

A Test of the Testing Effect: Acquiring Problem-Solving Skills From Worked Examples

Citation for published version (APA):

Van Gog, T., & Kester, L. (2012). A Test of the Testing Effect: Acquiring Problem-Solving Skills From Worked Examples. *Cognitive Science*, 36(8), 1532-1541. <https://doi.org/10.1111/cogs.12002>

DOI:

[10.1111/cogs.12002](https://doi.org/10.1111/cogs.12002)

Document status and date:

Published: 01/11/2012

Document Version:

Peer reviewed version

Document license:

CC BY-NC-SA

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

<https://www.ou.nl/taverne-agreement>

Take down policy

If you believe that this document breaches copyright please contact us at:

pure-support@ou.nl

providing details and we will investigate your claim.

Downloaded from <https://research.ou.nl/> on date: 30 Nov. 2022

Open Universiteit
www.ou.nl



Running head: A TEST OF THE TESTING EFFECT WITH WORKED EXAMPLES

A Test of the Testing Effect: Acquiring Problem-Solving Skills from Worked Examples

Tamara van Gog^a and Liesbeth Kester^b

^a Institute of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands

^b CELSTEC, Open University of The Netherlands, Heerlen, The Netherlands

Author Note:

Correspondence concerning this manuscript should be addressed to Tamara van Gog, Institute of Psychology, Erasmus University Rotterdam, The Netherlands, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. T: +31 1040889041; E: vangog@fsw.eur.nl

Acknowledgement. This research was funded by a Veni Grant from the Netherlands Organization for Scientific Research (NWO) awarded to Tamara van Gog (# 451-08-003). During the realization of this work, Liesbeth Kester was also supported by a Veni grant from NWO (# 451-07-007). The authors would like to thank Jeroen Mikkers for his assistance with this experiment.

Abstract

The ‘testing effect’ refers to the finding that after an initial study opportunity, testing is more effective for long-term retention than restudying. The testing effect seems robust and is a finding from the field of cognitive science that has important implications for education. However, it is unclear whether this effect also applies to the acquisition of problem-solving skills, which is important to establish given the key role problem solving plays in for instance math and science education. Worked examples are an effective and efficient way of acquiring problem-solving skills. Forty students either only studied worked examples (SSSS) or engaged in testing after studying an example by solving an isomorphic problem (STST). Surprisingly, results showed equal performance in both conditions on an immediate retention test after five minutes, but the SSSS condition outperformed the STST condition on a delayed retention test after one week. These findings suggest the testing effect might not apply to acquiring problem-solving skills from worked examples.

Keywords: testing effect; worked examples; problem solving

1. Introduction

The testing effect demonstrates that after an initial study opportunity, testing is more effective for long-term retention than restudying (for a review, see Roediger & Karpicke, 2006a). At an immediate retention test, there may be no performance differences between students in a condition that only studied and a condition that also engaged in testing (i.e., retrieving the studied information from memory), or those that only studied might even perform better. After a delay of one week, however, students in the testing condition outperform their counterparts that only studied (e.g., Roediger & Karpicke, 2006b).

The testing effect seems to be quite robust, as it has been demonstrated with a variety of learning tasks. Most research has used text materials such as word lists (e.g., Wheeler, Ewers, & Buonanno, 2003), facts (e.g., Carpenter, Pashler, Wixted, & Vul, 2008) or prose passages (e.g., Roediger & Karpicke, 2006b), but the effect has also been shown to apply to symbol-word pairs (Coppens, Verkoeijen, & Rikers, 2011), videotaped lectures (Butler & Roediger, 2007), visuo-spatial materials such as maps (Carpenter & Pashler, 2007), and multimedia materials such as animations (Johnson & Mayer, 2009). Given the clear implications of the testing effect for educational practice, it is not surprising that research has moved away from the cognitive psychology lab and into the classroom (see McDaniel, Roediger, & McDermott, 2007). What is surprising, however, is that the potential occurrence of a testing effect when acquiring problem-solving skills has –to the best of our knowledge– not yet been investigated, despite the key role that problems play in important areas of education such as math and science.

