Educ Psychol Rev (2015) 27:265–289
DOI 10.1007/s10648-015-9297-3

REPLICATION

Testing After Worked Example Study Does Not Enhance Delayed Problem-Solving Performance Compared to Restudy

Tamara van Gog · Liesbeth Kester · Kim Dirkx · Vincent Hoogerheide · Joris Boerboom · Peter P. J. L. Verkoeijen

Published online: 24 February 2015

Abstract Four experiments investigated whether the testing effect also applies to the acquisition of problem-solving skills from worked examples. Experiment 1 (n=120) showed no beneficial effects of testing consisting of *isomorphic* problem solving or *example recall* on final test performance, which consisted of isomorphic problem solving, compared to continued study of isomorphic examples. Experiment 2 (n=124) showed no beneficial effects of testing consisting of *identical* problem solving compared to restudying an identical example. Interestingly, participants who took both an immediate and a delayed final test outperformed those taking only a delayed test. This finding suggested that testing might become beneficial for retention but only after a certain level of schema acquisition has taken place through restudying several examples. However, experiment 2 had no control condition restudying examples instead of taking the immediate test. Experiment 3 (n=129) included such a restudy condition, and there was no evidence that testing after studying four examples was more effective for final delayed test performance than restudying, regardless of whether restudied/tested problems were isomorphic or identical. Experiment 4 (n=75) used a similar design as experiment 3 (i.e., testing/restudy after four examples), but with examples on a different topic and with a different participant population. Again, no evidence of a testing effect was found. Thus, across four experiments, with different types of initial tests, different problem-solving domains, and different participant populations, we found no evidence that testing enhanced delayed test

Experiments 1, 2, and 3 by Van Gog, Kester, Dirkx, and Hoogerheide, experiment 4 by Van Gog, Boerboom, and Verkoeijen

T. van Gog (✉) · V. Hoogerheide · J. Boerboom · P. P. J. L. Verkoeijen
Institute of Psychology, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
e-mail: vangog@fsw.eur.nl

L. Kester · K. Dirkx
Welten Institute, Open University of The Netherlands, Heerlen, The Netherlands

L. Kester
Department of Pedagogical and Educational Sciences, Utrecht University, Utrecht, The Netherlands

K. Dirkx
Fontys Educatief Centrum, Fontys University of Applied Sciences, Eindhoven, The Netherlands

performance compared to restudy. These findings suggest that the testing effect might not apply to acquiring problem-solving skills from worked examples.

Keywords Testing effect · Worked examples · Problem solving

The testing effect refers to the finding that after an initial study opportunity, testing is more effective for long-term retention than not testing and even than restudying (for a review, see Roediger and Karpicke 2006a). This effect seems robust, as it has been demonstrated with a variety of learning materials, such as word lists (e.g., Wheeler et al. 2003), facts (e.g., Carpenter et al. 2008), prose passages (e.g., Roediger and Karpicke 2006b), symbol–word pairs (Coppens et al. 2011), videotaped lectures (Butler and Roediger 2007), visuospatial materials such as maps (Carpenter and Pashler 2007), numerical materials such as functions (Kang et al. 2011), and multimedia materials such as animations (Johnson and Mayer 2009). However, problem-solving tasks, which play a central role in school subjects such as math, chemistry, physics, or economics, have received very little attention in research on the testing effect. Given the relevance of the testing effect for education (see, e.g., Agarwal et al. 2012; Karpicke and Grimaldi 2012; McDaniel et al. 2007; Roediger and Karpicke 2006a), it is important to establish whether the testing effect would also apply to the acquisition of problem-solving skills from worked examples. However, there seems to be only one study to date that investigated whether testing after worked example study would enhance problem-solving performance on a delayed test (Van Gog and Kester 2012), and this study showed no evidence for a testing effect. The present study builds on those findings by investigating the effects of different practice test conditions in four experiments, with the aim of getting more insight in conditions under which the testing effect may or may not occur when learning to solve problems.

