Comparing Worked Examples and Tutored Problem Solving: Pure vs. Mixed Approaches

Rob Weitz (weitzrob@shu.edu)
Department of Computing and Decision Sciences, Seton Hall University
South Orange, NJ 07079 USA

Ron J. C. M. Salden (rons@cs.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Ryung S. Kim (rkim@wpi.edu)
Department of Epidemiology and Population Health, Albert Einstein College of Medicine
1300 Morris Park Ave, Bronx, NY 10461 USA

Neil T. Heffernan (nth@wpi.edu)
Department of Computer Science, Worcester Polytech Institute, 100 Institute Road
Worcester, MA 01609 USA

Abstract

This paper extends our previous work (Kim, Weitz, Heffernan & Krach, 2009) which compared a "classic" worked examples (WE) condition with a tutored problem solving (TPS) condition. By classic we mean the WE condition does not include tutoring, a self-explanation component, or fading. The aim of the current study was to compare the WE and TPS conditions with a mixed condition, which presents students with WE-TPS pairs. More specifically, for conceptual problems a pure WE condition was compared with a WE-TPS condition and for procedural problems a pure TPS condition was compared with a WE-TPS condition. While overall learning occurred in all conditions no significant differences were found between conditions. Further, our findings echo the results of earlier studies, that students who receive worked examples learn more efficiently – that is, they need significantly less time to complete the same learning material. This is an important finding for educators because building classic worked examples is considerably easier than building tutoring.

Keywords: tutored problem solving; worked examples

Introduction

Research on worked examples (e.g., Sweller & Cooper, 1985; Ward & Sweller, 1990) has demonstrated that when students were presented with example-problem pairs rather than problems only, they could attain higher learning outcomes because their working memory capacity was not overloaded. Worked examples reduce problem solving demands by providing worked-out solutions. Therefore, more of the learners' limited processing capacity (i.e., working memory capacity) can be devoted to understanding the domain principles and their application to the problem at hand (Renkl & Atkinson, 2007).

In recent years, a considerable number of studies have explored the conditions under which examples aid in acquiring cognitive skills (for a review, see Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2005, 2009). While the impressive body of research on worked examples to date has been quite successful, it also has two important shortcomings. Firstly, the studies are mostly conducted in a laboratory setting without being extended to the more challenging authentic classroom setting and secondly, the studies have almost exclusively compared learning by studying examples to untutored problem solving.

One very successful tutored problem-solving approach is the use of Cognitive Tutors (Koedinger, Anderson, Hadley, & Mark, 1997; Koedinger & Aleven, 2007). These computer-based tutors provide individualized support for learning by doing (i.e., solving problems) by selecting appropriate problems to be solved, by providing feedback and problem-solving hints, and by on-line assessment of the student's learning progress. Because such a tutored environment offers a significant amount of guidance it is a much more challenging control condition than traditional problem solving against which to measure the possible beneficial effects of worked examples. Additionally, research on Cognitive Tutors aims to be examined in the authentic classroom setting (*in vivo* experimentation) which creates a much richer and challenging testing environment compared to a laboratory setting.

Several recent studies have embedded worked examples in a variety of Cognitive Tutors and investigated whether the examples still had beneficial effects over the tougher tutored control condition (e.g., Salden, Aleven, Schwonke, & Renkl, in press; Schwonke et al., 2009). More specifically, these studies proved that replacing some problems with worked examples further enhances student learning by reducing instructional time to the same outcome and/or increasing student outcomes than tutored problem solving.

Of particular interest for the current paper are the studies by McLaren, Lim, and Koedinger (2008) which compared worked examples with pure TPS (tutored problem-solving) within a Stoichiometry Cognitive Tutor. The results across three studies showed that the students who received worked examples did not learn more than the students who received pure TPS. This reinforces the prior claim that TPS poses a new challenge for the research on worked examples in being a much harder control condition. However, an important consistent finding in the McLaren et al. studies is that the students who received worked examples did learn more efficiently, using 21% less time to complete the same problem set. If these results were to scale across a 20-week course, students could save 4 weeks of time – yet learn just as much.

Another educational system that provides tutored problem solving in classroom settings is the Assistment system (e.g., Razzaq & Heffernan, 2009). Additionally, a further similarity between the Cognitive Tutors and Assistment is their focus on *in vivo* experimentation which allows for an examination of student learning in its most authentic environment. In a previous *in vivo* study Kim, Weitz, Heffernan and Krach (2009) explored the benefits and limitations of worked examples by comparing a "pure worked-example" (pure WE) condition with a pure TPS condition on conceptual and procedural learning. "Pure" means that students in the TPS condition received only TPS remediation while students in the WE condition received solely WE remediation. Note that in contrast to the Cognitive Tutor studies cited above, neither condition included a self-explanation component. The results showed that for conceptual problems students learned more in the pure WE condition and for procedural problems students learned more in the pure TPS condition. In agreement with the findings by McLaren et al. (2008), pure WE was more efficient – that is, it took students less time to do pure WE than TPS.

