

Tutored Problem Solving vs. “Pure” Worked Examples

Ryung S. Kim (rkim@wpi.edu)

Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road Worcester, MA 01609 USA

Rob Weitz (weitzrob@shu.edu)

Department of Computing and Decision Sciences, Seton Hall University, South Orange, NJ 07079

Neil T. Heffernan (nth@wpi.edu)

Department of Computer Science, Worcester Polytechnic Institute

Nathan Krach (nkrach@wpi.edu)

Department of Computer Science, Worcester Polytechnic Institute

Abstract

At present a handful of comparisons have been made of different variants of worked examples and tutored problem solving. There is some evidence to report a benefit of adding worked examples (WE) to current tutored problem solving (TPS) environments. Our research investigated how a “pure” WE condition could compete with a TPS condition. By pure we mean the WE condition does not include tutoring, a self-explanation component, or fading. We report on two experiments. We showed statistically significant evidence of learning benefits, both in terms of amount learned and rate of learning, from assigning WE to conceptual problems and TPS to procedural problems. Higher prior knowledge students tended to learn more with WE, and those with low knowledge tended to learn more with TPS, but these results were not significant. We found no statistically significant interaction between student preferences for one approach or the other and their performance. These results have important practical ramifications and raise interesting questions regarding the nature of student learning.

Keywords: tutored problem solving, worked examples.

Introduction

This study compares student learning using two approaches: tutored problem solving (TPS) and worked examples (WE). We measured how much each student learned as well as the time spent they spent in each condition. The study included both procedural and conceptual problems. This research is relevant to those working in instructional technology and to anyone interested in the nature of student learning. Certainly, from a practical perspective if learning under worked examples can be comparable or better than tutored problem solving, then we can save the significantly more time, money and effort on building tutors.

There is a long history of research in TPS and WE separately, and very little research comparing the two. Our work contributes in three ways. First we compare “pure” TPS with “pure” WE conditions. Students in the TPS condition received TPS remediation, while students in the WE condition received solely WE remediation (as

opposed to TPS remediation, which appears to be the case in previous studies). Also, in this study neither condition included a self-explanation component. In other words, the WE condition was a “passive instructional event” (Koedinger and Alevan 2007). Second we examined the effect of prior knowledge in the subject area as a mediating factor. Third we investigated how well a student’s preference for a particular form of instruction predicted which approach was actually superior for that student in terms of learning outcomes.

The area of instruction was college-level introductory statistics. Problem solutions generally required multiple steps. The domain is naturally suited to both procedural and conceptual problems. We conducted two experiments; the first focused on the application of the binomial and Normal probability distributions and the second dealt with confidence intervals.

Simple Problem Solving vs. Worked Examples

A number of studies have shown the benefits of learning from WE. Ward and Sweller (1990) and Sweller and Cooper (1985) compared simple problem solving with a condition alternating WE with problem solving. Atkinson, Derry, Renkl, Wortham (2000) provides a comprehensive review of the WE (vs. simple problem solving) literature with a focus on how best to design WE. One of their overarching conclusions (p. 197) is that “students who self-explain tend to outperform student who do not.”

Renkl, Atkinson and Maier (2000) and Renkl, Atkinson, Maier, & Staley (2002) explored the effectiveness of fading (successively removing worked-out solution steps) WE vs. traditional WE. Atkinson, Renkl and Merrill (2003) combined fading with prompts “designed to encourage learners to identify the underlying principle illustrated in each worked-out solution step.” They reported improved far transfer over WE with fading alone.

Intelligent Tutoring vs. Worked Examples

Koedinger and Alevan (2007) review the literature regarding adding worked examples to cognitive tutors.

McLaren, Lim and Koedinger (2008a, 2008b) compare a cognitive tutor with a WE in the domain of chemistry (stoichiometry). The WE condition included an interactive self-explanation component. (The explanations are checked for correctness.) They found that students in the WE conditions did not learn significantly more than students in the TPS condition, however the WE condition was more efficient,

Schwonke et al. (2007) and Schwonke et al. (2009) describe two studies, both comparing a standard cognitive tutor with one augmented by faded worked examples (in the field of high school geometry). Both conditions included an interactive self-explanation element. The results of the first study showed no difference in conceptual or transfer learning though the WE group took less time. The second experiment indicated an advantage to WE for conceptual learning, no difference regarding procedural learning and, again, a time advantage for WE. There was no difference in students' transfer knowledge.