Learning from Worked Examples

Worked examples provide learners with a written demonstration of how to solve a problem or complete a task. Research has shown that for novices, who have little if any prior knowledge of a task, studying worked examples (often alternated with problem-solving practice) is more effective and more efficient for learning than mere problem-solving practice (i.e., the "worked example effect"; for reviews, see Atkinson et al. 2000; Renkl 2011, 2014; Sweller et al. 1998; Van Gog and Rummel 2010). That is, after example study, students perform better on retention (and even transfer) tests, often with less learning time and less investment of mental effort during the learning and/or test phase. Efficiency benefits (i.e., equal or better performance attained in less study time) are even found when example study alternated with (tutored) problem-solving practice is compared to tutored problem solving. This is a stronger control condition than conventional, untutored problem solving because it provides more instructional guidance (i.e., scaffolding, feedback; Koedinger and Aleven 2007) to learners. Efficiency benefits were found in studies comparing, for instance, example–tutored problem pairs vs. tutored problems only (McLaren et al. 2008), worked examples vs. tutored problems (McLaren et al. 2014), and faded examples with increasingly more steps for the learner to complete with tutor support vs. tutored problem solving only (Schwonke et al. 2009; for a review of examples in tutoring systems, see Salden et al. 2010).

While most studies on the worked example effect have contrasted problem-solving practice with example–problem pairs, in which a worked example is immediately followed by an isomorphic problem to solve (e.g., Carroll 1994; Cooper and Sweller 1987; Kalyuga et al. 2001; Sweller and Cooper 1985), other studies compared problem-solving practice to example

study only (e.g., Nievelstein et al. 2013; Van Gerven et al. 2002; Van Gog et al. 2006). Comparisons of the effects of examples only, example–problem pairs, problem–example pairs, and problem solving only showed that examples only and example–problem pairs were equally effective for immediate retention test performance and were both more effective than problem–example pairs and problem solving only (Leppink et al. 2014; Van Gog et al. 2011).

The lack of a performance difference between the examples-only and example–problem pairs conditions is intriguing from a testing effect perspective, because the examples-only condition can be seen as a repeated-study condition whereas the example–problem pairs condition can be seen as a study-testing condition; that is, example–problem pairs provide an opportunity for retrieval practice that is not present when only studying worked examples. It should be noted, though, that beneficial effects of testing over restudy tend to become apparent primarily on a delayed test; on an immediate test, there may be no performance differences or restudy might even be more effective (Roediger and Karpicke 2006a).

The Testing Effect in Learning from Worked Examples

To investigate whether a testing effect would become evident at a delayed test, Van Gog and Kester (2012) compared the effects of examples only and example–problem pairs on both an immediate and a delayed test. They found no significant difference between these two conditions on the immediate test 5 min. after the study phase, in line with the findings by Van Gog et al. (2011). On the delayed test after 1 week, surprisingly, the examples-only condition outperformed the example–problem pairs condition.

While this seemed to suggest that the testing effect might not apply to the acquisition of problem-solving skills from worked examples, it would be too early to draw that conclusion. For instance, in many studies on the testing effect, final test moment (immediate vs. delayed) is a between-subjects (e.g., Coppens et al. 2011) or between-items (e.g., Carpenter et al. 2008) factor instead of a repeated measures within-subjects factor as in the Van Gog and Kester (2012) study. This cannot explain why no testing effect was found (because all participants took an immediate as well as a delayed test), but it cannot be ruled out that it was this repeated final test that caused the higher performance on the delayed test in the examples-only condition. Therefore, it is important to repeat the Van Gog and Kester study with final test moment as between-subjects factor.

Moreover, there is an important difference between learning materials previously used in studying the testing effect and problem solving. Materials such as word lists (e.g., Wheeler et al. 2003), facts (e.g., Carpenter et al. 2008), or symbol–word pairs (Coppens et al. 2011) are low-element interactivity materials that often require literal retrieval during testing. Problem-solving tasks, in contrast, are often high in element interactivity, and tests of problem-solving skill require not only recall but also generation/(re)construction. That is, the most relevant type of test to determine whether students have really learned a problem solution procedure is to have them solve *isomorphic* problems. Such problems have the same structural features and require the use of the same solution procedure as the problems demonstrated in the worked examples, but they have different surface features (e.g., different cover stories or different values). When solving isomorphic problems, the worked examples that were studied could 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, 2014). However, the exact numbers or values used in each step of the procedure in the worked example are not important because the test problem contains different values. As a consequence, what students need to recall during testing is the solution procedure (i.e., the steps to be taken), but they still have to execute each of the steps themselves (i.e., make the required calculations).

