PS pairs. They found that the FWE condition significantly outperformed the WE-PS condition in their learning outcomes and spent significantly less time on task (Najar et al., 2014). Finally, Salden et al. compared three conditions: FWE with a fixed fading policy, FWE with an adaptive fading policy, and PS-only (Salden et al., 2010). They found that the adaptive FWE group outperformed the fixed FWE who, in turn, outperformed PS-only and there is no significant time-on-task differences among three groups.

Thus prior researchers have shown that FWEs with *effective* pedagogical polices can outperform fixed WE-PS pairs. It has also been shown that the former may need significantly less time on task than the latter. However all of these studies relied on hand-coded tutor pedagogical polices whereas in this study, we investigated how students with different levels of incoming competence would differ on pedagogical decision-making at both problem and step level.

Students Pedagogical Decision on ITS: Prior research on student problem-level decision-making has primarily focused on decisions of choosing instructional content, e.g. problem selection, but not how, e.g. WE vs. PS. Mitrovic et al. showed that learners, even college students, often make poor problem selections (Mitrovic & Martin, 2003). Long et al. compared the impact of joint student/system control again full system control over problem selection (Long & Aleven, 2014). In joint control, the system adaptively selects the problem type while the students select the individual problems. They found no significant difference on learning between the joint control groups and the full control group. In another study, Long et al. augmented a ITS with features that help students develop effective problem selection strategies with shared student/system control and compared its effectiveness with full system control ITS (Long & Aleven, 2016). They found that students in the shared control group learned significantly better than those in the full system control group. The results for student step-level pedagogical decision-making are inconclusive. Aleven & Koedinger studied students' help-seeking behaviors in the Cognitive Tutor (Aleven & Koedinger, 2000). They found that students cannot use hints effectively in that they tended to wait too long before asking for hints. Roll et al. by contrast examined the relationship between students' help-seeking patterns and their learning (Roll et al., 2014). They found that asking for help on challenging steps was generally productive while help abusing behaviors were correlated with poor learning. Finally, Wood et al. found that learners with high prior knowledge can exhibit more effective help-seeking behaviors than those with low prior knowledge learners(Wood & Wood, 1999).

Therefore prior research on students' decision suggests that

students can benefit substantially from effective pedagogical decision-making. Yet they often lack the necessary metacognitive skills to do so. On the other hand, help in ITSs is generally provided on demand, and some students might never need to request. In this study, we controlled for this possible conflict by focusing on WE/PS decisions, and by examining both problem and step-level decision-making. By letting both High vs. Low students make pedagogical decisions, we can fully investigate the impact of decision agency and granularity on learning across students of different levels of incoming competence.

Our Approach

We will investigate the impact of students' pedagogical decisions on learning by comparing students' decisions to tutors' **random** decisions at either problem or step level in order to avoid the impact of possibly misguided pedagogical policies. This study is 2 {Student, Tutor} $\times$ 2 {Problem, Step} design with four conditions: 1) $Stud_{Prob}$: problem-level student decisions; 2) $Stud_{Step}$: step-level student decisions; 3) Tut_{Prob}: problem-level random tutor decisions and 4) Tut_{Step}: step-level random tutor decisions.

Methods

Participants: This study was conducted in the undergraduate Discrete Mathematics course at the Department of Computer Science at NC State University in the Fall of 2015. 279 students participated in this study, which was given as their *final* homework assignment.

Conditions: The students were assigned to the four conditions via balanced random assignment based upon their course section and performance on the class mid-term exam. Since the two tutor-random decision groups were already compared in our prior study (Zhou et al., 2015) and the primary goal of this work is to examine the nature and effectiveness of students' pedagogical decision-making and ATI effect, we assigned twice more students to the two student-decision groups, $Stud_{Prob}$ & $Stud_{Step}$, than the two tutor-random groups, Tut_{Prob} & Tut_{Step}. The final group sizes are as follows: $N = 92$ for $Stud_{Prob}$, $N = 93$ for $Stud_{Step}$, $N = 47$ for Tut_{Prob}, and $N = 47$ for Tut_{Step}.

Due to the holiday break, preparations for final exams, and length of the experiment, 212 students completed the experiment. 11 students were excluded from our subsequent analysis because they performed perfectly on the pretest. The remaining 201 students were distributed as follows: $N = 70$ for $Stud_{Prob}$; $N = 59$ for $Stud_{Step}$; $N = 38$ for Tut_{Prob}; $N = 34$ for Tut_{Step}. A χ^2 test examining the relation between condition and completion rate showd no significant difference: $\chi^2(3) = 1.159, p = 0.763$.