The current paper addresses a study which extends this research by comparing the best pure condition from the previous study with mixed approaches. That is, for conceptual problems we compare learning resulting in a pure WE condition to one that mixes WE and TPS. For procedural problems we compare a pure TPS approach to a condition that mixes WE and TPS. With these conditions we examine whether the findings of the previous study will still hold. More specifically, if pure WE is better than WE-TPS for conceptual problems and pure TPS is better than WE-TPS for procedural problems it could provide further evidence that examples are always better for conceptual learning and tutored problem solving is always better for procedural learning.

Overall, the outcomes of this study will suggest important guidelines for designing intelligent tutors and provide meaningful insights into the students' learning process. In practical terms, building worked examples is significantly less time consuming than building tutoring; if worked examples are as good as or better than traditional intelligent tutoring – and more efficient – this is valuable information.

The Experiment

Our study involved college students taking an introductory statistics course. Statistics is a good domain for this research as it includes both procedural and conceptual components.

Student Characteristics

Participating students were enrolled in an introductory statistics course at Worcester Polytechnic Institute (WPI), a private university specializing in engineering and the sciences. Eighty-four students, mostly first-year engineering students, participated in the experiment, which was conducted as one of the course's regular lab session.

Design

The tutorial and test problems were typical of problems given in introductory statistics courses. The subject matter concerned one-sample confidence intervals of the mean and was taught on days preceding the experiment. There were no assignments or tests on these topics due before the experiment. At the start of the experiment, students were randomly assigned to one of four groups with equal probability; the resulting student numbers are outlined in Table 1. Note that the mild non-uniformity in numbers is caused by randomness.

Table1: Initial Student Allocation to Groups

Group	Procedural Problem Tutorials	Conceptual Problem Tutorials	No. Students
1	WE-TPS	WE-WE (pure WE)	29
2	WE-TPS	WE-TPS	21
3	TPS-TPS (pure TPS)	WE-WE (pure WE)	17
4	TPS-TPS (pure TPS)	WE-TPS	17

This design allows the comparison of WE-TPS with pure TPS on procedural problems by comparing the performance of students in groups 1 and 2 with that of students in groups 3 and 4. Likewise pure WE may be compared with WE-TPS for conceptual problems by comparing the performance of students in groups 1 and 3 with that of students in groups 2 and 4.

An example of a procedural problem is one that asks the student to calculate a confidence interval. A conceptual problem might ask about the impact on the width of a confidence interval if the sample size is doubled. Procedural problems align with the NSF-Funded ARTIST project guidelines (https://app.gen.umn.edu/artist/glossary.html) for "statistical literacy," and conceptual problems with "statistical reasoning" and "statistical thinking" (delMas, 2002).

Of the eighty-four students that participated we excluded ten students who spent less than 5 minutes in the post-test from our analysis due to time and motivation issues. Further, we eliminated eleven students from the conceptual part of the analysis as they did not complete the conceptual problems in the tutorial. Note that the conceptual problems were towards the end of the tutorial. The final number of students used in each condition of the analysis is provided in Table 2. In both instances where we eliminated students from the analysis, roughly the same numbers were removed from each group.

Table 2: No. Students in Each Group

Group	Procedural Problem Tutorials	No. Students	Conceptual Problem Tutorials	No. Students
1	WE-TPS	25	WE-WE (pure WE)	21
2	WE-TPS	19	WE-TPS	18
3	TPS-TPS (pure TPS)	13	WE-WE (pure WE)	10
4	TPS-TPS (pure TPS)	15	WE-TPS	12
Total		72		61

The ASSISTment System

Our experiment was conducted via the ASSISTment intelligent tutoring system (http://assistment.org). It is similar to the CTAT system (Koedinger Aleven, Heffernan, McLaren, & Hockenberry, 2004), used in some of the previously mentioned studies (McLaren, et al., 2008), in that the system provides the student with tutoring on the individual steps of a problem, generally breaking a problem down into 3-4 steps. For each step, a student is asked to provide an answer, and receives feedback on their answer until they get it correct. Our system differs from the CTAT structure in several ways including that there is only one solution path and the intermediate solution goals are highlighted. A further difference is that our system does not contain a self-explanation component.