Salden, Alevén, Renkl, and Schwonke (2008) built on the above work, this time adding an adaptive fading WE condition to the cognitive tutor and fixed-fading WE conditions. (Adaptive here means that the rate of fading is based on student's level of understanding.) The two experiments (lab and classroom) they conducted indicated an advantage to the adaptively faded condition.

In all of the above cases, the WE example condition included self-explanation requirements and the WE condition provided tutoring support when the student was unable to solve the isomorphic problem. In the experiments described below we instead use a "pure" worked example condition that does not include any intelligent tutoring, self explanations or fading of prompts. This condition is meant to represent a "cleaner" test of the WE condition compared to TPS alone.

The Experiments

As noted previously, 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. The problems we categorized as conceptual measure what Garfield (2002) has called as the third level of statistical reasoning, or transitional reasoning. They measures student's ability "to correctly identify one or two dimensions of a statistical process without fully integrating these dimensions, such as, that a larger sample size leads to a narrower confidence interval, that a smaller standard error leads to a narrower confidence interval."

We performed two experiments: one for probability distributions (Binomial and Normal) and one for confidence intervals. The methodology undertaken for each experiment is described below.

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. Ninety-five students participated in each experiment. The tutorials and associated assessments were conducted as part of the course's regular statistics lab sessions and as such were integrated elements of the course. Students in this study comprise freshmen (17%), sophomores (61%), juniors (15%), and seniors (7%). Student majors comprise Engineering (65%), Math/Physics/Chemistry (7%), and Social Science/Computer Science/ Biology (27%).

Experiment 1

In the first experiment, we compared the effect TPS and WE on learning of Binomial and Normal probability distribution. The problems were all procedural in nature, and are typical of problems given in introductory statistics courses. The subject matter was taught on days preceding the experiment. There were no assignments or tests on these topics due before the experiment.

Each student was randomly assigned to one of the conditions listed in Table 1. Each student experienced both tutorial types. Each tutorial (TPS or WE) was composed of two, two-part isomorphic problems.

Table1: No. Students in Each Condition of Experiment 1

First tutorial (Method/Topic)	Second Tutorial	Students numbers
TPS/Binomial	WE/Normal	20
TPS/Normal	WE/Binomial	30
WE/Binomial	TPS/Normal	30
WE/Normal	TPS/Binomial	16

The ASSISTment System

Our experiment was conducted via the ASSISTment.org intelligent tutoring system built by a team lead by Heffernan and Koedinger. It's an intelligent tutoring system similar to the CTAT (Koedinger et al. 2004) used in some of the previously mentioned studies (McLaren, Lim & Koedinger, 2008a). It is similar 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 would be asked to provide an answer, and would get feedback on their answer until they got it correct. In this study ASSISTments was used for the TPS condition and the WE condition. In order to help others understand, and possibly replicate our work, we have archived all the study materials (Heffernan 2009). 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.

Tutored Problem Solving Condition

In this study the 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 question 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 (and records that the student was unable to solve the problem).

“Pure” Worked Example Condition

Student is presented with the first problem (same as the first problem under the TPS condition) and a worked solution to that problem. The student is then presented with an isomorphic problem (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 question 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 (and records that the student was unable to solve the problem).

Table 2: A Comparison of Intelligent Tutoring and Worked Examples

	Tutored Problem Solving (TPS)	Worked Examples (WE)
First Problem	Student studies with forced TPS	Student studies WE.
Second Problem	Student is given opportunity to answer the question. If student answer is incorrect, the problem is marked incorrect and,	
	TPS is provided.	WE is provided.

Due to relatively little workload for these tutorials, the students were allowed to work though both tutorials at their own pace. Less than 5% of the students failed to finish the tutorials on time. Students were allowed to move to the second tutorial once they completed the first tutorial.

Statistical Models

Embretson and Reise (2000) suggest the use of statistical techniques that have more power than approaches typically taken in cognitive science research. They advocate the use of item response models, and we take that a step further. Our main measurements are 1) repeated, because the same set of seven problems were used in the pre-test and post-test and 2) binary, because

students either answer each question correctly or not. We use two regression models for repeated binary data: the marginal regression model (Liang and Zeger 1986) and the Generalized Linear Mixed Model, or GLMM (Bates and Sarkar 2007). If we follow the typical approach to use the total score of each test for each student and perform repeated measures ANOVA, we would lose power in our analysis because information of how each student performed on each of the seven pairs of problems is lost. We include details of the statistical models used in this article in Appendix 1.

Result of Experiment 1

In this experiment students worked on two problems, each with two parts. The design for the experiment is a pair-matched randomized design with two conditions (TPS and WE), each with two problems. We define learning in each case (TPS and WE) if the student gets the second isomorphic problem correct. From the number of students with discordant performances between two tutorials (i.e. off-diagonal numbers in table 3), it is clear more students did better under TPS. For example, there were 10 students who got both questions correct after TPS but no problems correct after WE.