This consecutive "answer construction" aspect may interrupt the recall process, which might perhaps explain why Van Gog and Kester (2012) did not find a testing effect. Therefore, it is important to investigate whether a testing effect would occur when other testing conditions are being used in the initial and final tests, such as free recall of the information provided in the worked example, or solving a problem fully identical to the one demonstrated in the worked example that was just studied, which allows for heavier reliance on recall than isomorphic problem solving. To shed light on the conditions under which a testing effect might or might not occur when acquiring problem-solving skills from worked examples, we conducted four experiments.

Experiment 1 replicated the study by Van Gog and Kester, using isomorphic problems and examples, with the difference that participants took either an immediate or a delayed test, instead of taking both tests, and next to a practice test consisting of solving an *isomorphic* problem, it included an additional practice test condition: *example recall*. Experiment 2 focussed on yet another practice test condition, by investigating the effects of restudying or solving an *identical* example or problem, respectively. Performance on a practice problem after having studied only one example was relatively low in experiments 1 and 2, and because the testing effect may depend on how well students performed on the practice test (i.e., testing can only boost performance if there is something to boost in the first place), experiment 3 investigated the effects of testing after having studied multiple examples. All participants first studied four (isomorphic) examples before restudying two more examples (i.e., restudy) or solving two practice problems (i.e., practice test). The two restudy examples or practice problems were either isomorphic or identical to two examples from the first set of four. In experiments 1, 2, and 3, we used the same materials and had university students as participants. To exclude the possibility that the results were somehow restricted to the materials or the participant population, experiment 4 replicated the isomorphic conditions from experiment 3 in a different population (vocational education) and with different materials.

Experiment 1

Experiment 1 replicated the study by Van Gog and Kester (2012) with students either studying four examples (EEEE) or two example–problem pairs (EPEP), but with two important differences: 1) test moment was implemented as between-subjects factor, so that participants in each condition received *either* an immediate or a delayed test, and 2) another testing condition was added: example recall (ERER), which did not require problem solving. This allowed us to investigate whether testing consisting of example recall would be more effective for delayed test performance than testing consisting of isomorphic problem solving (i.e., example–problem pairs) or no testing (i.e., studying examples only).

In line with the findings by Van Gog et al. (2011) and Van Gog and Kester (2012), we did not expect to find differences on an immediate test between the examples-only and the example–problem pairs condition. On the delayed test, the examples-only condition might outperform the example–problem pairs condition in line with the findings by Van Gog and Kester (2012), unless the repeated testing (i.e., immediate+delayed) in that study was responsible for the higher delayed test performance in the worked-examples condition, in which case no differences between conditions would be expected in the present study. Finally, if the interruption of the recall process due to answer construction in isomorphic problem solving would explain why Van Gog and Kester did not find a testing effect, then a testing effect would be expected to arise with the example–recall condition. That is, this condition would be expected to outperform the examples-only condition and the example–problem pairs condition

on the delayed test. Because testing can be assumed to be more effortful than restudying (Bjork 1994), potential differences in invested mental effort across conditions were explored.

Method

Participants and Design

Participants were 120 students (age M=20.89, SD=2.77; 67 female, 53 male) enrolled in various programs at a Dutch university. They were randomly assigned to one of six conditions (n=20 per condition) resulting from a 3×2 factorial design with between-subjects factors instruction condition (examples only, EEEE; example–problem pairs, EPEP; example–recall pairs, ERER) and final test moment (immediate, 5 min; delayed, 1 week).

Materials

The materials, which were paper based, focused on learning to solve electrical circuits troubleshooting problems. With the exception of the recall task used in the ERER condition, they were the same as the materials used by Van Gog et al. (2011) and Van Gog and Kester (2012).