Probability Tutor –Pyrenees Pyrenees is a web-based ITS for probability. It covers 10 major principles of probability, such as the Complement Theorem and Bayes' Rule. Pyrenees provides step-by-step instruction, immediate feedback and on-demand hints prompting students with what they should do next. As with other systems, help in Pyrenees is provided via a sequence of increasingly specific hints. The last hint in the sequence, the bottom-out hint, tells the student exactly what to do. For the purposes of this study we incorporated four distinct pedagogical decision modes into Pyrenees to match the four conditions.

Procedure In this experiment, students were required to complete 4 phases: 1) pre-training, 2) pre-test, 3) training on Pyrenees, and 4) post-test. During the pre-training phase, all students studied the domain principles through a probability textbook, reviewed some examples, and solved certain training problems. The students then took a pre-test which contained 10 problems. The textbook was not available at this phase and students were not given feedback on their answers, nor were they allowed to go back to earlier questions. This was also true of the post-test.

During phase 3, students in all four conditions received the same 12 problems in the same order on Pyrenees. Each primary domain principle was applied at least twice. The minimum number of steps needed to solve each training problem ranged from 20 to 50. The steps included variable definitions, principle applications and equation solving. The number of domain principles required to solve each problem ranged from 3 to 11. For the FWE problems, the $Stud_{Step}$ students were asked to make decision only on two types of steps: **principle selection** and **principle application**. To apply a principle, students need to first choose the principle they will use *(principle selection)* and then write the appropriate equation to apply it *(principle application)*. We evaluated the students' decisions on both types of steps in our analysis below. The only procedural differences among the four conditions were the decision agency: Student vs. Tutor and the granularity of the decision: Problem vs. Step. Apart from this, the system was identical.

Finally, all of the students took a post-test with 16 problems. Ten of the problems were isomorphic to the pre-test problems given in phase 2. Note that the rest of six questions are non-isomorphic complicated problems.

Grading Criteria: The test problems required students to derive an answer by writing and solving one or more equations. We used three scoring rubrics: binary, partial credit, and one-point-per-principle. Under the binary rubric, a solution was worth 1 point if it was completely correct or 0 if not. Under the partial credit rubric, each problem score was defined by the proportion of correct principle applications evident in the solution. A student who correctly applied 4 of 5 possible principles would get a score of 0.8. The One-point-per-principle rubric in turn gave a point for each correct principle application. All of the tests were graded in a double-blind manner by a single experienced grader. The results presented below were based upon the partial-credit rubric but the same results hold for the other two. For comparison purposes, all test scores were normalized to the range of [0,1].

Table 1: Learning Performance

High Group Students			
Cond	Pre	Iso Post	Overall Post
$Stud_{Prob}(31)$	.851(.059)	.909(.111)	.843(.143)
$Stud_{Step}(28)$	.846(.074)	.936(.062)	.882(.104)
$Tut_{Prob}(20)$	.857(.074)	.889(.088)	.785(.141)
$Tut_{Step}(20)$	.868(.058)	.931(.061)	.877(.113)
Low Group Students			
Cond	Pre	Iso Post	Overall Post
$Stud_{Prob}(39)$	.551(.144)	.863(.107)	.731(.126)
$Stud_{Step}(31)$	.512(.164)	.772(.182)	.658(.195)
$Tut_{Prob}(18)$	.603(.188)	.764(.272)	.693(.282)
$Tut_{Step}(14)$	.591(.132)	.856(.158)	.773(.167)

Results

We split students into High and Low groups based on their pre-test scores. Using a median split of 0.75 and students were divided into: High (n = 99) and Low (n = 102) groups. As expected, the High group scored significant higher than the Low group: $t(199) = 17.462$, $p < 0.0001$, $d = 2.464$. The numbers in the parentheses in the first column of Table 1 shows the numbers of High vs. Low students across the four conditions. No significant difference was found among the four conditions on the distribution of High vs. Low students: $\chi^2(3) = 1.1879, p = 0.7559$.

