The Impact of Granularity on the Effectiveness of Students' Pedagogical Decision

Guojing Zhou, Collin F. Lynch, Thomas W. Price, Tiffany Barnes, Min Chi
Department of Computer Science
North Carolina State University
{gzhou3,cflynch,twprice, tmbarnes,mchi}@ncsu.edu

Abstract

In this study we explored the impact of student versus tutor pedagogical decision-making on learning. More specifically, we examined what would happen if we let students decide how to handle the next task: to view it as a worked example or to solve it as a problem solving. We examined this impact at two levels of task granularity: problem vs. step. This 2×2 study was conducted on an existing Intelligent Tutoring System (ITS) called Pyrenees. 279 students were randomly assigned to four conditions and the domain content and required steps were strictly controlled to be equivalent across four conditions: all students used the same system, 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 {Student vs. Tutor} and granularity {Problem vs. Step}. That is: who decided to present an example or to solve a problem; and was the decision made problem-by-problem or step-by-step? Our results showed that there was a significant interaction effect between decision agency and granularity on student learning and a significant main effect of granularity on time on training. That is, 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 step level. In general, on both problem and step levels, the students were significantly more likely to decide to do problem solving rather than study it as a worked example.

Keywords: pedagogical policy, student-centered learning, problem solving, faded worked example, granularity

Introduction

Human one-on-one tutoring is one of the most effective way to improve student learning (Bloom, 1984). Intelligent Tutoring Systems (ITSs) are computer systems that mimic aspects of human tutors and have also shown to be successful as well (Koedinger, Anderson, Hadley, & Mark, 1997; Vanlehn, 2006). Most 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 the 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" (M. T. H. Chi, Siler, & Jeong, 2004). Most ITSs generally employ fixed pedagogical policies that do not adapt to users' needs, or they rely on hand-coded rules that seek to implement existing cognitive or instructional theories that may not have been well-evaluated. For example, in 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 (Sweller & Cooper, 1985; McLaren & Isotani, 2011).

On the other hand, much previous research showed that it is desirable for student to experience a sense of control over their own learning (Harackiewicz, Sansone, Blair, Epstein, & Manderlink, 1987). People are likely to persist at doing constructive things, like learning, exercising, quitting smoking, or fighting cavities, when they are given the choice and when they can make decisions. 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. Therefore, in this paper we provided the students with two different yet both reasonable choices and let them decide how they want to solve the problem next. So the question is: can students make effective pedagogical decisions that would promote their learning?

Moreover, we investigated the impact of students' decisions across two levels of granularity: problem versus step. 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 task 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, a student faces difficulty with a single principle then a complete worked example 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 many fine grain decisions can be more frustrating than beneficial over time.

In order to investigate the effectiveness of students' pedagogical decision-making at both levels of granularity, it is necessary to separate the pedagogical decisions from the instructional content, strictly controlling the content so that it is equivalent for all participants. To strictly control the content to be equivalent, 1) we used an ITS which provides equal support for all learners; and 2) we focused 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, student were given a detailed example showing the expert solution for the problem or were shown

the best step to take given their current solution state. In PS, by contrast,the students were tasked with solving the same problem using the ITS or completing an individual problem-solving step. While engaging students in decision-making within an ITS is not novel, prior researchers have generally focused on letting students dictate content by letting them decide what problem they wish to solve but not how they wished to solve it (Koedinger et al., 1997). So as far as we know, no prior research has investigated pedagogical decision-making independently of content selection.

In short, our primary research question is: will the granularity of the pedagogical decisions have an impact on the effectiveness of students' pedagogical decision-making? To investigate this question we will compare students' pedagogical decisions against tutor's decisions.

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, Atkinson, Maier, & Staley, 2002; Schwonke et al., 2009; Najar, Mitrovic, & McLaren, 2014; Salden, Aleven, Schwonke, & Renkl, 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 the fixed fading policy significantly outperformed the WE-PS pairs. They found no significant time-on-task differences between the two groups. 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. They found no significant time-on-task differences among three groups.

Thus prior researchers have shown that FWE with *effective* pedagogical polices can outperform fixed WE-PS pairs. It has also been shown that the former may require 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 whether students can make effective pedagogical decisions on whether to do PS or study a WE at either problem level or step level.