Regarding the acquisition of problem-solving skills, research has demonstrated that for novices, instruction consisting mainly of problem-solving practice is not very efficient compared to instruction that relies more heavily on worked example study (for reviews, see Atkinson, Derry, Renkl, & Wortham, 2000; Sweller, Van Merriënboer, & Paas, 1998; Van

Gog & Rummel, 2010). This is known as ‘the worked example effect’ (Sweller et al., 1998) and is explained by the fact that compared to problem solving, worked examples reduce the high cognitive load imposed by ineffective problem-solving strategies novices use (see Sweller, 1988), and instead allow them to devote their attention to building a cognitive schema of how to solve that type of problem (Sweller et al., 1998). Recent studies show that the beneficial effects of worked examples on the effectiveness and efficiency of novices’ learning are even found compared to tutored problem solving, which is a stronger control condition than conventional, untutored problem solving in the sense that it also provides instructional guidance to learners (Salden, Koedinger, Renkl, Alevan, & McLaren, 2010).

In research on the worked example effect, a heavier reliance on example study than on problem solving is usually implemented by means of example-problem pairs, in which a worked example is immediately followed by an isomorphic problem to solve (e.g., Carroll, 1994; Cooper & Sweller, 1987; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Mwangi & Sweller, 1998; Sweller & Cooper, 1985), although some studies have also used examples only (e.g., Van Gerven, Paas, Van Merriënboer, & Schmidt, 2002; Van Gog, Paas, & Van Merriënboer, 2006). Sweller & Cooper (1985) mention that engaging in solving a similar problem immediately after example study may be more motivating for students, because it requires students to be more actively engaged than studying another example would (and if learners are not motivated to study examples, they will not learn much from them). However, they did not test this assumption. Another potential benefit of example-problem pairs compared to examples only relevant to the present study, is that example-problem pairs provide an opportunity for retrieval practice (i.e., testing) that is not present when only studying worked examples. It could be expected that this retrieval practice opportunity provided by example-problem pairs would lead to better learning outcomes than studying examples only, as problem solving following example study can have several benefits. For

instance, it allows students to practice with retrieving information on the solution procedure used in the example (source problem) from memory in order to solve an isomorphic (target) problem by analogy (Holyoak, 2005; Renkl, 2011), which is also required on the retention test (i.e., transfer-appropriate processing). It may also provide students with feedback regarding the quality of the knowledge (problem schema) they have acquired by the ease or difficulty they experience during problem solving, and hence, allow them to determine to which aspect of a next example they should pay more attention.

In a recent study, however, Van Gog, Kester, and Paas (2011) compared the effects on learning of instruction consisting of examples only (i.e., repeated study) or example-problem pairs (i.e., testing) and found no differences in performance on an immediate retention test. This study did not include a delayed retention test though, and as mentioned above, research on the testing effect suggests that benefits of retrieval practice on test performance may not become evident until several days later. Therefore, the present study investigated the effects on immediate and delayed retention test performance of example study only and example-problem pairs. It is hypothesized that there will be no performance differences on the immediate retention test after five minutes, but that the example-problem pairs condition will outperform the examples only condition on the delayed retention test after one week.

2. Method

2.1. Participants and Design

Participants were 40 Dutch university students (4 male; age $M = 20.65$, $SD = 4.02$) who did not take science classes in the later years of secondary education. They were randomly assigned to one of the conditions: (1) study only: worked examples study followed by an immediate (5 min.) and delayed (one week) retention test (SSSS-T-T; $n = 20$) or (2) study-

test: example-problem pairs followed by an immediate (5 min.) and delayed (one week) retention test (STST-T-T; $n = 20$).

