Big, Little, or Both? Exploring the Impact of Granularity on Learning for Students with Different Incoming Competence

Guojing Zhou, Xi Yang, and Min Chi
⟨*gzhou*3, *yxi*2, *mchi*@*ncsu.edu*⟩
Department of Computer Science, College of Engineering
North Carolina State University, Raleigh, NC, USA

Abstract

We explored the impact of three types of decision granularity, problem level (Prob), step level (Step), and both problem and step levels (Both), on student learning. We first conducted an empirical study to directly compare the three conditions and then three subsequent studies to evaluate one or two of the three conditions. Overall our empirical results showed there was no significant difference among the three conditions. We further split students into different groups based on their performances on the single-principle and the multiple-principle problems in the pre-test. Solving the single-principle problems only involves one step while solving the multiple-principle ones involves generating multiple steps in a logic order. We define High students as those who were correct on *all* single-principle problems and *at least one* multiple-principle ones in the pre-test, Low students as those who were correct on *some or all* single-principle problems *but no* multiple-principle ones, and the rest are in the Medium group. Our empirical results showed that for Low students, Both can be better than Step. For the Medium and High students, no clear conclusions could be drawn because of small sample sizes. As a result, in a post-hoc analysis all students were combined by their assigned conditions. Overall, while no significant difference was found among the three conditions, we found that the impact of three types of granularity, Prob, Step, and Both differs significantly for High vs. Low students: $Both, Step > Prob$ for the High students and $Both, Prob > Step$ for the Low students. No clear conclusions could be drawn for the Medium group due to its small sample sizes. In short, while Prob could be effective for Low students but ineffective for High ones and Step could be effective for High students but ineffective for Low ones, Both seemed to be effective for both High and Low students.

Keywords: granularity, worked example, problem solving, student competence

Introduction

In STEM domains like math, probability and science, solving a *problem* often requires producing an argument, proof or derivation consisting of one or more inference steps, and each *step* is the result of applying a domain principle, operator or rule. For instance, an algebraic equation 2x+5=21 can be solved via two steps: 1) subtract the same term 5 from both sides of the equation; and 2) divide both sides by the non-zero term 2. As a result, tutoring in such domains is often structured as a two-loop procedure. An outer loop selects the *problem* or task the student should work on next, while the inner loop governs *step* level decisions such as whether or not to give a hint (Vanlehn, 2006).

In this paper, we directly explored the impact of three types of decision granularity on student learning by comparing three conditions: problem level (Prob), step level (Step), and both problem and step levels (Both). In the Prob condition, the tutor randomly decides whether the next problem is worked example (WE) or problem solving (PS). In WE, students observe how the tutor solves a problem, while in PS the students solve the problem themselves. In the Step condition, a random decision is made on whether the next *step* should be WE or PS. To differentiate it from the problem level PS and WE, we refer to such step level interleaving as *Faded Worked Example (FWE)*. Finally, the Both condition involves both levels of decisions: at the problem level, it randomly decides whether the next problem should be WE, PS or FWE; if FWE is selected, step level decisions will be randomly made.

A series of studies were conducted to evaluate the three types of decision granularity in the domain of probability using an Intelligent Tutoring System (ITS) named Pyrenees from 2014-2017. Pyrenees allowed us to rigorously control the content and vary only the types of decision granularity. In Fall 2014 (Fall'14), all three conditions were *empirically* compared; for the subsequent studies, only one or two conditions were examined.[1] In a post-hoc comparison, students from all studies were combined by their conditions because all conditions across different years went through the same standard 4-phase procedure: textbook, pre-test, training on ITS, and post-test, and all materials in each of the four phases were kept to be *identical* across different years. Overall, our results showed that there was no significant difference among the three conditions either in Fall'14 (Zhou, Price, Lynch, Barnes, & Chi, 2015) or in the post-hoc analysis.