Students Pedagogical Decision on ITS

Prior research on problem-level decision-making has primarily focused on the impact of letting the students dictate content, e.g which problem to solve but not let students to decide how, e.g. WE vs. PS. The results for student step-level pedagogical decision-making are inclusive. Aleven & Koedinger studied students' help-seeking behaviors in the Cognitive Tutor (Aleven & Koedinger, 2000) where tutor permits students to request help when they do not know what step to take next. Help is provided via a sequence of hints that progress from general top-level hints that prompt the student to consider a principle or variable, to bottom-out hints that tell them exactly what action to take. They found that students do not always have the necessary metacognitive skills to know when they need help. They tend to wait too long before requesting information, and then focus only on applying the bottom-out action rather than processing the top-level conceptual guidance. Roll et al. by contrast examined the relationship between students' help-seeking patterns and their learning (Roll, Baker, Aleven, & Koedinger, 2014). They found that asking for help on challenging steps was generally productive while help abusing behaviors were correlated with poor learning.

Therefore prior research on students' help-seeking suggests that the students can benefit substantially from effective pedagogical decision-making. Yet they often lack the metacognitive skills that are required to do so. On the other hand, help in ITSs is generally provided on demand, and better-performing students are less likely to ask for it. Thus some students may simply never need to do so. 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. This allows us to evaluate *all* students' decision-making, not just the lower-performers and help-abusers. It also allows us to investigate the impact of granularity on student learning outcomes.

Our Approach

Previous studies on problem-level decision-making, PS vs. WE, mainly employed some fixed pedagogical polices (either WE-PS or PS-WE) and prior studies on step-level decision-making, FWE, either used fixed fading polices or relied on hand-coded adaptive polices. With adaptive policies, the system decides whether the next step is WE or PS based on a realtime assessment of the student's concept mastery. For example, students may be asked to solve steps involving the same concepts repeatedly until they demonstrate mastery and then such steps would be faded away by presented as WEs only. However there is no clear consensus on how or when students should be given a WE, nor how the faded policy should change on each level.

Therefore in this study 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.
4. Tut_{Step}: step-level random tutor decisions.

All students in this study were given the same problems in the same order. We compared the four groups using pre- and post-tests as well as their time on task.

Methods

Participants

This study was conducted in the undergraduate Discrete Mathematics course at the Department of Computer Science at North Carolina State University in the Fall of 2015. 279 students were enrolled in the course and this study was their *final* homework assignment. The students had two weeks to complete it and they were graded based upon their effort in completing the assignment, not their post-test scores.

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, Price, Lynch, Barnes, & Chi, 2015) and the primary goal of this work is to examine the nature and effectiveness of students' pedagogical decision-making, 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 probability 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}. We performed a χ^2 test of the relationship between students' condition and their rate of completion and found no significant difference among the groups: $\chi^2(3) = 1.159, p = 0.763$.

Probability Tutor

Pyrenees is a web-based ITS for probability. It covers 10 major principles of probability, such as the Complement Theorem and Bayes' Rule. In prior studies Pyrenees was compared against Andes, another well-evaluated ITS (Vanlehn et al., 2005). Results showed that Pyrenees significantly outperformed Andes in both physics (VanLehn et al., 2004) and probability (M. Chi & VanLehn, 2007). This improvement was observed in part because Pyrenees teaches students domain-general problem-solving strategies, which draw students' attention to the conditions under which each domain principle is applicable. The differences were apparent on all types of test problems: simple/complex problems and isomorphic/non-isomorphic problems, and the effects were large, with Cohen's d=1.17 for overall post-test scores.

Figure 1 shows the interface of Pyrenees, which is divided into multiple windows. Through the dialogue window, Pyrenees provides messages to the students such as explaining a worked example step, or prompting them to complete the next step. Students can enter their inputs, such as writing an equation or selecting the answer to a multiple-choice question, through the response text box below. Any variables or equations that are defined through this process are displayed on left side of the screen for reference. Any time an answer is submitted, Pyrenees provides immediate feedback on whether or not it is correct.