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

Introduction and Formula Sheet On one page of A4-sized paper, the abbreviations of the components of the circuit drawing were explained, along with a description of Ohm's law and the different forms of the formula (i.e., $R=U/I$; $U=R*I$; $I=U/R$).

Acquisition Phase Tasks 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 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 to solve this problem themselves; the solutions were fully worked out and students had to study the solution procedure. An example of a worked example is provided in the Appendix.

In the EEEE condition, examples 1 and 2 contained the same fault (i.e., *lower* current was measured in a particular parallel branch, which is indicative of *higher* resistance in that branch) and examples 3 and 4 contained the same fault (i.e., *higher* current was measured in a particular parallel branch, which is indicative of *lower* resistance in that branch). Examples 1 and 2, as well as examples 3 and 4, were isomorphic, so they contained the same structural features (i.e., same solution procedure, same fault) but different surface features (i.e., different circuit drawing, different values of resistors, different voltage supplied by the power source). The EPEP condition

received the same tasks, with the difference that only tasks 1 and 3 were in example format while tasks 2 and 4 were in problem format for the students to solve. The ERER condition received examples 1 and 3, and—during the time in which the other conditions worked on tasks 2 and 4—was asked to recall and write down as much as possible from the example just studied.

Final Test The final retention test 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 in two different branches, both of which had been encountered in the training.

Mental Effort Rating Scale Invested mental effort was measured using the nine-point subjective rating scale developed by Paas (1992), ranging from (1) very, very low effort to (9) very, very high effort.

Procedure

The study was run in small group sessions, with maximally eight participants per session, seated in individual cubicles. Participants 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 participants in the EPEP and the ERER conditions from using the examples during problem solving or recall. Participants were given 4 min per task. The experimenter kept time with a stopwatch and indicated when participants were allowed to proceed to the next task. After the acquisition phase when the tasks were completed, the experimenter collected the booklets. Participants were then given a filler task for 5 min (a Sudoku puzzle) after which participants in the immediate final test condition received the retention test, while participants in the delayed final test condition received another filler task (a different Sudoku puzzle). One week later, all participants returned to the lab, and those in the delayed final test condition received the retention test, while participants in the immediate final test condition received a filler task. Immediately after each task in the acquisition and test phase, participants rated the amount of mental effort they invested in studying the example/solving the problem/recalling the example. Participants were allowed to use a calculator (if they did not have one, they were provided with one by the experimenter) and to use the formula sheet throughout the acquisition and final test phase; they were provided with a new formula sheet in each phase, and the experimenter checked that they did not make notes on the sheet during the acquisition phase.

Data Analysis

The maximum total score on the conceptual prior knowledge test was ten points. For the final test task with only one fault, the maximum score was three 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). Half points were given for partially correct but incomplete answers. So for instance, for the step of calculating the current at all ammeters (step 1 of the problem in the Appendix), one point would be given if the answer contained "5 mA+10 mA+50 mA=65 m," and a half point was

given if a participant only filled out the formula, i.e., "5/1000+5/500+5/100." For the step of correctly indicating the faulty component, one point was given if the answer indicated that "the current at AM2 is lower than it should be and that therefore R_2 has a higher resistance indicated" or "R_2 has a higher resistance then indicated," and a half point was given when a participant only said that "the current at AM2 was lower than it should be." For the step of calculating how high the resistance actually is, one point was given for "5/7.14=0.7 kΩ or 700Ω" or "5/7.14," and a half point was given if they only gave the formula "$R=U/I$".

For the final test task containing two faults, the maximum score was five points (i.e., there were two faulty components and two faults to identify). The test scores were summed and then converted to percentages for ease of interpretation and comparison with experiment 2. The scoring procedure for the prior knowledge and retention test was based on a model answer sheet that was also used in the studies by Van Gog et al. (2011) and Van Gog and Kester (2012), and since this scoring model was very straightforward and did not leave much room for interpretation (as the example above indicates), scoring was done by a single rater.