On the other hand, the *aptitude-treatment interaction (ATI)* effect states that some instructional interventions can be more or less effective for particular students depending upon their specific abilities or knowledge (Cronbach & Snow, 1977; Snow, 1991). Here we argue that WE, PS, and FWE involve different *learning mechanisms*. More specifically, in WEs, students learn by *observing* how the tutor solves a problem; in PSs, students learn by *doing* – solving the problem with the tutor's assistance; in FWEs, students learn by *collaboratively constructing* the solution with the tutor. As a result, we argue that in the Prob condition students switched between learning by observing (WE) and learning by doing (PS); in the

[1] Please note that another purpose of the subsequent studies was to compare reinforcement learning induced policies with random policies. Due to participant limit, we were not able to compare the three conditions again.

Step condition, students learn by *collaboratively constructing* answers for (FWEs) with the tutor; in the Both condition, students experienced all three types of learning mechanisms. Therefore, we expect that Prob, Step, and Both can be more or less effective for different students.

To investigate whether there is indeed an ATI effect, we split students into High, Medium and Low groups based on their incoming competence measured by their performances on six single-principle and four multiple-principle problems in the pre-test. Solving the single-principle problems only involves one step while solving the multiple-principle ones involves generating multiple steps in a logic order. We define High students as those who were correct on *all six* single-principle problems and *at least one* multiple-principle ones in the pre-test, Low students as those who were correct on *some or all* single-principle problems *but no* multiple-principle ones, and the rest are in the Medium group. Our results from Fall'14 showed that for the Low students, both levels of the granularity (Both) is significantly more effective than the step level decisions (Step); for the Medium and High students, no clear conclusions could be drawn because of small sample sizes. In the post-hoc analysis, no clear conclusions could be drawn for the Medium group due to its small sample sizes and for the other two groups, we have: $Both, Step > Prob$ for the High students and $Both, Prob > Step$ for the Low students. In short, our post-hoc analysis suggested that the problem level decisions (Prob) could be effective for Low students but ineffective for High ones; on the other hand, the step level decisions (Step) could be effective for High students but ineffective for Low ones; finally, the both level decisions (Both) seemed to be effective for both High and Low students.

Background and Related Work

The Impact of Granularity Involving WE, PS, FWE

Much of prior research has investigated the effectiveness of WE, PS, FWE, and their various combinations (Sweller & Cooper, 1985; McLaren, Lim, & Koedinger, 2008; McLaren & Isotani, 2011; McLaren, van Gog, Ganoe, Yaron, & Karabinos, 2014; Van Gog, Kester, & Paas, 2011; Renkl, Atkinson, Maier, & Staley, 2002; Schwonke et al., 2009; Najar, Mitrovic, & McLaren, 2014; Salden, Aleven, Schwonke, & Renkl, 2010; Zhou et al., 2015; Zhou, Lynch, Price, Barnes, & Chi, 2016; Zhou & Chi, 2017; Zhou, Wang, Lynch, & Chi, 2017; Zhou, Azizsoltani, Ausin, Barnes, & Chi, 2019). Here we only include those that involved any of the three types of granularity. At the problem level granularity, for example, McLaren et al. (2008) found no significant difference in learning performance between Prob (WE-PS pairs) and PS-only, but the former spent significantly less time than the latter. In a subsequent study, McLaren and Isotani (2011) compared three conditions: WE-only, PS-only, and Prob (WE-PS pairs). Similarly, no significant differences were found among them in terms of learning gains, but the WE condition spent significantly less time than the other two; and no significant time on task difference was found between the PS and the Prob (WE-PS pairs) condition.

A series of studies compared the Step level and the Both level granularity with PS only (Schwonke et al., 2009; Salden et al., 2010). Results showed that the former two can be more effective than the latter. For example, Salden et al. compared three conditions: Both (WE-FWE-PS), Step (FWE), and PS-only (Salden et al., 2010). Their results showed that Step outperformed Both, which in turn outperformed PS-only, and no significant time on task difference was found among the three conditions. Note that in this study, the order of WE, FWE, and PS was fixed in Both; while in Step, the tutor used an adaptive pedagogical policy, expert rules combined with data-driven student models, to determine whether the next step should be WE or PS. Therefore, it is was not clear whether it was the adaption or the granularity that made the Step condition more effective than the other two conditions. In our studies, we factored out the impact of adaption by employing random policies.