In addition to providing immediate feedback, Pyrenees can also provide on-demand hints prompting the student 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. They read a general description of each principle, reviewed some examples of its application, and solved some single- and multiple-principle practice problems. After solving each problem, the student's answer was marked in green if it was correct and red if incorrect. They were also shown an expert solution at the

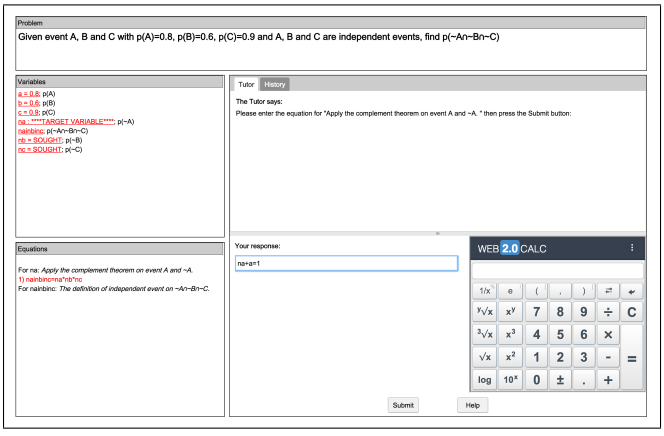

Figure 1: The interface of the Pyrenees probability tutor used in this study.

same time. If the students failed to solve a single-principle problem then they were asked to solve an isomorphic one; this process was repeated until they either failed three times or succeeded once. The students had only one chance to solve each multiple-principle problem and were not asked to solve an isomorphic problem if their answer was incorrect.

The students then took a pre-test which contained 10 problems. The textbook was not available. They 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 each principle, students need to first do principle selection: to choose the principle that they will use and then do principle application: to write the appropriate equation to apply it. 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. The remainder were non-isomorphic complicated multiple-principle 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].

Results

A one-way ANOVA test on students' pre-test score show that there is no significant difference among the four groups. $F(3,197) = 1.969$, $p = 0.12$. The second column in Table 1 showed students' pretest scores and as we can see, the two Tutor decision groups, Tut_{Prob} and Tut_{Step}, had higher pretest scores than the two Student decision groups: $Stud_{Prob}$ and $Stud_{Step}$ but the difference is not significant. Next we will compare students' learning performance in the post-test and training time across the four conditions. We discuss each comparison in turn.

Learning Performance

A repeated measures analysis using test type (pre-test and isomorphic post-test) as factors and test score as the dependent measure showed a main effect for test type $F(3,197) = 163.160$, $p < 0.0001$. On the isomorphic questions, all four groups of students scored significantly higher on the post-test than on the pre-test, $F(1,69) = 68.04$, $p < 0.0001$ for $Stud_{Prob}$; $F(1,58) = 65.35$, $p < 0.0001$ for $Stud_{Step}$; $F(1,37) = 8.349$, $p = 0.004$ for Tut_{Prob}; and $F(1,33) = 32.04$, $p < 0.0001$ for Tut_{Step}. Therefore all four conditions made significant gains from pre- to post-test by training on Pyrenees. This suggests that the basic practice and problems, domain exposure, and interactivity of Pyrenees might help students to learn even when the problem- and step-level decisions are made randomly.

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. We calculated a two-way ANCOVA analysis on decision agency (Student vs. Tutor) $\times$ granularity (Problem vs. Step) using pretest scores as a covariate. We found a significant interaction effect on the isomorphic post-test scores: $F(1,196) = 5.664$, $p = 0.018$. However, there was no significant main effect on either decision agency or the granularity alone. Pairwise t-tests showed a significant difference between $Stud_{Prob}$ and Tut_{Prob} groups: $t(106) = 2.514$, $p = 0.013$, $d = 0.477$, that is, the $Stud_{Prob}$ scored significantly higher than the Tut_{Prob} on isomorphic post-test scores. Additionally, there is a trend that Tut_{Step} group out-performed Tut_{Prob} group: $t(70) = -1.853$, $p = 0.068$, $d = 0.444$. Therefore, this result showed that students were able to make effective problem-level decisions in that $Stud_{Prob}$ group learned significantly more than random decision Tut_{Prob} group but not step level decisions in that $Stud_{Step}$ is not significantly better than those trained with the random decisions Tut_{Step}.