Fortunately, random assignment balanced the four conditions and this balance persisted even with the groups were subdivided into High and Low. The second column in Table 1 shows the pretest scores of High and Low groups. A one-way ANOVA test on students' pre-test score shows that there is no significant difference among the four conditions: $F(3,197) = 1.969$, $p = 0.12$, or the four conditions in High group: $F(3,95) = 0.581$, $p = 0.629$ or the four conditions in the Low group: $F(3,95) = 0.449$, $p = 0.719$.

Learning Performance

Table 1 shows a comparison of the pre-test, isomorphic post-test (10 isomorphic questions), and overall post-test scores among the four conditions, showing the mean (and SD) for each score.

To investigate the impact of decision agency and granularity on learning performance across High and Low students, two three-way ANOVAs were conducted using decision agency (tutor vs. student), granularity (problem vs. step), and incoming competence (High vs. Low) on the isomorphic post-test scores and the overall post-test scores respectively. For the isomorphic post-test scores, there is a significant three-way interaction effect: $F(1,193) = 4.079$, $p = 0.045$, a significant two-way interaction effect on decision agency and granularity: $F(1,193) = 5.324$, $p = 0.022$, a significant main effect on incoming competence: $F(1,193) = 26.23$, $p < 0.0001$ and a marginal interaction effect on granularity and incoming competence: $F(1,193) = 2.854$, $p = 0.093$. For the overall post-test scores, there is a significant two-way interaction effect on decision agency and granularity: $F(1,193) = 4.415$, $p = 0.037$, a significant main effect on incoming competence: $F(1,193) = 38.96$, $p < 0.0001$ and a marginal significant interaction effect on granularity and incoming competence: $F(1,193) = 3.521$, $p = 0.062$. Overall, our results showed that the impact of decision agency and granularity on learning performance differs significantly between the High and the Low groups. Next we will examine the learning performance of High and Low groups separately.

High Groups A repeated measure analysis using test type (pre-test vs. isomorphic post-test) as factors and test score as the dependent measure showed a main effect for test type $F(1,95) = 34.74$, $p < 0.0001$. Thus, overall the High students learned significantly by training on Pyrenees. However, further comparisons on the condition by condition basis revealed that: no significant improvement was found from pre-test to isomorphic post-test for the High Tut_{Prob} group:$F(1,19) = 1.817$, $p = 0.194$, but the remaining three High groups showed significant improvement: $F(1,30) = 6.385$, $p = 0.017$ for $Stud_{Prob}$; $F(1,27) = 22.58$, $p < 0.0001$ for $Stud_{Step}$ and $F(1,19) = 16.37$, $p = 0.0007$ for Tut_{Step}. This suggests that random problem level pedagogical decisions may not be very effective for High students.

To fully compare the learning performance among the four High groups, a two-way ANOVA analysis using decision agency and granularity as factors was conducted on the overall post-test scores. Our results showed while there is no significant interaction effect, there is a significant main effect on granularity: $F(1,95) = 5.504$, $p = 0.021$, that is, the step level decisions are significantly more effective than problem level decisions across the decision agencies. More specifically, the two step level decision groups, $Stud_{Step}$ and Tut_{Step} scored significantly higher than the Tut_{Prob} group: $t(38) = -2.263$, $p = 0.029$, $d = 0.716$ for the Tut_{Step} group and $t(46) = -2.749$, $p = 0.009$, $d = 0.805$ for the $Stud_{Step}$ group respectively. For isomorphic post-test scores. Two-way ANOVA analysis showed a marginal main effect on granularity: $F(1,95) = 3.563$, $p = 0.062$. Pairwise t-tests showed the $Stud_{Step}$ group outperformed the Tut_{Prob} group significantly: $t(46) = -2.178$, $p = 0.035$, $d = 0.638$ and the Tut_{Step} group tended to outperform the Tut_{Prob} group: $t(38) = -1.757$, $p = 0.087$, $d = 0.556$. Therefore, our results showed that step-level decisions are more effective for High group students than problem-level ones.