From the regression model (M1; Appendix 1), the probability of a student solving the problem after a WE (pooled over two topics) is estimated as 53% and that after a TPS is 63%. This difference between the two tutorials was significant ($p=0.047$).

Table 3: Number of question answered correctly by condition

		Number of questions answered correctly after WE		
		0	1	2
Number of questions answered correctly after TPS	0	14	6	2
	1	11	5	11
	2	10	9	27

Experiment 2

The second experiment utilizes questions from the domain of one-sample confidence intervals of the mean (with continuous observations). There were two types of problems: procedural and conceptual in nature. As in experiment 1, the general concepts of the topic were taught during the days preceding the trial, there were no assignments or tests on this topic due before the trial, and on the day of experiment there was no additional teaching from the instructor prior to the tutorials. The experiment consisted of three parts: pre-test, tutorial, and post-test. The pre-test and post-test were identical, and comprised of three conceptual problems and four procedural problems. In the experiment, we used a completely

randomized design: approximately half the students took the TPS version of the tutorial and the other half took a WE version. The students were given 20 minutes to go through the pre-test without any feedback, 40 minutes for one of two types of tutorial, and 20 minutes for the post test (Table 4). In order to control time, students were not allowed to move to next step until a designated time passed. The design of the tutorials is equivalent to that of experiment 1. That is to say, the problems are presented in pairs using the same approach as experiment 1; the contents of the two tutorials were as equivalent as possible. The tutorials in experiment 2 were comprised of three problems. The first two problems were procedural and the last one (composed of four sub-problems) was conceptual.

Table 4: Outline of Experiment 2
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 (TPS or WE)
 - 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 of Experiment 2

Item-wise Learning Pooled Over the Two Conditions

Student learning was clearly shown in all items. The probability to solve the seven problems in two tests (pre-test/post-test) were estimated at 22%/56%, 11%/21%, 35%/71%, 15%/46%, 75%/87%, 42%/61%, and 33%/69%, respectively (M2c; Appendix 1).

Itemized Learning Pooled Over the Two Conditions

Table 5 shows WE improves learning significantly for both problem types ($p=0.031$ and 0.020), and TPS improves learning significantly for procedural problems ($p < 0.0001$) but only moderately in conceptual problems ($p=0.740$). The table shows the size of learning, in probability estimated by our model (Appendix 1; M2a). For example, on average, TPS helps 23% more students to answer the procedural problems correctly (column four, row three) and only 1% for the conceptual problems (column 4, row 1).

Table 5: Probabilities to solve problem

	PRE	POST	Difference	p-value
--	-----	------	------------	---------

			(learning)	
conceptual:TPS	43%	44%	1%	0.740
conceptual:WE	41%	52%	11%	0.031
procedural:TPS	25%	48%	23%	<0.0001
procedural:WE	31%	44%	13%	0.020

Interaction of Tutor Type and Problem Type We note the following trends in Table 5: in procedural problems, learning from TPS is 10% (11% vs. 1%) greater than that from WE. In conceptual problems, on the contrary, learning from WE is 10% (23% vs. 13%) greater than that from TPS. Neither of these two main differences was significant. However the interaction in learning between tutor types and problem types was significant ($p = 0.0347$) -- that is, γ_1 and γ_2 are significantly different in M2a Appendix 1). In other words, the experiment shows significant evidence of learning benefit from changing tutorial types according to problem types. Put simply, WE was more effective for conceptual problems, while TPS was more effective for procedural problems.

Comparison of Learning Rates

In addition to comparing the amount of learning, we took into account time students spent on the tutorials. While we tried to control for the number of problems done, and not for time, for practical reasons of running a classroom we set a 40 minute time window to complete the problems. We thought 40 minutes represents a reasonable amount of time to complete the problems. Unfortunately, of the 95 student in the experiment, 16 did not complete the problems (13 in TPS condition and 3 in WE condition). Hence, student that did not finish the tutorial were recorded at 40 minutes. This represents uneven censoring of the data. We address this issue at the end of this section. On average, students spent 31 minutes ($s=10.4$) on TPS and 22 minutes ($s=10.0$) on WE. We estimated the learning rates per minute in the two tutorial types. The rate in odds ratio per ten minutes by each condition is in Table 6 (M2b in Appendix 1). Under TPS, odds of solving conceptual problems become 1.43 times greater as students learn 10 more minutes, whereas odds of solving procedural problems stays unchanged. Under WE, odds of solving conceptual problems become 1.29 times greater as students learn 10 more minutes, whereas odds of solving procedural problems become 1.33 times greater. (Table 6)

WE is significantly more efficient in conceptual learning than TPS ($p=0.008$). On the contrary, two tutorial types have similar efficiencies in procedural learning. Figure 1 is a direct representation of Table 6 by converting odds ratio scale to more intuitive probability scale. The low learning rate of TPS for conceptual problems is notable in the figure (flat line). To see if uneven censoring caused bias, we repeated our analyses comparing only students

who completed the tutorials, and the results remained unchanged.

