The Impact of Granularity on Worked Examples and Problem Solving

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

Abstract

In this paper, we explore the impact of two types of instructional interventions, *worked examples* and *problem solving*, at two levels of granularity: problems and steps. This study drew on an existing Intelligent Tutoring System (ITS) for Probability called Pyrenees and involved 266 students who were randomly assigned to five conditions. All students experienced the same procedure, studied the same training problems in the same order, and used the same ITS. The conditions differed only in how the training problems were presented. Our results show that when the domain content and required steps are strictly equivalent, different granularities of pedagogical decisions can significantly impact students' time on task. More specifically, the fine-grained step level decisions can have a stronger pedagogical impact than the problem-level ones.

Keywords: worked example, problem solving, faded worked example, granularity

Introduction

A great deal of research has investigated the different impacts of worked examples (WE) and problem solving (PS) on student learning (Sweller & Cooper, 1985; McLaren, Lim, & Koedinger, 2008; McLaren & Isotani, 2011; McLaren, van Gog, Ganoe, Yaron, & Karabinos, 2014; Renkl, Atkinson, Maier, & Staley, 2002; Schwonke et al., 2009; Najar, Mitrovic, & McLaren, 2014; Salden, Aleven, Schwonke, & Renkl, 2010). In PS students are given tasks to complete either independently or with assistance while in WE, students are given detailed solutions. When comparing WE to PS, we often need to control for content. Sweller and Cooper, for example, compared the learning effects of WE-PS pairs with PS-only (Sweller & Cooper, 1985). In the WE-PS condition, students studied a worked example and then solved a practice problem. Their results showed that the WE-PS condition not only learned significantly more but spent significantly less time than the PS-only condition. However, it is possible that the primary benefit of the WE-PS training was that students received additional domain content that was not given to the PS-only ones. Therefore, in this paper we will focus on research that controlled for learning content across the conditions.

Several techniques have been employed to control for learning content. One approach is to use a tutor such as an Intelligent Tutoring System (ITS). ITSs are generally designed to give students on-demand hints, and to give immediate or delayed feedback on submitted solutions. In this paper we will focus on comparisons between in-tutor WE and tutor-assisted PS, and we will explicitly state when this is not the case.

Tutoring in domains such as math and science can be viewed as a two-loop procedure (Vanlehn, 2006). The outer loop makes problem or task level decisions, such as deciding which problem or example to provide next, while the inner loop governs step level decisions during problem solving. In the educational literature, the term "step" often refers to the application of a major domain principle or equation, such as Newton's Third Law of Thermodynamics, during problem solving. Solving a whole problem generally involves carrying out many individual steps in a logical order. Based on this two-loop structure, we further divide the prior research into two levels of granularity: problem level and step level. Research on the impact of step level decisions has generally been focused on the impact of *faded worked examples* (FWEs). FWEs interleave problem solving with step-level examples within a problem. In the remainder of this section we will describe prior work on WEs vs. PS at both levels of granularity and we will focus on two types of outcome measures: learning performance and time on task.

Problem Level Decisions

McLaren and colleagues compared problem-level WE-PS pairs with PS-only (McLaren et al., 2008). Every student was given a total of 10 training problems. Students in the PS-only condition were required to solve every problem while students in the WE-PS condition were given 5 example-problem pairs. Each pair consisted of an initial worked example problem followed by tutored problem solving. They found no significant difference in learning performance between the two conditions, however the WE-PS group spent significantly less time than the PS group.

McLaren and his colleagues found similar results in two subsequent studies (McLaren & Isotani, 2011; McLaren et al., 2014). In the former, the authors compared three conditions: WE, PS and WE-PS pairs, in the domain of high school chemistry. All students were given 10 identical problems. Students in the PS group were required to solve each problem in an ITS. Students in the WE group viewed them as examples, and students in the WE-PS group alternated worked examples with problem solving. As before, the authors found no significant differences among the three groups in terms of learning gains but the WE group spent significantly less time than the other two conditions; and no significant time on task difference was found between the PS and WE-PS conditions.