2.2. Materials

The materials, which were paper-based, focused on learning to solve electrical circuits troubleshooting problems. They were the same as the materials used by Van Gog et al. (2011), with the addition of an isomorphic test that was newly created.

2.2.1. Conceptual prior knowledge test

The conceptual prior knowledge test consisted of seven open-ended questions on troubleshooting and parallel circuits principles.

2.2.2. Formula sheet

On one page A4 paper, Ohm's law was explained and the different forms of the formula were given (i.e., $R=U/I$; $U=R*I$; $I=U/R$).

2.2.3. Acquisition phase examples and problems

The troubleshooting problems consisted of a malfunctioning parallel electrical circuit. In the circuit drawing, it was indicated how much voltage the power source delivered and how much resistance each resistor provided.

In the problem format, participants then had to answer the following questions: 'Determine how this circuit should function using Ohm's law, that is, determine what the current is that you should measure at each of the ammeters'; (this was given) 'Suppose the ammeters indicate the following measurements: ...'; 'What is the fault and in which component is it located?'. Based on the information in the circuit and the formula sheet, the current that *should* be measured (i.e., if the system were functioning correctly) in each of the parallel branches as well as overall could be calculated. By comparing the measurements given at step 2 to those calculated at step 1, it could be inferred in which branch the resistance differed from the resistance indicated in the diagram, and the actual measurement at step 2

could be used to find the actual value of the resistor. In the example format, participants did not have solve this problem themselves; the solutions were fully worked-out and students had to study the solution procedure.

In the first pair of tasks (i.e., either two examples or one example followed by one problem, depending on assigned condition), the fault was that *lower* current was measured in a particular parallel branch, which is indicative of *higher* resistance in that branch. In the second pair of tasks, the fault was that *higher* current was measured in a particular parallel branch, which is indicative of *lower* resistance in that branch.

The two tasks in each pair were isomorphic (i.e., a similar problem solving procedure was required, but surface features such as the values of resistors and voltage supplied by the power source in the circuit differed).

2.2.4. Retention tests

The immediate and delayed retention tests consisted of two troubleshooting tasks in problem format. While there was one familiar fault in the first test task (i.e., it was isomorphic to one pair of the training tasks), the second test task was slightly different: it contained two faults –though both faults were encountered in the training. Two isomorphic versions of the tests were developed (A and B), in which only surface features such as the values of resistors and voltage differed.

2.3. Procedure

Participants were randomly assigned to conditions. They first filled out demographic data, after which they completed the conceptual prior knowledge test. Then, they received the troubleshooting tasks associated with their assigned condition. These were provided in a booklet, with each task printed on a separate page. Participants were instructed to perform the tasks sequentially, and not to look back at previous tasks or look ahead to the next task. This was monitored by the experimenter and was done in order to prevent the example-problem

pairs condition from using the examples during problem solving. Participants were given 3 minutes per task (as in the study by Van Gog et al., 2011). Time was kept by the experimenter using a stopwatch, and the experimenter indicated when participants were allowed to proceed to the next task. After the acquisition phase tasks were completed, the experimenter collected the booklet with the training tasks. Participants were then given a filler task (5 min.) after which they were given the booklet with troubleshooting test tasks for the immediate retention test. There were two equivalent versions (A and B) of the test. Half of the participants in each condition received version A, the other half version B. One week later, participants returned for the delayed retention test, receiving version B when they had had version A on the immediate test, and version A when they had had version B on the immediate test. Participants were allowed to use a calculator (provided by the experimenter) and to use the formula sheet throughout all phases of the experiment.

2.4. Data Analysis

The maximum total score on the conceptual prior knowledge test was 10 points. For the troubleshooting retention test tasks with only one fault, the maximum score was 3 points: One point for correctly calculating the current at all ammeters, one point for correctly indicating the faulty component, and one point for indicating what the fault was (i.e., what the actual resistance was). For the troubleshooting retention test tasks containing two faults, the maximum score was 5 points (i.e., there were two faulty components and two faults to identify). The scoring procedure for the prior knowledge and retention tests was based on a model answer sheet that was also used in the study by Van Gog et al. (2011).