Results

Data are provided in Table 1. Performance on the initial test problems in the EPEP condition was M=34.58 % (SD=27.83) for problem 1 and M=66.25 % (SD=28.11) for problem 2. One participant in the ERER-delayed condition did not show up for the delayed test session, leaving 19 participants in this condition. One participant in the EPEP condition failed to fill out the mental effort rating on one acquisition phase task; this was replaced with the average of her other three ratings. As for mental effort ratings on the test, next to the missing data from the one participant who failed to show up, one participant in the EPEP-delayed condition failed to fill out mental effort on one of the test tasks and was therefore excluded from this analysis.

Unexpectedly, an ANOVA showed a marginally significant difference across the six conditions in performance on the conceptual prior knowledge test, $F(5,114)=2.28$, $p=0.051$, $\eta_p^2=0.09$. Bonferroni post hoc tests showed that performance in the EEEE-immediate condition was significantly lower than performance in the ERER-delayed condition ($p=0.017$). No other differences were significant. Note though, that in general, prior knowledge was very low, and that this would only be problematic if a testing effect would be found for the ERER-delayed condition.

Final Test Performance A 3×2 ANOVA on final test performance, with factors instruction condition (EEEE, EPEP, ERER) and final test moment (immediate, delayed) showed no effect

Table 1 Mean (SD) performance and mental effort scores in experiment 1

	EEEE		EPEP		ERER	
	Immediate	Delayed	Immediate	Delayed	Immediate	Delayed
Prior knowledge (0–10)	0.63 (0.87)	1.33 (1.02)	1.53 (1.63)	1.45 (1.36)	1.45 (1.77)	2.10 (1.52)
Final test performance (%)	75.31 (26.40)	49.69 (35.67)	81.25 (22.76)	50.94 (28.12)	72.50 (29.34)	57.57 (34.46)
Effort acquisition phase (1–9)	4.43 (1.76)	4.53 (1.91)	4.93 (1.88)	5.54 (1.91)	4.79 (1.85)	5.01 (1.64)
Effort final test (1–9)	4.68 (1.66)	5.60 (1.38)	3.88 (1.86)	5.11 (2.46)	4.40 (2.47)	5.11 (2.11)

of instruction condition, $F<1$, $p=0.858$, $\eta_p^2=0.003$, but did show a significant main effect of final test moment, with performance in the delayed final test conditions ($M=52.65$ %, SD=32.50) being significantly lower than performance in the immediate final test conditions ($M=76.35$ %, SD=26.11), $F(1,113)=18.75$, $p<0.001$, $\eta_p^2=0.142$. There was no significant interaction, $F<1$, $p=0.503$, $\eta_p^2=0.012$.

Mental Effort An ANOVA with instruction condition as factor showed that while the mean mental effort invested during the acquisition phase seemed lower in the EEEE condition than in the other conditions (Table 1), this difference was not significant, $F(2,116)=1.79$, $p=0.172$, $\eta_p^2=0.030$.

A 3×2 ANOVA on effort invested in the final test with factors instruction condition (EEEE, EPEP, ERER) and final test moment (immediate, delayed) showed no effect of instruction condition, $F(1,112)=1.02$, $p=0.364$, $\eta_p^2=0.018$, but did show a significant main effect of final test moment, with effort invested in the delayed final test conditions ($M=5.28$, SD=2.00) being significantly higher than effort invested in the immediate final test conditions ($M=4.32$, SD=2.02), $F(1,112)=6.52$, $p=0.012$, $\eta_p^2=0.055$. There was no significant interaction, $F<1$, $p=0.848$, $\eta_p^2=0.003$.

Discussion

This experiment did not show any evidence of a testing effect, either with isomorphic problem solving or with example recall, when acquiring problem-solving skills from worked examples. There was no significant difference across instruction conditions on either immediate or delayed final test performance. With regard to examples only and example–problem pairs, the findings on the immediate test replicated the findings by Van Gog et al. (2011) and Van Gog and Kester (2012).