In a follow-up study, conducted in the domain of high school stoichiometry, McLaren and colleagues compared four conditions: WE, tutored PS, untutored PS, and Erroneous Examples (McLaren et al., 2014). Students in the Erroneous Examples condition were given *incorrect* worked examples containing between 1 and 4 errors and were tasked with cor-

recting them. Again the authors found no significant differences among the conditions in terms of learning gains, and as before the WE students spent significantly less time than the other groups. More specifically, for time on task they found that: *WE* < *Erroneous Examples* < *untutored PS* < *tutored PS*. In fact, the WE students took only 30% of the total time that the tutored PS students did. $M = 19.8$, $SD = 5.8$ and $M = 62.4$, $SD = 17.2$ respectively.

The advantages of worked examples were also demonstrated in another study in the domain of electrical circuits (Van Gog, Kester, & Paas, 2011). The authors of that study compared four conditions: WE, WE-PS pairs, PS-WE pairs (problem-solving followed by an example problem), and PS only. They found that the WE and WE-PS students significantly outperformed the other two groups, and found no significant differences was found among four conditions in terms of time on task.

In short, prior research has shown that problem-level worked examples can be as or more effective than problem solving or alternating problems with examples, and the former can take significantly less time than the latter two (Sweller & Cooper, 1985; McLaren et al., 2008; McLaren & Isotani, 2011; McLaren et al., 2014; Renkl et al., 2002; Schwonke et al., 2009).

Step Level Decisions

With respect to step level decisions, the results from previous research are mixed. For example, Renkl et al. compared WE-PS pairs with FWE using a fixed fading policy (Renkl et al., 2002). For FWEs with a fixed fading policy, the study designer predefined which steps to give as examples and which steps to task students with solving. The number of examples and tasks provided was equal in both conditions. They found that a FWE with the fixed fading policy significantly outperformed WE-PS pairs. No significant difference was found between the two groups on time on task.

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 difference in terms of learning outcomes between the two conditions, however the FWE group spent significantly less time than tutored PS group.

Najar and colleagues (Najar et al., 2014) compared FWE with an adaptive fading policy to WE-PS pairs. They found that the FWE condition significantly outperformed the WE-PS condition in terms of their learning outcomes and the former also spent significantly less time on task than the latter.

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). With respect to learning outcomes, they found that FWE with the adaptive fading policy outperformed FWE with the fixed fading policy, which in turn outperformed PS-only. They found no significant time on task differences among the groups.

In short, for step-level worked examples, while the results have been generally mixed, it has been shown that FWE with effective fading policies can outperform either PS or WE-PS pairs. It has also been shown that the former may require significantly less time than either of the latter two.

Our Approach

In this study, we compared five conditions:

1. **Worked Examples (WE)**: where the tutor guides the student through a complete solution.
2. **Problem Solving (PS)**: where the student is required to solve each problem with the assistance of an ITS.
3. **Faded Worked Examples (FWE)**: where problem solving steps are interspersed with step-level worked examples.
4. **WE/PS**: where students receive both WE and PS problems.
5. **ALL**: where students receive WE, FWE and PS problems.

Most of the prior research focused on comparing the effectiveness of two or three conditions. To our knowledge, no prior study has compared all five conditions directly, especially WE vs. FWE.

For the WE/PS, FWE and ALL conditions, there are many ways to make problem-level decisions, such as when to provide a WE, PS or FWE. For FWEs, there are also step-level decisions, such as whether to provide the next step as a worked example or as a problem solving task. *Pedagogical* strategies are *policies* used to decide the next system action when there are multiple actions available.