3. Results

Data from one participant in the SSSS-T-T was incomplete because of absence at the delayed test and is therefore not used in the analyses. For all analyses, a significance level of

.05 is used and Cohen's d is provided as a measure of effect size, with 0.2, 0.5, and 0.8 corresponding to small, medium, and large effect sizes, respectively (Cohen, 1988).

As expected, performance on the conceptual prior knowledge test was low ($M = 1.48$, $SD = 1.29$) and did not differ significantly between conditions $t(38) = 1.77$, *ns*.

Performance scores during the acquisition phase (STST-T-T condition) and on the tests (both conditions) are presented in Table 1 as percentages for reasons of comparison; the analyses on the test scores presented below were conducted on the sum scores (max. score = 8). In the STST-T-T condition, performance on the first problem in the learning phase did not correlate significantly with performance on either the immediate or delayed test, however, performance on the second problem in the learning phase did correlate significantly with the immediate ($r = .60$, $p < .01$) as well as delayed test ($r = .45$, $p < .05$).

-----insert Table 1 about here-----

In line with our hypothesis and the findings from Van Gog et al. (2011), there was no significant difference in performance on the immediate retention test (after 5 min.) between the SSSS-T-T ($M = 5.83$, $SD = 2.85$) and the STST-T-T ($M = 5.15$, $SD = 2.57$) conditions, $t(38) = .79$, *ns*. On the delayed retention test (after one week), however, performance did differ significantly between conditions, but not in the direction one would expect based on the testing effect: the SSSS-T-T condition ($M = 5.61$, $SD = 2.36$) outperformed the STST-T-T condition ($M = 4.10$, $SD = 2.20$), $t(37) = 2.06$, $p < .05$, $d = .66$.

4. Discussion

This study investigated if the testing effect also applies to the acquisition of problem solving skills from worked examples, and our findings suggest it does not: example study only was more effective after one week than example study alternated with testing (i.e., problem solving). The finding by Van Gog et al. (2011) that acquiring problem-solving skills

by means of example study or from example-problem pairs was equally effective as measured by performance on an immediate retention test, was replicated. According to the testing effect, however, the benefits of the multiple retrieval practice opportunities that are present in the example-problem pairs condition but not in the examples only condition, would manifest themselves only after a delay (Roediger & Karpicke, 2006a). To our surprise, however, this was not the case; it was even reversed: delayed retention test performance was lower in the example-problem pairs condition.

As mentioned in the introduction, the testing effect has been demonstrated with different types of learning materials, such as word lists (e.g., Wheeler et al., 2003), facts (Carpenter, et al., 2008), prose (Roediger & Karpicke, 2006b), lectures (Butler & Roediger, 2007), symbols (Coppens et al., 2011), maps (Carpenter & Pashler, 2007), and animations (Johnson & Mayer, 2009). An important difference between for instance word lists or facts, and worked examples, is that word lists or facts require literal retrieval of an item previously studied. This item needs to be retrieved from memory and written down (at least in the most widely used free recall test). However, testing following worked example study requires students to solve an isomorphic problem, which involves more than recall. The example can serve as a source analog for solving the target problem, by mapping the *solution procedure* of the example onto the target problem (Holyoak, 2005; Renkl, 2011). This implies that students' attention during example study needs to be focussed on the solution procedure in order to build a schema of that procedure; the exact values used in each step in the procedure are not important because the test problem contains different values. As a consequence, after mapping each step in the solution procedure from the example (source) to the test (target) problem, learners still need to execute that step themselves and make the required calculations. This consecutive 'answer construction' aspect is not present in the above mentioned materials. Construction seems to play a role when final test questions go beyond

free recall and require inferences to be made (e.g., Butler, 2010; Johnson & Mayer, 2009); however, this would primarily involve making novel combinations of information present in the mental representation of a text or animation, which is different from making calculations. Moreover, failing to make an inference on one question does not necessarily affect responses to other questions, whereas with problem solving, each step builds on the previous one.