Low Groups A repeated measure analysis using test type as the repeated factor shows that Low group students learned significantly after training on Pyrenees. $F(1,98) = 200.01$, $p < 0.0001$. In fact, all four groups made significant improvement from pre-test to isomorphic test: $F(1,38) = 117.99$, $p < 0.0001$ for $Stud_{Prob}$; $F(1,30) = 63.89$, $p < 0.0001$ for $Stud_{Step}$; $F(1,17) = 8.537$, $p = 0.010$ for Tut_{Prob}; and $F(1,13) = 39.98$, $p < 0.0001$ for Tut_{Step}. This suggests that for Low students, the basic practice and problems, domain exposure, and interactivity of Pyrenees might help students

Table 2: Student Decisions

Problem Level Decisions			
Competence	WE	PS	Total
High	1.55(1.31)	8.45(1.31)	10
Low	1.56(1.48)	8.44(1.48)	10
Step Level Decisions			
Competence	WE	PS	Total
High	21.14(24.34)	115.07(27.03)	136
Low	18.84(17.28)	114.26(16.84)	133

to learn even when the problem- and step-level decisions are made randomly.

A two-way ANOVA analysis using decision agency and granularity on isomorphic post-test showed a significant interaction effect across the four Low groups: $F(1,98) = 5.819$, $p = 0.018$. Post hoc Pairwise t-test reveals that the $Stud_{Prob}$ Low group scored significantly higher than the $Stud_{Step}$ Low group: $t(68) = 2.591, p = 0.012, d = 0.624$. For overall post-test, our two-way ANOVA showed a marginal interaction effect: $F(1,98) = 3.591$, $p = 0.061$. Pairwise t-tests showed a trend that the $Stud_{Prob}$ group outperformed the $Stud_{Step}$ group: $t(68) = 1.903$, $p = 0.061$, $d = 0.458$. Therefore, our results showed that Low group students benefited more from making problem level decisions than step level ones and no significantly difference was found between the two tutor decision groups: Tut_{Prob} and Tut_{Step}.

To summarize, our results showed that: 1) for the High group, step level decision was more effective 2) for the Low group, letting students make problem level decisions can be more beneficial than letting them make step level decisions.

Student Pedagogical Decisions and Training Time

Student Decisions Much to our surprise, our analysis on students' decision-making preference revealed that both High and Low Groups are far more likely to choose problem solving than worked examples. For the tutor decision groups, our random policies generated a balanced 50-50 selection of WE and PS. Table 2 shows the number of pedagogical decisions made by students at both problem level and step level. Columns 2 and 3 show the average number of worked examples and problem-solving decisions made by each group. We required all students to solve two problems in order to familiarize them with Pyrenees. Therefore each student in problem-level condition made or received 10 problem-level decisions. Within each of the 10 problems, there are 6 to 24 step-level decisions. Therefore each student in step-level condition made or received about 135 step level decisions. In the following, we will compare the decision making preference across High and Low groups.

We compared the percentage of WEs students selected among different groups. For problem level decisions, both High and Low groups selected around 15%-16% of WEs on average. That is, both groups chose significantly less WEs than the two corresponding tutor groups: $t(49) = 8.717$, $p < 0.0001$, $d = 2.500$ for the High groups and $t(55) = -10.668$, $p < 0.0001$, $d = 3.040$ for the Low groups. The results for step level decisions are similar. High group students chose an average of 15.52% WE steps; while Low group students chose 14.16%. Again, both groups chose significantly less WEs than the two corresponding tutor decision groups: $t(46) = 8.920, p < 0.0001, d = 2.612$ for the High groups and $t(43) = 10.27$, $p < 0.0001$, $d = 3.308$ for the Low groups. The results suggested that students were significantly more likely to choose PS than WE at both levels.

Training Time Given that our results showed that the type of student decisions was not impacted by granularity and our preliminary results showed that similar patterns were found across the two different granularities on the training time. In the following, we will combine the step and problem level decision groups and mainly focus on the impact of the decision agency on time on task for High vs. Low students.

Despite the fact that students selected more PS, surprisingly, not all of them spent more time on learning comparing to those received equal number of PS and WE from the tutor. Table 3 shows the average total training time on Pyrenees (in seconds). A two-way ANOVA analysis examining the effect of incoming competence and decision agency shows a marginal significant interaction effect: $F(1,196) = 3.345$, $p = 0.069$. More specifically, while no significant difference was found between the two High groups, there is a significant difference between the two Low groups in that the student decision group spent significantly more time on training than the tutor decision group: $t(99) = -2.272, p = 0.025, d = 0.490$.