While the studies described above mainly used PS-only as baselines, several studies directly compared different types of granularity. Overall, results suggested that the Both level granularity could be more effective than the Prob level (Renkl et al., 2002; Najar et al., 2014). For example, Renkl et al. (2002) compared Both (WE-FWE-PS) with Prob (WE-PS pairs) and the former significantly outperformed the latter on student learning performance while no significant difference was found between them on time on task. Similarly, Najar et al. (2014) compared Both (adaptive WE/FWE/PS) with Prob (WE-PS pairs). They found that the former significantly outperformed the latter in terms of learning outcomes and the former also spent significantly less time on task. Here, an adaptive pedagogical policy was also employed to make both the problem and step level decisions. Thus, it is quite possible that the superiority of Both over Prob stemmed from the adaption rather than from the granularity. In sum, while different decision granularities were involved in prior studies, the WEs and PSs were provided following some fixed or adaptive pedagogical policies. In this work, we factor out the impact of pedagogical policies by employing a random policy for all three types of granularity.

The ATI Effect of WE, PS, FWE

Some prior studies have also investigated the ATI effect of WE, PS, FWE, and their combinations (Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Najar & Mitrovic, 2013; Najar, Mitrovic, & McLaren, 2016). For example, Najar and Mitrovic (2013) compared three conditions: 1) WE-only, 2) PS-only and 3) Prob (WE-PS pairs) in the domain of Structured Query Language and students were split into High vs. Low groups based on their pre-test scores. The results showed that for the High students: Prob, PS-only $>$ WE-only; while for their Low peers: Prob $>$ PS-only, WE-only. In a subsequent study, Najar et al. (2016) compared Both (adaptive WE/FWE/PS) with Prob (WE-PS pairs) and students were divided into High and Low groups by a median split on pre-test scores. Results showed that for the High students, Both

Table 1: Single-principle Problem vs. Multiple-principle Problem

Type	Single-principle Problem	Multiple-principle Problem
Question	If $p(A \cap B) = 0.2$ and $p(B) = 0.5$, find $P(A\|B)$.	If $p(B) = 0.06$, $p(\sim A \cap \sim B) = 0.87$ and $p(A \cap B) = 0.03$, find $p(A)$.
Answer	Apply the definition of conditional probability: $p(A\|B) = p(A \cap B)/p(B) = 0.2/0.5 = 0.4$	1) Apply the complement theorem: $p(\sim B \cap \sim A) + p(\sim (\sim B \cap \sim A)) = 1$ 2) Apply the de morgan's law: $p(A \cup B) = p(\sim (\sim B \cap \sim A)) = 1 - 0.87 = 0.13$ 3) Apply the addition theorem: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, $p(A) = 0.13 + 0.03 - 0.06 = 0.1$.

is more effective than Prob; while for the Low students, no significant difference was found.

In short, prior research investigating the ATI effect of WE, PS, FWE, and their combinations showed that for Low students, Prob could be more effective than doing WE and PS only; but for High students, Both can be more effective than Prob. While much of prior ATI research involved one or two types of granularity, to the best of our knowledge, no prior study has investigated the ATI effect when comparing the three types of granularity directly.

High, Medium, vs. Low Students

To investigate the ATI effect, we need to first distinguish students based on some specific abilities or knowledge. Learning in STEM domains such as math and science often involves acquiring two types of knowledge: declarative and procedural (Anderson, 1993). Declarative knowledge includes facts that we know and that can be described to others, for example, "the probability of TRUE is always 1". Procedural knowledge specifies how to retrieve and use declarative knowledge to solve problems. It is a type of knowledge that display with behaviors and often times cannot be explicitly described. Procedural knowledge often requires the interplay of many cognitive factors including but not limited to the following five ones in order of occurrence: 1) acquisition of declarative knowledge, 2) identification and retrieval of the proper declarative knowledge, 3) application of declarative knowledge, 4) organization and production of solution plans; 5) execution of solution plans and evaluation of answers.