Table 6: Learning Rate (odds ratio per ten minutes)

	TPS	WE	Ratio (WE/TPS)	p-value
Conceptual	1.002	1.29	1.29	0.008
Procedural	1.430	1.33	0.93	0.550

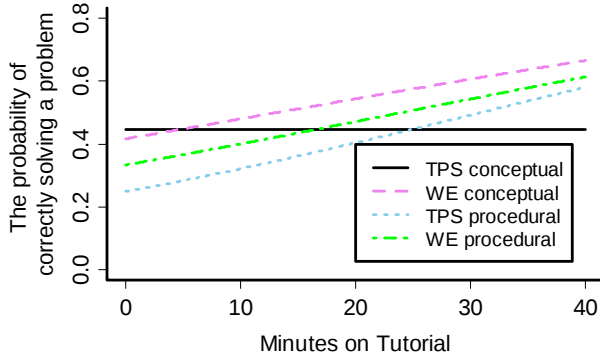


Figure 1: Change of probabilities to solve a problem type as student spends time on each tutorial type.

Interaction of Tutorial Effect and Student Preference

We investigated whether there were interactions between tutorial type and student preference. Students were overwhelmingly positive about the laboratory experience. We asked students to indicate which method of instruction they preferred after the first experiment. Student responses were about evenly split. There was a trend indicating that students performed better on experiment two when given their preferred method of instruction but these results were not statistically significant.

Interaction of Tutorial Effect and Student Prior Knowledge

We also investigated the interaction between prior knowledge and tutor type. We measured student prior knowledge from previous quizzes in class. The topics of quizzes were not related to confidence interval: they were theoretical probability distributions and statistical design. The trend indicated that students with high prior knowledge learned more with WE, and those with low knowledge learned more with TPS. However, these results were not significant.

Conclusions and Future Research

We report on two experiments and compared both student procedural and conceptual knowledge on tutored and “pure” WE conditions. We found that the pure WE was superior to TPS on conceptual knowledge, and also used

less time (i.e. was more efficient). On the other hand, TPS was superior at producing more learning on procedural knowledge but took more time. On procedural questions the efficiency ratings of TPS and WE were not reliably different from each other. Given that creating worked examples is so much more cost effective than creating intelligent tutoring system, these results are of practical importance. However, we hesitate to generalize these conclusions to other environments where the self-discipline to use WE might be lower. In fact in a separate paper submitted to this conference we report (Razzaq et al. submitted) that in an inner city middle school, albeit a slightly different implementation of a worked example condition, we found that TPS was more effective than the worked example condition. This suggests to us that self-discipline to actively engage with the material may be a factor in the effectiveness of worked examples. Our results support previous research indicating that worked examples are superior to cognitive tutors in regards to supporting conceptual learning. It has been suggested that worked examples yield improved learning as compared to simple problem solving due to reduced cognitive load on the student (Merrienboër & Sweller 2005). McClaren et al. (2008) posit that the *level of assistance* is important in predicting learning results and that research indicates “a tendency for *mid-level* assistance being most beneficial to learning.” (In their rubric, tutored problem solving and worked examples with no explanations are both counted as mid-level assistance.) More work is required to further explore why and under which conditions this kind of straightforward WE condition performs so well.

Appendix 1 – Statistical Models

We describe here the statistical models more in detail. We use two regression models for repeated binary data: the marginal regression model and the GLMM. While the former requires fewer model assumptions, the latter enables us to model multi-level effects and to predict individual performance. In the first experiment, we modeled p_{ij} , the i^{th} student’s probability of solving the j^{th} problem correctly, to be determined by the tutorial type ($k = 1$ for TPS, 2 for WE). That is,