Generally speaking, prior research studying problem-level decisions employed fixed pedagogical policies: either WE-PS (a worked example first followed by problem solving) or PS-WE. Studies of step-level decisions generally used a fixed fading policy or an adaptive fading policy. In the former case the order of steps was pre-specified and did not adapt to the students' learning experience. For adaptive fading policies, such decisions are made based upon a real time evaluation of the student's mastery of the subject knowledge. For example, a student may be asked to solve a step until he/she has demonstrated mastery of the knowledge involved in it. Note that in prior studies both fixed fading policies and adaptive fading policies have been defined by hand-coded rules.

We have previously investigated the application of data-driven methodologies to induce pedagogical policies directly from student-system interaction data (Chi, Jordan, & VanLehn, 2014; M. Chi, VanLehn, Litman, & Jordan, 2012, 2011). In those studies we applied Reinforcement Learning (RL) to induce the policies directly from an exploratory corpus. The exploratory corpus was collected by having the ITS make random decisions when interacting with students. In our prior work (M. Chi et al., 2012, 2011), we used the induced pedagogical policies to decide when to provide an example step and when to require students to solve it themselves. We found that when students were all given the same FWEs, RL-induced policies significantly improved students' learning gains compared to poor pedagogical policies and

random decisions. On the other hand, we also found that students can still learn from these FWEs even with poor policies. This was likely due to the content exposure and available practice opportunities. In post-hoc comparisons, different versions of "poor" faded policies were compared, and no significant difference was found between them either in terms of learning outcomes or time on task.

In this study, we will investigate the impact of pedagogical policies on learning across two different granularities of decisions. For the purposes of this study we used a random pedagogical policy on both problem-level and step-level decisions. By making random decisions, we expect the number of example steps to be equivalent among the FWE, WE/PS, and ALL conditions. We are interested in investigating the impact of random pedagogical decisions on student learning across FWE, WE/PS, and ALL conditions, and how they will differ from the WE and PS-only groups. Therefore that content will be strictly controlled to be the same across conditions.

We will examine students' performance on a pre- and post-test, as well as their time on task. In light of prior research, we expect that there will be no significant learning difference among the five conditions, since the system is making random decisions on the WE/PS, FWE and ALL conditions. For time on task, given the number of steps that students need to complete, we expect: $WE < WE/PS = FWE = ALL < PS$.

Methods

Participants

The study was conducted in two sections of the Discrete Mathematics for Computer Science course offered at North Carolina State University in the Fall of 2014. 266 undergraduate students were assigned to complete the task as one of their regular homework assignments during the last two weeks of the class.

Conditions

The participants were randomly distributed into five conditions. We used balanced random assignment stratified by course section and performance on a prior class exam. The group sizes were as follows: $N = 31$ for WE[1], $N = 58$ for WE/PS, $N = 59$ for FWE, $N = 59$ for ALL, and $N = 59$ for PS.

Due in part to a holiday break, preparations for final exams, and length of the experiment, only 163 students completed the experiment. Four students were excluded from our subsequent analysis because they performed perfectly on the probability pre-test. The remaining 159 students were distributed as follows: $N = 21$ for WE, $N = 38$ for WE/PS, $N = 37$ for FWE, $N = 34$ for ALL, and $N = 29$ for PS.

We performed a χ^2 test of independence to examine the relation between completion rate and condition. We found no significant differences among five groups: $\chi^2(4, N = 266) = 4.12, p = 0.39$.

[1]Note that a smaller portion of students were assigned to the WE condition. This is because another purpose of this study was to collect exploratory data in order to apply RL to induce adaptive pedagogical policies.

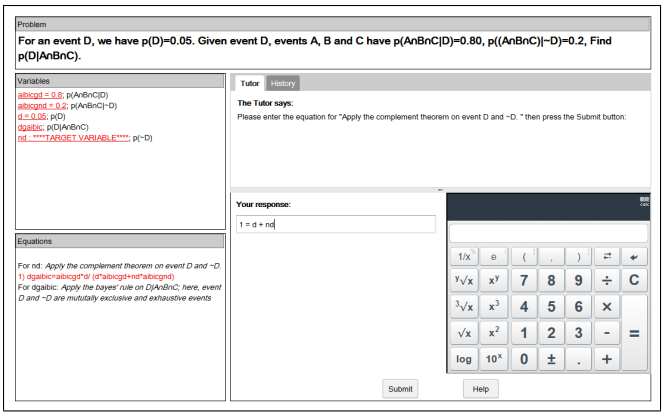

Figure 1: The Pyrenees tutor's interface.