The tutorials were comprised of three pairs of problems.[1] Each pair was comprised of two isomorphic problems. The first two pairs were procedural problems and the last problem pair was conceptual in nature.

TPS-TPS (Pure TPS) Condition

For this study the ASSISTment system was modified to force students to work through the TPS for the first problem of each pair. This "forced TPS" approach ensures that each student experiences tutoring. After completion of the first problem of the pair, the student is presented with an isomorphic problem and is asked by the system to provide the answer. If the student gets this second problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides TPS support.

In terms of tutoring, the system gives immediate corrective feedback for each attempt at solving a problem. The student can choose to answer the problem or ask the system to break it into steps. However, if the student answers incorrectly the system automatically breaks the problem into steps. For each step, the student will receive immediate feedback and has the possibility to request hints.

WE-WE (Pure Worked Example) Condition

Firstly, it should be noted that "classic" worked examples are used which do not contain tutoring, a self-explanation component, or fading.

The student is presented with the same first problem as in the TPS condition, and a worked solution including the necessary steps to take in that problem. After studying the worked example, the student is then presented with an isomorphic problem, the exact same second problem as in the TPS condition, which the student is expected to solve. The student has access to the first WE while trying to solve the second. If the student gets this second problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides the worked solution for the problem for review by the student.

WE-TPS (Mixed) Condition

The student is presented with the first problem and a worked solution to that problem, similar to the WE-WE condition. After studying the worked example, the student is then presented with the second problem. If the student gets this isomorphic problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides TPS support. See Table 3 for an overview of the problem pairs for each experimental condition.

Table 3: A Comparison of Intelligent Tutoring and Worked Examples

<table>
<tr><th></th><th>Pure TPS (TPS-TPS)</th><th>Pure WE (WE-WE)</th><th>Mixed (WE-TPS)</th></tr>
<tr><td>First Problem</td><td>Student studies with forced TPS</td><td>Student studies WE</td><td>Student studies WE</td></tr>
<tr><td rowspan="2">Second Problem</td><td colspan="3">Student is given opportunity to solve the problem. If student answer is incorrect, the problem is marked incorrect and,</td></tr>
<tr><td>TPS is provided</td><td>WE is provided</td><td>TPS is provided</td></tr>
</table>

The students were allowed to work though both tutorials at their own pace. One week before the experiment students were given a ten minute tutorial on how to use the ASSISTment software for which they were allowed to work through at their own pace. They created an account for

[1] All of our materials are available at http://teacherwiki.assistment.org/wiki/index.php/CogSci2010 so other researchers can inspect them.

themselves, and enrolled in their professor's class. They got a few minutes of practice with the system during which they did one worked example and one tutored problem solving.

The experiment consisted of three parts: pre-test, tutorial, and post-test. The pre-test and post-test were identical, and were comprised of four procedural problems and three conceptual problems.

The students were given 20 minutes to go through the pre-test without any feedback, 40 minutes for their tutorials, and 20 minutes for the post test (see Table 4). In order to control time, students were not supposed to be allowed to move to the next part of the experiment until a designated time passed. However, in practice we actually had some students not following the directions when asked to move to the next part of the experiment.

Table 4: Outline of Experiment

One Sample Confidence Interval for the Mean

Several Days Prior to Lab Session
- Lecture on the topic

During Lab Session
1. Pre-Test (20 min; students' initial knowledge)
    - 20 minutes
    - Four procedural and three conceptual.
2. Condition (Tutorials)
    - 40 minutes
    - 3 pairs of Problems: 2 procedural, one conceptual (3 parts)
3. Post-Test (20 min; students' knowledge after trial)
    - Same problems as Pre-Test

Results

Learning by Problem

Table 5 provides the percentage of students across all conditions getting each problem correct on the pre- and post-tests. Student learning is clearly evident for all items ($z = 3.78$, $p < .001$, $d = 1.36$).

Following the approach in item response theory (Embretson & Reise, 2002), throughout the remainder of this section, we summarize student performance on a problem or on a category of problems by the adjusted percent correct, that is, the percent correct adjusted by problem difficulty. We then define learning for problems as the difference in adjusted percent correct between post-test problems and the corresponding pre-test problems.

Qualitatively speaking, this means that students who correctly answer harder items will get more credit than students who correctly answer easier items.

We determined these adjusted values using a generalized linear mixed effects model, also referred to as a generalized linear multilevel model (Bates & Sarkar, 2007; Rabe-Hesketh, Skrondal, & Pickles, 2005).