Similarly, a two-way ANCOVA on the factors of granularity and decision using pretest scores as a covariate also showed significant interaction effect on the overall post-test score: $F(1,196) = 4.375$, $p = 0.038$. Again there was no significant main effect on either the granularity or decision

Table 1: Learning Performance

Cond	pre	Iso Post	Overall Post
$Stud_{Prob}(70)$	.684(.186)	.890(.119)	0.788(.137)
$Stud_{Step}(59)$	.671(.212)	.861(.129)	0.778(.152)
$Tut_{Prob}(38)$	.737(.189)	.818(.177)	0.726(.198)
$Tut_{Step}(34)$	.754(.167)	.882(.101)	0.811(.133)

Table 2: Time on task (in minutes)

Cond	# Stud	Total Time
$Stud_{Prob}$	70	130.39(28.26)
$Stud_{Step}$	59	148.54(42.31)
Tut_{Prob}	38	121.53(47.15)
Tut_{Step}	34	136.44(30.27)

agency alone. Post-hoc pairwise t-tests showed the Tut_{Step} group had significantly higher scores than the Tut_{Prob} group: $t(70) = -2.107$, $p = 0.039$, $d = 0.503$ and a trend that the $Stud_{Prob}$ group out-performed the Tut_{Prob} group: $t(106) = 1.933$, $p = 0.056$, $d = 0.368$. Therefore, it seems that tutor's step-level decision is more effective than tutor's problem-level decision. But no significant difference was found between the two student decision making groups.

To summarize, our results showed that: 1) the granularity can make a significant difference on student learning in that tutor's step-level decisions can be more effective than tutor's problem-level decisions; and 2) students can make better problem-level decision than random, but not better step level decisions. Therefore, one potential explanation for the lack of the difference between the two student decision groups is that: while the step level decisions can indeed be more effective than problem-level decisions, the students cannot make effective step level decisions to fully take advantage of the learning power that step level decisions can provide. Further research is needed to investigate this hypothesis.

Training Time

Table 2 shows the average amount of total training time (in minutes) students spent on Pyrenees for each condition. A two-way ANOVA analysis on granularity and decision agency revealed there is no significant interaction effect. However, there is a significant main effect of granularity: $F(1,197) = 10.283$, $p = 0.0015$ and a marginal main effect of decision agency: $F(1,197) = 3.609$, $p = 0.059$. Subsequent pairwise t-tests showed that the $Stud_{Step}$ condition spent significantly more time than the $Stud_{Prob}$ and Tut_{Prob} conditions: $t(127) = -2.902$, $p = 0.004$, $d = 0.504$ ($Stud_{Prob}$); $t(95) = -2.937$, $p = 0.004$, $d = 0.603$ (Tut_{Prob}).

Overall, we found that decision granularity can make a difference on the time on task: 1) students spent more time with step-level decisions than problem-level decisions in that the two step-level groups spent significantly more time than the two problem-level groups; 2) the two student decision groups seemingly spent more time on task than the two random tutor groups, but the difference was only marginally-significant.

Student Decisions

Our preliminary analysis on students' decision-making preference suggested that students are far more likely to choose problem solving than worked examples.

Problem Level Decisions: Table 3 shows the number of different types of problem level decisions made by the $Stud_{Prob}$ and the Tut_{Prob} groups. Columns 2 and 3 show the average number of worked examples and problem-solving problems that each condition experienced. We required each student to solve two problems in order to familiarize them with Pyrenees. Therefore each student made 10 problem-level decisions. For the $Stud_{Prob}$ group, the students chose less than two WEs on average; while the Tut_{Prob} group, the students received an almost equal number of WEs and PSs (5.45 vs. 4.55) since the tutor makes random decisions. That is, the $Stud_{Prob}$ group only received 15.8% of worked examples; while the Tut_{Prob} group received 54.5% worked examples . This difference was statistically significant: $t(106) = -13.203$, $p < 0.0001$, $d = 2.614$

Table 3: Number of problem-level decisions

Cond	WE	PS	Total
$Stud_{Prob}$	1.58(1.40)	8.44(1.40)	10
Tut_{Prob}	5.45(1.57)	4.55(1.57)	10