Since student decision groups chose more PS than WE and PS is generally more time consuming than WE, we further investigated the impact of decision agency on training time by comparing the *average time* on each WE step and PS step. The third and fourth columns in Table 3 shows the average amount of time students spent on each WE and PS steps respectively. For the average WE step time, no significant difference was found among the four groups. For the average time on PS steps, a two-way ANOVA on decision agency and incoming competence showed a significant main effect of decision agency: $F(1,196) = 14.53$, $p = 0.0002$. That is the student decision groups spent significantly less time on each PS step than the tutor decision groups. Pairwise t-test showed that this difference is significant for both High and Low groups: $t(97) = 6.118$, $p < 0.0001$, $d = 1.253$ for the

Table 3: Time Results

High Group Students			
Cond	Total	WE	PS
$High_{Stud}$	7977(1811)	9(10)	35(8)
$High_{Tut}$	8041(2503)	8(5)	51(18)
Low Group Students			
Low_{Stud}	8612(2428)	9(9)	39(10)
Low_{Tut}	7457(2179)	9(7)	50(16)

High group and $t(99) = 3.888$, $p = 0.0002$, $d = 0.839$ for the Low group. Therefore, students worked faster on PS steps when they made decisions than when tutor decided.

Discussion

In this study, we investigated the impact of decision agency (student vs. tutor) and granularity (problem vs. step) on learning across students with different levels of incoming competence (High vs. Low). Students were randomly assigned to four experiment conditions and split into High and Low groups based on their pre-test scores. Our results showed that all four Low groups and three out of four High groups (except the High Tut_{Prob} group) learned significantly after training on Pyrenees. In general, the Low students learned more than their High peers. This suggests that the training of Pyrenees is generally effective especially for low students.

We found that there were significantly different impacts of decision agency and granularity across High and Low students. For the High ones, granularity is the more dominant factor in that the two step-level groups significantly outperformed the two problem-level decision groups on the overall post-test. For the Low groups, there is a significant decision agency and granularity interaction effect: while no significant difference was found between the two Low tutor decision groups, the Low Student Problem-level group learned significantly more than the Low Student Step-level group. The results suggest that for High students step level decisions can be more effective than problem level decisions, but for Low students making problem level decisions are more beneficial than making step level ones.

Surprisingly, both High and Low students selected more problem solving than worked example at both problem and step level. However, students worked faster on PS steps when they selected them than received them. A potential explanation is that the control of their own learning process produced increases in motivation and depth of engagement. Currently, we are applying Reinforcement Learning (RL) to induce effective pedagogical policies based on which we will derive a methodology for teaching effective pedagogical decision-making strategy. After that, we will augment our ITS with decision-making development features to help students learn those strategies and examine its effectiveness.

Acknowledgements

This research was supported by the NSF Grant #1432156 and #1651909.

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In *Intelligent tutoring systems* (pp. 292–303).

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363-387.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology*, *88*(4), 715.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* Irvington.

Koedinger, K. R., & et al. (1997). Intelligent tutoring goes to school in the big city. *IJAIED*, *8*(1), 30-43.

Long, Y., & Aleven, V. (2014). Gamification of joint student/system control over problem selection in a linear equation tutor. In *Its* (pp. 378–387).

Long, Y., & Aleven, V. (2016). Mastery-oriented shared student/system control over problem selection in a linear equation tutor. In *Its* (pp. 90–100).

McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In *Aied* (pp. 222–229).

Mitrovic, A., & Martin, B. (2003). Scaffolding and fading problem selection in sql-tutor. In *Aied* (pp. 479–481).

Najar, A. S., Mitrovic, A., & McLaren, B. M. (2014). Adaptive support versus alternating worked examples and tutored problems: Which leads to better learning? In *Umap* (pp. 171–182).

Renkl, & et al. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education*, *70*(4), 293–315.

Roll, I., Baker, R. S. d., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, *23*(4), 537–560.

Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices learning. *Contemporary Educational Psychology*, *36*(3), 212–218.

Vanlehn, K. (2006). The behavior of tutoring systems. *IJAIED*, *16*(3), 227–265.

Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, *33*(2), 153–169.

Young, J. D. (1996). The effect of self-regulated learning strategies on performance in learner controlled computer-based instruction. *Educational Technology Research and Development*, *44*(2), 17–27.

Zhou, & et al. (2015). The impact of granularity on worked examples and problem solving. In *Cogsci* (pp. 2817–2822).

Zhou, & et al. (2016). The impact of granularity on the effectiveness of students' pedagogical decisions. In *Cogsci* (pp. 2801–2806).