Table 5: Learning by Problem

Problem	**Percent Students Correct**	
Procedural	**Pre-Test**	**Post-Test**
1	5.6%	58.3%
2	16.7%	73.6%
3	15.3%	43.1%
4	30.6%	54.2%
Conceptual		
1	11.5%	27.9%
2	73.8%	91.8%
3	37.7%	52.5%

Learning by Condition

Table 6 below summarizes the learning results by type of tutorial. So, for example, for procedural problems, students in the WE-TPS improved their performance by 40.1% (54.9% - 14.8%).

Table 6: The Adjusted Percent Correct

		Percent Correct
Procedural Problems	Pre-Test	14.8%
	WE-TPS	54.9%
	TPS-TPS	63.3%
Conceptual problems	Pre-Test	37.8%
	WE-TPS	61.2%
	WE-WE	61.7%

For procedural problems, students in the pure TPS condition outperformed students in the WE-TPS condition. However, this difference (63.3% vs. 54.9%) is not significant ($p = 0.23$). Likewise, for conceptual problems, the results indicate a small benefit for the pure WE condition over the WE-TPS condition (61.7% vs. 61.2%); these results are clearly not statistically significant ($p = 0.95$).

Learning Time

As noted earlier, previous research has consistently indicated that doing worked examples requires significantly less time for students than tutored problem solving.

Table 7: Times for Students to do the Tutorial Problems

		n	Mean	SD
Procedural Problems	WE-TPS	44	18.03	7.79
	TPS-TPS	28	26.00	10.63
Conceptual Problems	WE-TPS	30	6.70	2.53
	WE-WE	31	6.60	3.17

Table 7 provides the mean and standard deviation of student times in each group for both types of problems in the tutorial. Focusing on the procedural problems, we see the same pattern here with students in the WE-TPS condition taking less time than those in the pure TPS condition. These results are statistically significant (t = 3.42, p < .01, d = 0.86).

As the conceptual problems were placed after the procedural problems in the tutorial (condition), the above-reported conceptual times may have been artificially constrained. We observed that procedural times and conceptual times are negatively correlated – an indication that individuals who spent a lot of time on the procedural problems ran out of time on the conceptual problems. Note (again) that we excluded students who did not finish the conceptual part of the tutorial from our post-test results.

Discussion

This paper extends our previous work (Kim et al., 2009) comparing pure WE with pure TPS approaches where the results showed that pure WE was more effective for conceptual problems, while pure TPS was more effective for procedural problems. Furthermore, pure WE was more efficient in that students took less time to work through the WE condition than the TPS condition. The aim of the current study was to compare these pure WE and TPS conditions with a mixed condition, which presents students with WE-TPS pairs. More specifically, for conceptual problems a pure WE condition was compared with a WE-TPS condition and for procedural problems a pure TPS condition was compared with a WE-TPS condition.

While overall learning occurred in all conditions and the pure methods come out ahead in terms of student learning, the results are not statistically significant. More specifically, there were small non-significant differences favoring the pure WE condition for conceptual problems and the pure TPS condition for procedural problems. Furthermore, the efficiency effect of the previous study was replicated meaning that students needed less time to complete the WE tutorial than the TPS tutorial. These results are similar to the findings of McLaren et al. (2008) who also did not find significant differences in student learning but who also found that students who received worked examples did learn more efficiently, using 21% less time to complete the same problem set.

It should be noted that McLaren et al. (2008) and other studies use worked examples in combination with tutoring, a self-explanation component, and/or fading. In contrast to those studies, the worked examples used in our experiments are "classic" worked examples which do not include these extra elements. While these elements can undoubtedly improve learning our studies shows that the use of classic worked examples in tutored problem solving can still result in similar outcomes without any detrimental effect on student learning. As such, the replication of the time efficiency effect makes a strong case for the use of classic worked examples in tutored problem solving.

A possible explanation for the lack of significant main differences could be offered by Rittle-Johnson, Siegler, and Alibali (2001) who stated that effects of worked examples on procedural tasks might be more indirect and need more time to materialize. In fact, other studies (e.g., Anthony, 2008; Salden, et al., 2009) that compared TPS and WE also did not find significant differences on the post-test but they did find positive effects favoring the WE conditions on a delayed post-test.

A further explanation might be found in the time limit that we imposed on the students. We had to exclude eleven students from our data analysis because they did not have enough time to complete the conceptual problems in the tutorial. Had we given them more time then we might have been able to observe possible conceptual learning differences.