It is important to note that even though construction of an answer is required at each step, this study is different from studies on the generation effect, an effect that is closely related to the testing effect, but different in nature (see Karpicke & Zaromb, 2010). Whereas we first provided students with a worked example, in which the solution steps were worked out and then tested their memory for the solution procedure by means of an isomorphic problem, studies on the generation effect that have looked at problem solving (e.g., multiplication; McNamara & Healy, 1995) do not first provide students with an example of a procedure. For instance, in the study by McNamara and Healy, learners were provided only with a problem statement and the answer ('read condition') or were provided only with the problem statement and had to generate the answer ('generation condition'). However, this meant they also had to generate the steps to take to attain that answer by themselves. In a subsequent test, identical problems were given, and it was found that having generated an answer led to better test performance (at least on more difficult problems) than only having read that answer. However, given that the test problems were identical, in this case, "solving" the test problem did not necessarily require renewed application of the procedure, it could also be based on literal recall of a previously constructed answer. Given that we investigated memory for a solution procedure after studying a worked example showing that procedure, rather than investigating effects of generating an answer on memory for that answer, we feel our study was an investigation of the testing effect rather than the generation effect.

In sum, even though the testing effect has been demonstrated with different types of tests (see Roediger & Karpicke, 2006a), and even with transfer questions that went beyond the information provided in the materials (Butler, 2010; Johnson & Mayer, 2009), the focus on recalling knowledge of the solution procedure and then applying that knowledge to execute each step in the procedure constitutes a fundamental difference between learning from worked examples and learning the other kinds of materials that were used thus far.

This difference might hold the key to the explanation of why we did not find a testing effect (and even the reverse) in the present study, and several potential explanations could be addressed in future studies. For instance, the fact that construction is required at each step might interrupt the recall of the solution procedure. An interesting question for future studies therefore, is whether asking students in the ‘testing’ condition to simply recall the example they studied during the acquisition phase would be more effective than asking them to solve an isomorphic problem. Another potential explanation is that the students in the examples conditions did better in the long term because they had studied more examples and therefore had more opportunities for self-explaining the examples, which leads to more knowledge elaboration and better understanding of the procedure (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Renkl, 1997, 2002), and in other domains it has been shown that effects of enhanced elaboration and understanding tend to show mainly at a delay (e.g., Mamede et al., in press; Van Blankenstein, Dolmans, Van der Vleuten, & Schmidt, 2011; Woods, Brooks, & Norman, 2007). Finally, our acquisition phase was relatively short, consisting of only four learning tasks in total. Even though one might argue that only studying examples provides more opportunities for schema acquisition, it seems unlikely that this would explain the difference between the conditions on the delayed test, as one would expect this to affect performance on the immediate test as well. However, neither in this study, nor in the Van Gog et al. (2011) study, was there a difference on the immediate test between participants in the

example study only group (who studied four examples) and the example-problem pairs condition (who studied only two examples alternated with practicing problem solving). Yet, it is possible that different results would be found with longer acquisition phases or different ratios of examples and problems (e.g., studying two examples and then solving a problem), which future studies might establish. When investigating whether a testing effect might occur in longer acquisition phases, it should be kept in mind though that worked examples are only effective for learning when students are novices. For students with more expertise, who have already acquired a problem schema, worked examples are no longer necessary and may even hamper learning (Kalyuga et al., 2001; this has become known as the “expertise reversal effect”, for a review, see Kalyuga, Ayres, Chandler, & Sweller, 2003).