Probability Tutor

The ITS involved in this study is called Pyrenees, a web-based ITS for probability. Pyrenees teaches students 10 major principles of probability, such as the Complement Theorem and Bayes' Rule. Prior studies have shown that Pyrenees is effective and have compared it to Andes, another well-evaluated ITS (Vanlehn et al., 2005). Pyrenees has outperformed Andes in both physics (VanLehn et al., 2004) and probability (Chi & Vanlehn, 2007; 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. In the dialog window, Pyrenees can provide messages to the student, such as explaining a worked example step, or prompting them to complete the next step. The student can enter responses below such as writing an equation or giving the answer to a multiple-choice question. Any variables or equations that are defined through this process are displayed on left side of the screen for reference. Once students submit an answer, Pyrenees provides immediate feedback on whether or not it was correct.

In addition to providing immediate feedback, Pyrenees can also provide on-demand hints, either explaining what is wrong with an incorrect step or prompting the student with what they should do next. Because Pyrenees requires students to follow the Target Variable Strategy, it knows exactly what step the student should be doing next, so it gives specific hints. In Pyrenees, help was 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

this study, Pyrenees had three basic modes. In the WE or PS modes, each step was performed either by the tutor or student throughout the problem. In the FWE mode, there was a 50% chance at each step for either the student or the tutor to solve the step.

Procedure

The study was organized into four phases: 1) pre-training, 2) pre-test, 3) training on Pyrenees, and 4) post-test.

During pre-training, 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 14 problems. 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 five conditions received the same 12 problems in the same order on Pyrenees. 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 PS, WE or FWE, based upon the students' experimental condition. All students could access the corresponding pre-training textbook.

Finally, all students took a post-test which had 20 problems in total. 14 of the problems were isomorphic to the pre-test problems given in phase 2. The remainder were non-isomorphic multiple-principle problems.

The only procedural differences among the five conditions occurred within Pyrenees when the system chose whether to provide a worked example problem, example step, or to require the student to engage in problem-solving. Apart from this behavioral difference the system was identical for each student.

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

The conditions were balanced in terms of students' incoming competence. Prior to the intervention in Phase 3 we found no significant differences among the five conditions according to a range of measures. These measures include (1) the probability pre-test with respect to students' test scores on three types of problems: single-principle, multiple-principle, and overall across all 3 scoring rubrics; and (2) the students' performance during probability pre-training on all three types of problems. Thus, despite attrition, the conditions remained balanced in terms of incoming competence. We will now compare students' learning performance in the post-test and training time across the five conditions. We discuss each comparison in turn.

Learning Performance

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 $F(4,154) = 118.59$, $p < 0.0001$. On the isomorphic questions, all five groups of students scored significantly higher on the post-test than on the pre-test, $F(1,20) = 8.75$, $p < 0.009$ for WE, $F(1,37) = 25.66$, $p < 0.001$ for WE/PS, $F(1,36) = 29.34$, $p < 0.001$ for FWE, $F(1,33) = 20.61$, $p < 0.001$ for ALL, and $F(1,28) = 55.04$, $p < 0.001$ for PS. Therefore all five conditions made significant gains from pre- to post-test. This suggests that the basic practices 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 compares the pre-test, isomorphic post-test and overall post-test scores among the five conditions. The second column in Table 1 lists the number of students in each condition who completed the study. The third, fourth, and fifth columns list the mean and SD for the pre-test, isomorphic post-test (14 isomorphic questions), and overall post-test scores. Overall, no significant differences were found among

Table 1: Test scores across conditions.