Similar to previous research, we used pre-test to measure students' incoming competence. Our pre-test contains single-principle problems which involve applying one domain principle once and multiple-principle problems which involve applying multiple domain principles and for some principles more than once. Table 1 shows an example for each of them. The second column shows the question and answer for a single-principle problem. As we can see, the problem can be solved by directly applying a single-principle. The third column shows a multiple-principle problem. Solving the problem needs to not only apply three algebraic principles but also organize them in a logical order.

Based on the five cognitive factors described above, we argue that solving single-principle problems mainly involves factors 1-3, while solving multiple-principle ones involves all five of them. Thus, students must be able to solve single-principle problems before they can solve multiple-principle problems. Our data supported this point, showing that students who could solve multiple-principle problems always had the perfect score on all single-principle problems in the pre-test. Therefore, in the following we refer to students who could solve at least one multiple-principle problem correctly as High students, those who could only solve some or all of the six single-principle problems correctly as Low students, and the rest as the Medium students.

Methods

Participants

Four studies were conducted in each of the Fall semesters from 2014-2017 to evaluate the three conditions: Prob, Step, and Both using an ITS named Pyrenees in the undergraduate-level Discrete Mathematics course at North Carolina State University. They were assigned to students as one of their regular homework assignments and the completion of the tutor was required for full credit. Students were told that the assignment will be graded based on their demonstrated effort rather than performance. In different studies, different conditions were evaluated and in each study, students were randomly assigned to each condition. In Fall'14, all three conditions were *empirically* compared while for the subsequent three studies, only one or two conditions were examined and in the post-hoc analysis, students from all studies across the four years were combined by their conditions.

Table 2 shows an overview of participants in the four studies and the post-hoc analysis: the first two columns show the semester of the study and its corresponding conditions; columns 3 and 4 list the number of students initially assigned and finally completed in each condition. Overall, Pearsons Chi-squared test showed that there was no significant difference among the three conditions on their completion of study: $\chi^2(2) = 1.13, p = 0.57$ for Fall'14 and $\chi^2(2) = 0.65, p = 0.72$ for the post-hoc analysis. Here we only focus on Fall'14 and the post-hoc analysis because all three conditions are present.

Finally, students with perfect pre-test scores were excluded because we could not measure the improvement they made through training. The last column in Table 2 shows the number of students included in the following analysis.

Probability Tutor

Pyrenees is a web-based tutor that teaches students a general problem solving strategy and 10 major probability principles, such as the Complement Theorem and Bayes' Rule. It provides students with step-by-step instruction, immediate feedback, and on-demand help. Specifically, the help is provided via a sequence of increasingly specific hints. The last hint in

Table 2: Participants for E ach Study and Condition

Study	Cond	Distributed	Completed	Included
Fall'14	Prob	58	38	37
	Step	59	39	37
	Both	59	34	34
Fall'15	Prob	47	38	38
	Step	47	35	34
Fall'16	Prob	40	32	31
	Step	41	35	35
Fall'17	Both	70	57	56
Post-hoc	Prob	145	108	106
	Step	147	109	106
	Both	129	91	90

the sequence, i.e., the bottom-out hint, tells student exactly what to do. The ITS has three basic modes. In the WE mode, all the steps in a problem were solved by the tutor while in the PS mode, they were solved by the student. In the FWE mode, each step has a 50% chance to be solved by the tutor and 50% chance by the student. Except for the decision granularity, the remaining components of the tutor, including the GUI interface, the training problems and the tutorial support were identical for all students.

Procedure

All four studies include the four identical phases: 1) textbook, 2) pre-test, 3) training, and 4) post-test. The only difference among the three conditions was the decision granularity level, problem level for Prob; step level for Step; and both the problem and the step level for Both.