A potential limitation of this study, which should also be addressed in future research, is the fact that all participants were given 3 min. per task in the learning phase. In testing effect studies, it is important to keep the acquisition phase time equal in both conditions to ensure that longer time on task would not be a potential cause of differences between conditions, and based on a prior study 3 min. seemed to be sufficient. Even though performance on the acquisition phase problems was lower than on the immediate test (Table 1) in the example-problem pairs condition, this unlikely to be due solely to time constraints, as it is also a matter of students still being in the process of skill acquisition. This seems to be supported by the finding that performance on the first learning phase problem did not correlate with test performance, whereas performance on the second learning phase problem did correlate highly with performance on both tests. Moreover, if the time constraint was a disadvantage, one would expect a performance difference between the conditions on the immediate test already, which we did not find. Nevertheless, we cannot entirely rule out that 3 min. was insufficient for some participants in the example-problem pairs condition to be able

to solve the problem. Hence, future studies might investigate the effects of providing students (in both conditions) with more time per task in the acquisition phase.

In sum, the findings from this study might potentially point towards a boundary of the testing effect, and identifying potential boundary conditions is not only of scientific interest to the cognitive science community, but also of practical relevance for the educational science community, as it can lead to more specific instructional guidelines regarding for which learning materials testing is or is not more useful than restudying. However, because this was –to the best of our knowledge- the first study to investigate the testing effect with worked examples, further research along the lines mentioned above is needed to corroborate these findings before we can conclude with certainty that the testing effect does not apply to the acquisition of problem-solving skills.

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*, 181-214.
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118-1133.
- Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514-527.
- Carpenter, S., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474-478.
- Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition, 36*, 438-448.
- Carroll, W. M. (1994). Using worked out examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*, 360-367.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347-362.
- Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 23*, 351-357.

- Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 117-142). Cambridge, UK: Cambridge University Press.
- Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621-629.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23-32.
- Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579-588.
- Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227-239.
- Mamede, S., Van Gog, T., Moura, A. S., De Farias, R. M. D., Peixoto, J. M., Rikers, R. M. J. P., & Schmidt, H. G. (in press). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical Education*.
- McDaniel, M.A., Roediger, H.L., & McDermott, K.B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206.
- McNamara, D. S., & Healy, A. F. (1995). A generation advantage for multiplication skill and non-word vocabulary acquisition. In A. F. Healy & L. E. Jr. Bourne (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 132-169). Thousand Oaks, CA: Sage Publications.
- Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction, 16*, 173-199.

- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1–29.
- Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*, 529–556.
- Renkl, A. (2011). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 272-295). New York: Routledge.
- Roediger, H.L., & Karpicke, J.D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210.
- Roediger, H.L., & Karpicke, J.D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.
- Salden, R., Koedinger, K. R., Renkl, A., Alevan, V., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review, 22*, 379-392.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
- 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.
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-295.
- Van Blankenstein, F. M., Dolmans, D. H. J. M., Van der Vleuten, C. P. M., & Schmidt, H. G. (2011). Which cognitive processes support learning during small-group discussion? The role of providing explanations and listening to others. *Instructional Science, 39*, 189-204.

- Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: Effects of worked examples on training efficiency. *Learning and Instruction, 12*, 87–105.
- 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*, 212–218.
- Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*, 154-164.
- Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174.
- Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571-580.
- Woods, N. N., Brooks, L. R., & Norman, G.R. (2007). It all make sense: Biomedical knowledge, causal connections and memory in the novice diagnostician. *Advances in Health Sciences Education, 12*, 405-415.

Table 1. *Mean (SD) of performance on the conceptual prior knowledge test, during the acquisition phase, and on the tests in %*

	Conceptual prior knowledge test	Acquisition Phase		Immediate Test	Delayed Test
		Problem 1	Problem 2		
SSSS-T-T	18.3 (12.6)	n.a.	n.a.	72.81 (35.60)	70.07 (29.51)
STST-T-T	11.3 (12.4)	19.17 (17.33)	45.00 (36.71)	64.37 (32.07)	51.25 (27.48)

April 21, 2012

Concerns: **Cognitive Science manuscript 11-1234R2**

Dear Editor, dear Dr. Moore,

Thank you for accepting our manuscript "A Test of the Testing Effect: Acquiring Problem-Solving Skills from Worked Examples" for publication in *Cognitive Science* as a brief report, pending some remaining modifications (Action Letter of January 4, 2012).