Cond	# Stud	pre-test	Iso Post	Overall Post
WE	21	.687(.160)	.789(.187)	.650(.197)
WE/PS	38	.658(.165)	.774(.130)	.630(.167)
FWE	37	.625(.134)	.736(.159)	.588(.145)
ALL	34	.664(.181)	.803(.136)	.651(.159)
PS	29	.618(.155)	.802(.118)	.645(.139)

the five conditions on any of the learning outcome measures: $F(4,154) = 0.871$, $p = 0.483$ (pre-test), $F(4,154) = 1.25$, $p = 0.29$ for (isomorphic post-test questions); and $F(4,154) = 0.98$, $p = 0.42$ (overall post-test).

We also compared the adjusted post-test and NLG scores across all five conditions. The adjusted post-test scores were compared via an ANCOVA with the corresponding pre-test score used as a covariate. The NLG score measures the students' learning gains *irrespective of their incoming competence*: $NLG = \frac{post-pre}{1-pre}$. Here 1 is the maximum score. Again, no significant difference was found among the conditions.

Training Time

Table 2 shows the average amount of total training time (in minutes) students spent on Pyrenees for each condition. A one-way ANOVA showed significant differences among the five groups: $F(4,154) = 26.91$, $p = 0.000$.

Subsequent pairwise t-tests showed that the WE condition spent significantly less time than the others: $t(57) = -5.22$, $p < 0.001$, $d = 1.33$ (WE/PS); $t(56) = -6.22$, $p < 0.001$, $d = 1.95$ (FWE); $t(53) = -6.26$, $p < 0.001$, $d = 1.70$ (ALL); and $t(48) = -8.93$, $p < 0.001$, $d = 2.55$ (PS).

Similarly, we found that the WE/PS condition spent significantly less time than FWE, ALL and PS conditions: $t(73) = -2.77$, $p < 0.008$, $d = 0.64$ (FWE); $t(70) = -2.49$, $p < 0.016$, $d = 0.58$ (ALL); $t(65) = -6.96$, $p < 0.001$, $d = 1.67$ (PS) respectively.

Finally, while we found no significant time on task differences between the FWE and ALL conditions, ($t(69) = 0.395$, $p = .69$, $d = 0.09$). They both took significantly less time than the PS condition: $t(64) = -3.60$, $p = .001$, $d = 0.89$ (FWE); $t(61) = -4.14$, $p < .001$, $d = 1.04$ (ALL) respectively.

Overall, with respect to time on task. we found that: $WE < WE/PS < FWE = ALL < PS$. In fact, the WE group only took around 43% as much training time as FWE and 32% as much as PS but reached the same learning gains as other conditions.

Finally, we conducted a one-way ANCOVA to determine if there was any statistically significant differences among the five groups on their overall post-test scores. We used both the pre-test score and total training time as covariates; no significant difference was found: $F(4,152) = 1.18$, $p = 0.32$.

Table 2: Time on task per condition.

Cond	# Student	Time (in minutes)
WE	21	47.96 (39.27)
WE/PS	38	92.23 (25.79)
FWE	37	112.80 (37.50)
ALL	34	109.48 (32.85)
PS	29	146.40 (37.88)

Discussion and Conclusion

In this study, we used an ITS called Pyrenees to compare five tutorial conditions: WE, PS, FWE, WE/PS, and ALL. For the WE/PS, FWE, ALL conditions, the tutor used a random policy to decide when to give students a worked example problem (or example step) or to ask them to solve the problem (or step). Our results showed that all five conditions learned significantly after training on Pyrenees, and no significant difference was found on all of our learning measures including the pre-test, isomorphic post-test, and overall post-test scores.