During textbook, all students studied the domain principles through a probability textbook. They read a general description of each principle, reviewed some examples of it, and solved some single- and multiple-principle 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 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. They were not given feedback on their answers, nor were allowed to go back to earlier questions (this was also true for the post-test).

During training, students in all three conditions received the same 12 problems in the same order. Each main domain principle was applied at least twice. The minimal number of steps needed to solve each training problem ranged from 20 to 50. Such steps included variable definitions, principle applications, and equation solving. The number of domain principles required to solve each problem ranged from 3 to 11. The problems were given as WE, PS, or FWE, based upon the students' experimental condition. All students could access the textbook.

Finally, all students took the post-test which contained 16 problems in total. 10 of the problems were isomorphic to the pre-test problems given in phase 2. The remainder were non-isomorphic multiple-principle problems.

Grading criteria

The pre- and post-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

The three conditions were compared on test scores. For the Fall'14 study, a One-way ANOVA analysis on the pre-test score showed no significant difference among the three condition: $F(2,105) = 1.12$, $p = 0.33$, $\eta = 0.021$. A One-way ANCOVA analyses on the post-test score using the pre-test score as a covariate also showed no significant difference: $F(2,104) = 1.70$, $p = 0.19$, $\eta = 0.021$. Similar insignificant results were found in the host-hoc analysis: $F(2,299) = 0.68$, $p = 0.51$, $\eta = 0.005$ for the pre-test and $F(2,298) = 0.98$, $p = 0.38$, $\eta = 0.004$ for the post-test. In terms of time on task, contrast analysis revealed that Prob spent significantly less time than Step in both Fall'14: $t(105) = -2.62$, $p = 0.010$, $d = 0.61$ and post-hoc: $t(299) = -3.00$, $p = 0.003$, $d = 0.40$.

To evaluate the ATI effect, we split students based on their pre-test scores. Our pre-test included six single-principle and four multiple-principle problems. Following our splitting criteria discussed above, we refer to students who could solve at least one multiple-principle problem correctly (pre ≥ 0.7) as High students [2]; those who could only solve some or all of the six single-principle problems correctly (pre ≤ 0.6) as Low students, and the rest as Medium ones. As expected, in the pre-test the High group scored significantly higher than the Medium group: $t(105) = 6.94$, $p < 0.0001$, $d = 3.16$ in Fall'14 and $t(299) = 9.71$, $p < 0.0001$, $d = 2.37$ in post-hoc; the Medium group significantly outperformed the Low group: $t(105) = 8.41$, $p < 0.0001$, $d = 2.14$ in Fall'14 and $t(299) = 11.82$, $p < 0.0001$, $d = 2.08$ in post-hoc.

Incoming competence combined with three conditions partitioned the students into nine groups for both Fall'14 and

[2]Note that in our grading rubrics, all problems were weighted equally in both pre- and post-tests.

Table 3: Students Performance and Time (minutes) on Fall'14 Empirical Study and Post-hoc Analysis

Cond	Fall'14 Empirical Study					Post-hoc Analysis				
	N	Pre	Iso	Post	Time	N	Pre	Iso	Post	Time
Prob_H	12	.857(.065)	.817(.125)	.700(.169)	85.5(19.9)	56	.822(.086)	.844(.138)	.740(.178)	111.1(42.5)
Step_H	8	.800(.080)	.868(.156)	.769(.154)	125.2(40.0)	47	.827(.077)	.908(.089)	.819(.126)	128.2(29.4)
Both_H	13	.826(.064)	.863(.136)	.767(.160)	113.2(30.1)	46	.818(.073)	.902(.110)	.821(.156)	106.6(29.3)
Prob_M	5	.636(.017)	.728(.134)	.598(.168)	96.8(17.4)	10	.652(.021)	.813(.148)	.688(.173)	101.1(24.8)
Step_M	6	.647(.017)	.687(.187)	.559(.124)	124.1(22.2)	12	.649(.022)	.818(.190)	.715(.191)	125.7(28.5)
Both_M	6	.653(.028)	.740(.163)	.618(.189)	111.5(25.8)	11	.652(.025)	.736(.216)	.616(.216)	113.1(26.3)
Prob_L	20	.453(.117)	.657(.190)	.528(.205)	95.8(29.3)	40	.455(.117)	.703(.234)	.596(.244)	98.7(35.7)
Step_L	23	.441(.103)	.592(.192)	.458(.153)	104.2(38.1)	47	.439(.110)	.628(.219)	.500(.190)	109.8(34.6)
Both_L	15	.414(.110)	.703(.170)	.550(.154)	105.1(37.4)	33	.415(.119)	.707(.208)	.565(.185)	110.3(36.3)