We have now made the required changes, detailed below in the response to the reviewers' comments. My apologies for taking so long to resubmit this revised manuscript; I had a small accident in January that nevertheless had a major impact on my ability to work for several weeks –as a consequence of which many tasks, including this one, have unfortunately experienced delays.

We have also added keywords as requested in the final submission guidelines, numbered headings as requested in the style guidelines, and uploaded as well as e-mailed the copyright transfer agreement.

With kind regards,

Tamara van Gog
vangog@fsw.eur.nl

Response to Reviewers:

Thank you very much for these additional comments and suggestions. Below we indicate the changes we made in the manuscript (plain text) in response to each of your comments and suggestions (italics).

Reviewer #1

Without doubt, this manuscript has improved by addressing the reviewers' comments. Nevertheless, there are still some open questions:

- Now the authors are more specific with respect to the processes that the learners in the STST condition presumably engage in. Although the assumptions on the problem solving processes are plausible at first glance, they are not really the processes that are identified by analogical reasoning research (retrieving the source, mapping, schema induction by performing mapping etc.; see the recent reviews of this research by Keith Holyoak). And I think analogical reasoning research applies very well to this situation where a problem has to be solved on the basis of a previous "source example". My recommendation is that the authors postulate the usual analogical reasoning processes (in any case). When they want additionally to present their own assumptions they should provide good reasons for that. Framing the problem solving process for "T" differently (i.e., in analogical reasoning terms) also affects the interpretation of the findings in the discussion. I guess what I am trying to say is that the interpretation of the findings should be connected to the assumed processes for "S/T".

We now rephrased some parts of the introduction and discussion section in terms of analogical reasoning (Holyoak, 2005; Renkl, 2011).

- The "lack of time problem" that was raised by reviewer 2 could also be addressed more "deeply". I suggest some post-hoc analyses whether the drop between the immediate and the delayed tests in the STST condition is related to problem solving performance in the acquisition phase. In addition, such analyses might also inform about possible reasons for the present findings.

Performance on both tests correlated highly and positively with performance on the **second** learning phase problem; we now added this to the results section: "In the STST-T-T condition, performance on the first problem in the learning phase did not correlate significantly with performance on either the immediate or delayed test, however, performance on the second problem in the learning phase did correlate significantly with the immediate ($r = .60$, $p < .01$) as well as delayed test ($r = .45$, $p < .05$)." We also refer to the absence of correlation with performance on the **first** learning phase problem in the discussion now (sentence printed in bold here was added): "Even though performance on the acquisition phase problems was lower than on the immediate test (Table 1) in the example-problem pairs condition, this unlikely to be due solely to time constraints, as it is also a matter of students still being in the process of skill acquisition. **This seems to be supported by the finding that performance on the first learning phase problem did not correlate with test performance, whereas performance on the second learning phase problem did correlate highly with performance on both tests.** Moreover, if the time constraint was a disadvantage, one would expect a performance difference between the conditions on the immediate test already, which we did not find."

- When looking at paper the second time I noticed that the t value for the prior knowledge scores is not too far away from the level of significance. So please provide the pretest scores (Ms and SDs) in the two groups, and check whether the results remain stable when including prior knowledge as covariate.

We added prior knowledge M & SD per condition to Table 1. When including conceptual prior knowledge as covariate the difference on the delayed test only nears significance ($p = .06$); however, even though it seems to have an effect statistically, the question is whether this difference is meaningful, as the pretest showed that conceptual knowledge was so low in both conditions ($M < 2$ out of 10) that it is not very likely to have played a meaningful role in differences in performance outcomes between conditions. Hence, we did not report the analysis with pretest score as covariate, but we would be happy to include it if you wish.