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_k \quad (\text{M1})$$

for $i = 1, \dots, 95, j = 1, 2$. For the second experiment, we used the following three models. First, we used model M2a where the probability is determined by problem type ($m = 1$ for procedural, 2 for conceptual), tutorial type ($k = 1, 2$) and the test ($s = 1$ for pre-test, 2 for post-test). Two interesting parameters in this model are the learning difference between two tutorials in procedural problems, $\gamma_1 = (\beta_{1,2,2} - \beta_{1,1,1})$, and that in conceptual problems, $\gamma_2 = (\beta_{2,2,2} - \beta_{2,1,1})$.

$$\log \frac{p_{ijs}}{1 - p_{ijs}} = \beta_{m,k,s} \quad (\text{M2a})$$

We used model M2b to take into account the time each student spent on a tutorial (t_i). In this model, $\beta_{1,m,k}$ is the rate of learning per minute for each problem type ($m=1,2$), and each tutorial type ($k=1,2$).

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_{0,m,k} + \beta_{1,m,k} t_i \quad (\text{M2b})$$

Finally, we used a multilevel model (M2c) to estimate item easiness in the second experiment. In this model, we have Gaussian random effects corresponding to student knowledge on pre-test, b_{i1} , and on post-test, b_{i2} , and problem difficulty, b_j . This is an extension of the Rasch IRT model (Doran et al. 2007) to measure student performance on each test. We defined item easiness, $\beta_{m,t} - b_{jt}$, averaging easiness over two tutorial types.

$$\log \frac{p_{ijs}}{1 - p_{ijs}} = \beta_{m,s} + b_{is} - b_j \quad (\text{M2c})$$

References

- Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181–214.
- Atkinson, R.K., Renkl, A. & Merrill, M.M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, 95(4), 774–783.
- Bates, D. & Sarkar, D. (2007). lme4: Linear mixed-effects models using s4 classes. R package version 0.9975-12, URL <http://CRAN.R-project.org/>.
- Doran H., Bates D., Bliese P., Dowling M. (2007). Estimating the multilevel rasch model: with the lme4 package. *Journal of Statistical Software*. 20 (2). <http://www.jstatsoft.org/v20/i02>
- Embretson, S. & Reise, S (2002) *Item Response Theory for Psychologists*. LEA: Mahwah, NJ.
- Fitzmaurice G. M., Laird N. M. & Ware J. H., (2004). *Applied Longitudinal Analysis*, Wiley-IEEE.
- Heffernan, N. (2009) WPI Department of Computer Science Technical Report Number 2009-01. <http://www.cs.wpi.edu/Research/techreports.html>
- Garfield, J. (2002) the challenge of developing statistical reasoning, *Journal of Statistics Education*. 10 (3).
- Koedinger, K. R., Alevan, V., Heffernan, T., McLaren, B. & 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.
- McLaren, B.M., Lim, S., & Koedinger, K.R. (2008a). When is assistance helpful to learning? Results in combining worked examples and intelligent tutoring. In B. Woolf, E. Aimeur, R. Nkambou, S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, Lecture Notes in Computer Science, 5091 (pp. 677-680). Berlin: Springer.
- McLaren, B.M., Lim, S., & Koedinger, K.R. (2008b). 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
- Merrienboër, J. and Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17 (2) pp. 147-177.
- Razzaq, L., Heffernan, N. Shrestha, P., Wei, X., Maharjan, A. & Heffernan, C. (2009). Are worked examples an effective feedback mechanism during problem solving? *Proceedings of the Annual Meeting of the Cognitive Science Society* (to appear).
- Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. In L. Gleitman & A. K. Joshi (Eds.), *Proceeding of the 22nd Annual Conference of the Cognitive Science Society* (pp. 393–398). Mahwah, NJ: Erlbaum.
- Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: smooth transitions help learning. *Journal of Experimental Education*, 70, 293–315.
- Salden, R., Alevan, V., Renkl, A., & Schwonke, R. (2008). Worked examples and tutored problem solving: redundant or synergistic forms of support? In C. Schunn (Ed.) *Proceedings of the 30th Annual Meeting of the Cognitive Science Society, CogSci 2008* (pp. 659-664). New York, NY: Lawrence Earlbaum
- Schwonke, R, Wittwer, J., Alevan, V., Salden, R., Krieg, C., & Renkel, A. (2007). Can tutored problem solving benefit from faded worked-out examples? In S. Vosniadou, D. Kayser & A. Protopapas (Eds.), *Proceedings of EuroCogSci 07. The European Cognitive Science Conference 2007* (pp. 59 – 64). New York, NY: Erlbaum.
- Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Alevan, V., & Salden, R. (2009). The worked-example effect: is it just an artefact of lousy control conditions? *Computers in Human Behavior*. 25 (2) pp. 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. and Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction*, 7(1), 1-39.