post-hoc. The number of students in each group is listed in the "N" column for Fall'14 in Table 3 (Left) and for post-hoc in Table 3 (Right). Fortunately, random assignment balanced the three conditions for ability, and this balance persisted even after the groups were subdivided into High, Medium, and Low. No significant difference was found on pre-test among the three High groups, the three Medium groups, or the three Low groups in both Fall'14 and post-hoc.

Empirical Fall'14 Study

In Table 3, the first column shows the condition-competence group and then followed by a section presenting the learning performance and time on task (in minutes) for Fall'14. Here it shows the number of students (N) and the mean and SD of pre-test score (Pre), isomorphic post-test score (Iso), overall post-test score (Post) and time on task (Time). A Chi-square test showed that there was no significant relation between condition and incoming competence $\chi^2(4) = 2.94, p = 0.57$.

To measure student learning improvement, we compared their isomorphic post-test scores with their pre-test scores. A repeated measures analysis using test type (pre-test vs. isomorphic post-test) as a factor and test score as the dependent measure showed that there is a main effect for test type: $F(1,107) = 50.82$, $p < 0.0001$, $\eta = 0.322$ in that they scored significantly higher on the isomorphic post-test problems than pre-test. Thus, our tutor is indeed effective on improving student learning. More specifically, all three conditions scored significantly higher in the isomorphic post-test than in the pre-test: $F(1,36) = 13.56$, $p = 0.0008$, $\eta = 0.274$ for Prob, $F(1,36) = 16.26$, $p = 0.0003$, $\eta = 0.311$ for Step, and $F(1,33) = 20.92$, $p < 0.0001$, $\eta = 0.388$ for Both respectively. This showed that the basic practices and problems, domain exposure, and interactivity of our ITS might be effective to help students acquire knowledge.

Finally, to obtain a comprehensive evaluation of students' final performance, analyses were performed on the overall post-test which contains six additional multiple-principles. A two-way ANCOVA analysis on the factors of granularity and incoming competence using the pre-test score as a covariate showed no significant interaction or main effect. A subsequent pairwise contrast analysis revealed that for Low students, the Both_L group scored significantly higher than the Step_L group: $t(98) = -2.01$, $p = 0.047$. The results suggested that the Both levels of decisions can be more effective than the step level decisions for the Low students.

In terms of time on task, a two-way ANOVA analysis on granularity and incoming competence showed a main effect on granularity: $F(2,99) = 3.97$, $p = 0.02$, $\eta = 0.071$ in that the Prob condition spent significantly less time than the Step condition $t(105) = -2.62$, $p = 0.01$, $d = 0.61$ and the Both condition $t(105) = -2.22$, $p = 0.029$, $d = 0.58$. Subsequent contrast analyses showed that such difference mainly came from the High students in that: Prob_H spent significantly less time than Step_H and Both_H: $t(99) = -2.72$, $p = 0.008$, $d = 1.35$ and $t(99) = -2.17$, $p = 0.03$, $d = 1.08$ respectively; no significant difference was found among the three Low groups.

Overall, Fall'14 results showed that on learning performance, Both was better than Step for the Low students; while on time on task, Prob spent less time than the other two for the High students. Note that since some of the groups are in small size, the absence of significant differences might be due to insufficient statistical power.