This happened despite the fact that the pedagogical strategies employed for the WE/PS, FWE, and ALL conditions were random and thus were rather ineffective. They did not adapt to the students and thus may not have been able to make a positive impact on students' performance beyond the baseline provided by content exposure. Here the basic practices and problems, domain exposure, and interactivity of Pyrenees set a minimum bar for students' learning that the pedagogical strategies, however poor, could not undercut. This lack of a significant difference among the five conditions supports our hypothesis and is consistent with results from prior studies (M. Chi et al., 2012, 2011).

Previously, we found that students' learning performance could be improved by employing effective pedagogical strategies (M. Chi et al., 2012, 2011). However, in that study no significant difference was found in terms of *time on task* between the students trained on the system with effective pedagogical policies and those with ineffective pedagogical policies. In this study, we showed that different granularities of pedagogical decisions can make a significant difference in students' time on task.

Much of the prior research has shown that WE can be as effective as tutored PS but the former often take significantly less time than the latter. One potential explanation for this time difference is that the students in the PS condition have to do more work. Given that the same amount of work was expected for students in the WE/PS, FWE, and ALL conditions, we hypothesized that: $WE/PS = FWE = ALL$. However, our results suggest that for time on task, $WE/PS < FWE = ALL$. WE/PS spent significantly less time than both FWE and ALL.

There are many possible explanations for why the FWE group took longer time than WE/PS group. Since both WE/PS and FWE groups get the same random decisions, we hypothesize that the granularity of the decision must therefore play an important role. Solving a problem in domains such as probability consists of applying domain principles in a valid logical order. Students' later steps are directly dependant upon what they have done previously. This partial dependence may force students in the FWE condition to pay more attention to not only tutor-solved steps but also what their own steps.

Additionally, Pyrenees' instructional methods may explain some of the extra time taken by the FWE condition compared with WE/PS condition. If the tutor solves a problem in a way that is unexpected to the student, the student will require

extra time to process the tutor's intentions and continue its progress. These tutor-solved steps may act as constraints on the student's problem solving process. There are many possible strategies for solving a problem, and Pyrenees uses one specific strategy which may not be intuitive for the student. Thus these solved steps may lead students onto a different solution path which is outside of their expectations. We are currently in the process of analyzing our log files to determine why this occurred. Why did the same random pedagogical policy improve efficiency when applied at the problem level more than at the step level?

Our results from this study suggested that step-level decisions are more sensitive to ineffective pedagogical strategies than problem level decisions. With random decisions, the FWE group not only failed to learn more than WE/PS, they also spent significantly more time.

Overall, this study suggests that different granularities of pedagogical decisions can have a significant impact on students' time on task. The fine-grained interaction steps can have a strong pedagogical impact. Our ultimate goal is to apply RL to induce effective pedagogical policies, at both the problem and step levels, directly from our dataset. This raises an interesting question: with effective pedagogical strategies, will there be a difference in time on task and learning among the five conditions? This is an promising question for future research.

Acknowledgements
This work is supported by the National Science Foundation award #1432156. We would also like to thank the anonymous reviewers for their valuable feedback.

References

Chi, M., Jordan, P. W., & VanLehn, K. (2014). When is tutorial dialogue more effective than step-based tutoring? In *Intelligent tutoring systems - 12th international conference, ITS 2014, honolulu, hi, usa, june 5-9, 2014. proceedings* (pp. 210–219). Retrieved from http://dx.doi.org/10.1007/978-3-319-07221-$0_2$5 doi: 10.1007/978-3-319-07221-$0_2$5

Chi, M., & Vanlehn, K. (2007). Accelerated future learning via explicit instruction of a problem solving strategy. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, *158*, 409.

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).

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

M. Chi, VanLehn, K., Litman, D. J., & Jordan, P. W. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model. User-Adapt. Interact.*, *21*(1-2), 137-180.

M. Chi, VanLehn, K., Litman, D. J., & Jordan, P. W. (2012). An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education.*.

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.

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.

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. *International Journal of Artificial Intelligence in Education*, *15*(3), 147–204.