Table 4: Number of step-level decisions

ST	Cond	WE	PS	Total
Principle Selection	$Stud_{Step}$	9(10)	58(11)	67
	Tut_{Step}	34(4)	33(4)	67
Principle Application	$Stud_{Step}$	11(11)	56(11)	67
	Tut_{Step}	34(5)	33(5)	67

Step Level Decisions: Table 4 shows the number of different types of step level decisions made by the $Stud_{Step}$ and Tut_{Step} groups on the principle selection and principle application steps. For both groups the number of WE and PS decisions sum to 67. The $Stud_{Step}$ group selected an average of only 9 WE steps and decided to do PS on the remaining 58 steps; while in the Tut_{Step} group, since the tutor makes random decisions, the students received an almost equal number of WE and PS steps (34 vs. 33). That is, the $Stud_{Step}$ group received 14.81% WE steps while the Tut_{Step} group received 50.92% WE steps. This difference was also statistically-significant: $t(91) = 13.67$, $p < 0.0001$. We found the similar patterns on the principle application steps as well.

To summarize, the two tutor decision groups received about equal number of WE vs. PS at either problem or step levels while the two student decision groups, either problem level or step level, are significantly more likely to decide to do Problem Solving than Worked Examples.

Discussion

In this study, we investigated the impact of students' pedagogical decision-making on learning. We focused on the decisions whether to give students a WE or to engage them in PS at two levels of granularity: problem versus step. We were able to strictly control the domain content and thus to isolate the impact of *pedagogy* from *content*. And we compared the

students' pedagogical decision making performance to a random baseline with a goal of factoring out the impact of hand-coded strategies on student learning.

Interestingly, our results showed that students can make effective problem-level decisions that enabled them to significantly outperform students with random decisions. However, the students were no better than a random tutor when making step-level decisions. When comparing across the four conditions we found that the Tutor random step-level decision group outperformed the Tutor random problem-level decision group ($Tut_{Step} > Tut_{Prob}$) but no significant difference was found between the two student decision groups.

Our results suggests that different granularity of pedagogical policies can significantly impact students' performance in that the step-level decisions can potentially be more beneficial than the step level ones; however, the students are more capable of making effective problem-level pedagogical decisions than making step-level ones. This may be due to the fact that students may lack the necessary metacognitive skills to make such fine-grain decisions or because they get overwhelmed by the number of decisions to make.

Surprisingly, students selected more problem solving than worked example on both problem and step levels. The feeling of engagement may partly explain their decisions. Prior work has shown that students are more likely to be engaged in the learning process when they experience a sense of control over it(Harackiewicz et al., 1987). Therefore, the students might decide to do problem solving simply because they feel more involved in problem solving than in worked example. However, much further research is needed to fully understand why.

Currently we are applying Reinforcement Learning (RL) to induce effective pedagogical policies directly from our data. We will investigate whether the RL-induced policies can be more effective than student decision making at both levels of granularity. We will also investigate whether it is possible to combine RL-induced policies with student decision making and thus give students both beneficial guidance and an all-important sense of agency.

Acknowledgements
This research was supported by the NSF Grant #1432156: "Educational Data Mining for Individualized Instruction in STEM Learning Environments".

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).

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

Chi, M., & VanLehn, K. (2007). The impact of explicit strategy instruction on problem-solving behaviors across intelligent tutoring systems. In *Proceedings of the 29th annual conference of the cognitive science society, nashville, tennessee* (pp. 167–172).

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

Harackiewicz, J. M., Sansone, C., Blair, L. W., Epstein, J. A., & Manderlink, G. (1987). Attributional processes in behavior change and maintenance: smoking cessation and continued abstinence. *Journal of Consulting and Clinical Psychology*, *55*(3), 372.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30-43.

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

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

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (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.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89.

Vanlehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227–265.

VanLehn, K., Bhembe, D., Chi, M., Lynch, C., Schulze, K., Shelby, R., ... Wintersgill, M. (2004). Implicit versus explicit learning of strategies in a non-procedural cognitive skill. In *Intelligent tutoring systems* (pp. 521–530).

Vanlehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... Wintersgill, M. (2005). The andes physics tutoring system: Lessons learned. *JAIED*, *15*(3), 147–204.

Zhou, G., Price, T. W., Lynch, C., Barnes, T., & Chi, M. (2015). The impact of granularity on worked examples and problem solving. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2817–2822).