Post-hoc Analysis

The right section of Table 3 presents the post-hoc analysis results. Numbers in the "N" column revealed that the three High and the three Low groups are in reasonable size while the three Medium groups remain small. A Chi-square test showed no significant relation between condition and incoming competence: $\chi^2(4) = 2.11, p = 0.72$.

A repeated measures analysis using test type (pre-test vs. isomorphic post-test) as a factor and test score as the dependent measure showed that there was a main effect for test type $F(1,301) = 177.38$, $p < 0.0001$, $\eta = 0.371$ in that students scored significant higher in the isomorphic post-test than in the pre-test. Similarly, for each of the three conditions, students scored significantly higher in the isomorphic post-test than in the pre-test: $F(1,105) = 42.79, p < 0.0001, \eta = 0.290$ for Prob; $F(1,105) = 72.27$, $p < 0.0001$, $\eta = 0.408$ for Step and $F(1,89) = 67.46$, $p < 0.0001$, $\eta = 0.431$ for Both. The

results confirmed that our tutor is effective over the years.

For the overall post-test scores, a two-way ANCOVA analysis on the factors of granularity and incoming competence using the pre-test score as a covariate showed a significant interaction effect: $F(4,292) = 3.66$, $p = 0.006$, $\eta = 0.029$. Subsequent contrast analyses showed that for High students, the Step_H group and the Both_H group scored significantly higher than the Prob_H group: $t(292) = 2.25$, $p = 0.03$ and $t(292) = -2.50$, $p = 0.01$ respectively. For Low students, the Prob_L group and the Both_L group scored significantly higher than the Step_L group: $t(292) = 2.29$, $p = 0.02$ and $t(292) = 2.19$, $p = 0.03$ respectively. The results suggest that for High students, the Step level decisions and the Both level decisions are more effective than the Prob level while for Low students, the Prob level decisions and the Both level decisions are more effective than the Step level.

For time on task, a two-way ANOVA analysis on granularity and incoming competence showed a significant main effect on granularity: $F(2,293) = 4.98$, $p = 0.007$, $\eta = 0.032$ in that the Step condition spent more time than the Prob condition: $t(299) = 3.00$, $p = 0.003$, $d = 0.40$ and the Both condition: $t(299) = 2.22$, $p = 0.027$, $d = 0.34$. Subsequent contrast analysis revealed that for High students: the Step_H group spent longer time than the Prob_H group: $t(293) = 2.51$, $p = 0.01$, $d = 0.46$ and the Both_H group: $t(293) = 3.03$, $p = 0.003$, $d = 0.74$. No such significant difference was found among the three Low groups.

Overall, the results suggest that on learning performance, the problem level decisions can be effective for Low students but ineffective for High students, the step level decisions could be effective for High students but ineffective for Low students, while Both level decisions seem to be effective for both High and Low students. For time on task, the High students, the Step_H group can spend more time than the Prob_H and the Both_H groups while no significant difference was found among the three Low groups.

Conclusion & Discussion

In this paper, we explored the impact of three types of decision granularity on student learning by comparing three conditions: Prob involving WE and PS, Step involving FWE only, and Both involving all WE, PS and FWE. Overall, while no significant difference was found among the three conditions on learning performance, a significant difference was found among them on time on task in that Prob spent significantly less time than Step for both Fall'14 and the post-hoc.

We hypothesized that different learning mechanisms are involved in WE, PS and FWE and thus there may exist an ATI effect. Students were then split into High, Medium and Low groups based on their pre-test performance. Results from Fall'14 show that on learning performance, for Low students Both is more effective than Step; on time on task, for High students Prob would spend less time than Step. Overall because of small sample sizes, more general conclusions cannot be drawn here. Furthermore, our post-hoc results suggest that on learning performance, Prob can be effective for Low students but ineffective for High ones on the other hand, Step could be effective for High students but ineffective for Low ones; finally, Both seemed to be effective for both High and Low students; as for time on task, while no significant difference was found among the three Low groups either in Fall'14 or post-hoc, significant difference was found among the three High groups in that Prob_H spent significantly less time than Step_H in both Fall'14 and post-hoc.