Reviewer #2

This review is for the revised version of the manuscript "A Test of the Testing Effect with Worked Examples". The authors did an excellent job in addressing the original concerns through substantial additional discussion, and carefully described the revisions in the "response to reviewers" document. I very much enjoyed reading the manuscript - thank you.

I have two very small suggestions. First, on page 10, the second last sentence in the first paragraph states "even when more complex materials than word lists are used and final test questions would go beyond free recall and would require, for instance, inferences to be made, these inferences can be expected to have been made at the time the information was studied". I did not understand the rationale behind the last part - why would one expect the inferences to be made when information was studied? Prior work (e.g., Chi et al., 1989) on the self-explanation effect, for instance, suggests that while some students do indeed make additional inferences when studying, this is not the case for all students?

We rephrased this, it now reads: "Construction seems to play a role when final test questions go beyond free recall and require inferences to be made; however, this would primarily involve making novel combinations of information present in the mental representation of a text or animation, which is different from making calculations. Moreover, failing to make an inference on one question does not necessarily affect responses to other questions, whereas with problem solving, each step builds on the previous one."

The second suggestion is purely editorial - on page 12, the middle paragraph starts with "Finally, our acquisition phrase was relatively short" and goes on to describe this potential limitation. The next paragraph, however, starts off by saying "A potential limitation of this study" - I would remove the word "finally" from the middle paragraph since it doesn't seem to include the last limitation discussed.

The "finally" paragraph was meant not so much as a potential limitation, but as a potential explanation of why we did not find a testing effect. We now attached it to that paragraph, which gets a little long as a consequence, but makes the structure of the arguments clearer.

The Cognitive Science Society, Inc.

Publication Agreement

The undersigned authors agree that the manuscript entitled "A Test of the Testing Effect: Acquiring Problem-Solving Skills from Worked Examples"

having been accepted for publication in the journal *Cognitive Science*, or *topiCS*, is now the property of the Cognitive Science Society, Inc. Specifically, the author(s) transfer(s) to the Society the exclusive rights comprised in the copyright of said manuscript, except that the author(s) retain(s) the following:

- 1) All proprietary rights, other than copyright.
- 2) The right to make oral presentations of the material.
- 3) The right to use, after publication, all or part of the material in any book authored or edited by the author(s).
- 4) Authors are entitled to distribute copies of the article for personal use, either on paper or electronically, through their own personal mailing or website, or through the website of an agency by which they are employed, but permission of the Cognitive Science Society is required to reproduce published papers in other sources, including electronic archives.

It is understood, in relation to Item 4, that the author(s) is/are required to obtain permission in writing from the Cognitive Science Society for all or a substantial part of the article to be reproduced for distribution by a commercial publisher or an agency in which the author(s) is/are not employed.

In the case of works prepared by employees or contractors of the U.S. Government, the U.S. Government will retain a non-exclusive, royalty-free license to use any or all of the material for any government purpose.

The author(s) warrant(s) that the manuscript is the original work of the author(s), contains no libelous or unlawful statements, and does not infringe any legal rights of any other party. In the case of an employee work for hire, both the author and a representative authorized to bind the owner of the work have signed below.

Cognitive Science and *topiCS* prefer all co-authors of jointly authored papers to sign the copyright transfer. In the event where only the principle co-author signs, Cognitive Science and *topiCS* will assume that the signing co-author does so with the authorization of the remaining co-author(s).

Intending to be legally bound, and understanding that the Society will grant specific license for republication of this manuscript without charge when agreed to by the authors and upon written request of the prospective licensee, the author(s) has/have subscribed below.

	21-04-2012	Tamara van Gog	Erasmus University Rotterdam
Author's Signature & Date		Printed Name	Institution or Organization

Please sign and upload this form to Editorial Manager with your final revision; or you may email it to the Editorial Office at cogscij@indiana.edu, or topicsj@indiana.edu