For future studies we would like to explore other factors which could deepen the insights on the beneficial effects of worked examples in TPS. One possible factor is students' prior knowledge which can have a mediating influence on their learning progress if students with differing prior knowledge levels work through the same training material. In line with the expertise reversal effect (Kalyuga, 2007), students who have a high knowledge level could even experience detrimental effects of worked examples. In future studies we could use the pre-test scores to check if such differences in prior knowledge exist and use this information to determine what experimental condition a student ought to be in.

Furthermore, in accordance with Schwonke et al. (2009) we could try to add thinking aloud to differentiate learning effects. In their first study Schwonke et al. also did not find student learning differences but they used thinking aloud protocols in their second study which subsequently showed a higher learning gain in terms of conceptual knowledge for the example-enriched TPS condition. It is plausible that students who were thinking aloud about the worked examples engaged in deeper processing of conceptual knowledge than the students in the control TPS condition without examples. Consequently, being able to talk aloud about the worked examples might have led to the observed higher learning gain.

Finally, adding a delayed post-test to our future studies might also enable us to differentiate differences between TPS and WE-TPS conditions. Rittle-Johnson et al.'s statement that the effects on procedural tasks might need time to materialize has been proven to be accurate in other studies compared tutored problem solving and worked examples (e.g., Anthony, 2008; Salden, et al., 2009). More specifically, if worked examples support students in engaging with the conceptual knowledge more deeply but only over longer period of time then this has significant implications for developing computer-based learning programs which use worked examples.

In conclusion, our results extend the previous findings of TPS and WE-TPS comparisons. The tutored problem solving environment poses a more challenging control condition than traditional problem solving conditions. Yet across two studies and in line with the McLaren et al. (2008) studies we consistently found that students needed less time to complete the training phase when being presented with worked examples without any loss of student learning on the post-test. These results are even more impressive as our experiments used classic worked examples, which do not offer tutoring or a self-explanation component, as those used by McLaren et al. (2008).

This is an important finding for educators because building classic worked examples is considerably easier than building tutoring and in fact is easier than building worked examples with more features. Future studies are needed to further investigate under what circumstances classic worked examples can make computer-based instructional materials more efficient.

Acknowledgments

This works was funded in part by the National Science Foundation CAREER Award the US Department of Education (R305K030140) but the opinions are solely those of the authors. Furthermore, the authors would like to express their gratitude to Professor Joseph D. Petruccelli for his invaluable help in running the study.

References

Anthony (2008). *Developing handwriting-based Intelligent Tutors to enhance mathematics learning*. Unpublished doctoral dissertation, Carnegie Mellon University, USA.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*, 181-214.

Bates, D. & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.9975-12, http://CRAN.R-project.org/.

delMas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, 10 (3). Retrieved from www.amstat.org/publications/jse/v10n3/delmas_discussion.html

Embretson, S. E. & Reise, S. (2002). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kalyuga, S., (2007). Expertise reversal effect and its implications for learner-entailed instruction. *Educational Psychology Review, 19*, 509-539.

Kim, R, Weitz, R., Heffernan, N. & Krach, N. (2009). Tutored problem solving vs. "pure" worked examples. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*, 239-264.

Koedinger, K. R., Aleven, V., Heffernan, N. T., McLaren, B. M., & Hockenberry, M. (2004). Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. In Lester et al (Eds.) *Proceedings of 7th Annual Intelligent Tutoring Systems Conference*, Springer, 162-173.

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*, 30-43.

McLaren, B. M., Lim, S., & 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 B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2176-2181). Austin, TX: Cognitive Science Society.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*, 301-323.

Razzaq, L., & Heffernan, N.T. (2009). To tutor or not to tutor: That is the question. In Dimitrova, Mizoguchi, du Boulay and Graesser (Eds.), *Proceedings of the Conference on Artificial Intelligence in Education*. pp. 457-464.

Renkl, A. (2005). The worked-out-example principle in multimedia learning. In R. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning*. Cambridge, UK: Cambridge University Press.

Renkl, A. (2009). *Towards an instructionally oriented theory of example-based learning*. Manuscript submitted for publication.

Renkl, A., & Atkinson, R. K. (2007). An example order for cognitive skill acquisition. In F. E. Ritter, J. Nerb, E. Lehtinen, T. O'Shea (Eds.), *In order to learn: How the sequence of topics influences learning* (pp. 95-105). New York , NY : Oxford University Press.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346-362.

Salden, R. J. C. M., Aleven, V., Renkl, A., & Schwonke, R. (2009). Worked examples and tutored problem solving: Redundant or synergistic forms of support? *Topics in Cognitive Science, 1*, 203-213.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., Salden, R. J. C. M. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior, 25*, 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*, 59-89.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*, 1-39.