Our results showed a difference between the Prob and Step granularity. In terms of time on task, students spent less time when learning with Prob than with Step. For learning performance, each of them can be effective for some students but ineffective for some other students, depending on students' knowledge level. This suggests that the granularity can have an impact on student learning. Additionally, results for the Both granularity suggest that mixing this two types of granularity together has the potential to get a more robust instructional intervention. The Prob granularity can be ineffective for the High students and the Step granularity can be ineffective for Low students, but our results suggest that Both can be effective for both High and Low students.

One possible explanation for our results is that different cognitive load were involved in the three conditions. At the problem level, students pay attention to either the tutor's solution in WE or their own solution in PS; while at the step level, they need to pay attention to both the tutor's solution and their own solution and integrate them. Compared with PSs, in FWEs the tutor may solve certain steps for students but on the other hand, students need to devote extra effort to understand and to integrate their answers with the tutor's answers. Thus, we hypothesized that in terms of cognitive load, WE <PS <FWE. This explains why the Step condition spent more time than the Prob condition (in both Fall'14 and post-hoc) despite that students in these two condition completed the same amount of work (as measured by the number of PS steps in our subsequent log analysis). Assuming that FWEs are more challenging than WEs or PSs, the results that Step benefits the High students more than Prob while Prob benefits the Low ones more than Step can be explained by the conjecture that High students have more prior knowledge and learning capacity than the Low ones. However, this is only our hypothesis and much more research is needed to fully understand it. More importantly, more research is needed to explain why the Both levels of granularity benefits both High and Low students.

Lots of prior research has shown that studying WEs help students learn. However, questions about how and when WEs should be presented remain open. Our findings inform researchers that the granularity can have an impact on student learning and the impact of granularity can differ for students at distinct knowledge levels. Thus, it urges researchers to consider the impact of granularity when designing instructions and adapt the instruction based on students' knowledge level.

Acknowledgements

This research was supported by the NSF Grants #1432156: "Educational Data Mining for Individualized Instruction in STEM Learning Environments", #1651909: "CAREER: Improving Adaptive Decision Making in Interactive Learning Environments", #1726550: "Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments", and #1916417: "MetaDash: A Teacher Dashboard Informed by Real-Time Multichannel Self-Regulated Learning Data". We would also like to thank the anonymous reviewers for their valuable feedback.

References

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35.

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

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of educational psychology*, *93*(3), 579.

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

McLaren, B. M., Lim, S.-J., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? new results and a summary of the current state of research. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2176–2181).

McLaren, B. M., van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2014). Exploring the assistance dilemma: Comparing instructional support in examples and problems. In *Intelligent tutoring systems* (pp. 354–361).

Najar, A. S., & Mitrovic, A. (2013). Do novices and advanced students benefit differently from worked examples and its. In *Proceedings of international conference icce* (pp. 20–29).

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). Springer.

Najar, A. S., Mitrovic, A., & McLaren, B. M. (2016). Learning with intelligent tutors and worked examples: selecting learning activities adaptively leads to better learning outcomes than a fixed curriculum. *UMUAI*, *26*(5), 459–491.

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.

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.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of consulting and clinical psychology*, *59*(2), 205.

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.

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. *International journal of artificial intelligence in education*, *16*(3), 227–265.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2019). Hierarchical reinforcement learning for pedagogical policy induction. In *International conference on artificial intelligence in education.*

Zhou, G., & Chi, M. (2017). The impact of decision agency & granularity on aptitude treatment interaction in tutoring. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 3652–3657).

Zhou, G., Lynch, C., Price, T. W., Barnes, T., & Chi, M. (2016). The impact of granularity on the effectiveness of students' pedagogical decision. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2801–2806).

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).

Zhou, G., Wang, J., Lynch, C., & Chi, M. (2017). Towards closing the loop: Bridging machine-induced pedagogical policies to learning theories. In *Edm* (pp. 112–119).