However, in the present experiment, we did not find a benefit of examples only on the delayed test as in the study by Van Gog and Kester (2012). Possibly, this is due to the fact that test moment was a within-subjects factor in their study, so that taking the immediate test may have affected performance on the delayed test in the example study only condition, but not in the example–problem pair condition. In the present experiment, students either took the immediate or the delayed test, in which case there no longer was a benefit of examples only. Another possible explanation is that the examples-only condition also received isomorphic tasks during acquisition (i.e., the problems to be solved in the EPEP condition, but in worked-out format). Thus, rather than restudying the exact same example, participants studied an isomorphic example, and some studies have shown that variability in examples might foster learning (e.g., Paas and Van Merriënboer 1994), although other studies did not find beneficial effects of variability (e.g., Renkl et al. 1998).

As such, results might have been different if a "pure" restudy condition had been used in which participants would receive the exact same example they had just studied for restudy. Moreover, testing consisting of completing a problem fully identical to the one presented in the example might be more effective than testing consisting of isomorphic problem solving or example recall because it allows for heavier reliance on recall than isomorphic problem solving, while providing more cues than free example recall.

Therefore, experiment 2 compared restudying identical examples in the EEEE condition (i.e., E1=E2 and E3=E4) with solving an identical problem in the EPEP condition. In addition, participants received a final test which consisted either of identical problems to those encountered in the learning phase or of isomorphic problems, and they received this final test either at a delay only, or both immediately and delayed.

Experiment 2

If the fact that the restudy conditions in prior experiments were not "pure" restudy conditions (i.e., isomorphic instead of identical examples), then a testing effect should arise in this experiment, in which the examples-only condition involves true restudy and the example–problem pairs condition requires identical instead of isomorphic problem solving. That is, the example–problem pairs condition would outperform the examples-only condition on the delayed test, at least when this test also consists of identical problems, but not necessarily when it consists of isomorphic problems. The effects of engaging in both an immediate and a delayed test on delayed test performance are explored because the findings of Van Gog and Kester (2012) suggested that possibly, taking a test does have a positive effect on longer-term retention, but only after a certain level of schema acquisition has taken place. Finally, because testing can be assumed to be more effortful than restudying, potential differences among conditions in invested mental effort were again explored.

Method

Participants and Design Participants were 124 Belgian (Dutch-speaking) university students from a Department of Pedagogical and Educational Sciences (age M=22.85, SD=3.88; 20 male, 104 female). Participants were randomly assigned to one of the eight conditions (n=15–16 per condition) resulting from a 2×2×2 design with between-subjects factors instruction condition (EEEE vs. EPEP), final test moment (delayed vs. immediate+delayed), and final test format (isomorphic vs. identical). Note that even though the same abbreviations (EEEE and EPEP) are used, there was an important difference between the conditions used here and those used in experiment 1 and in the study by Van Gog and Kester (2012), which both used isomorphic examples and practice test problems. In experiment 2, in contrast, the second example was identical to the first, and the fourth example was identical to the third. The same applied to the problems to be solved: the problem in each pair was identical to the one explained in the preceding example.

Materials The same materials were used as in experiment 1, with the exception that only acquisition phase tasks 1 and 3 were used (either in example or in example–problem format), and with the exception that two different final tests were used in this experiment: the isomorphic retention test was the same as used in experiment 1; the identical retention test consisted simply of the two acquisition phase tasks in problem format.

Procedure The procedure was identical to the procedure of experiment 1, with the exception that experiment 2 was run in three group sessions with 36 to 51 participants per session and that participants in the immediate+delayed final test condition received a test instead of a filler task during the second session.

Data Analysis The same scoring method for the isomorphic final test format was used as in experiment 1, resulting in a maximum score of eight points. On the identical final test format, a maximum score of six points could be gained because each task contained only one fault. Scores were converted to percentages for reasons of comparability. A model answer sheet was used and as in experiment 1, scoring was done by a single rater.

Results

Data are provided in Table 2. Performance on the initial test problems across the EPEP condition was M=47.54 % (SD=31.30) for problem 1 and M=82.24 % (SD=24.88) for problem 2. One participant in the EPEP-isomorphic-delayed condition did not show up for the delayed test session, leaving 14 participants in this condition. One participant in the EPEP condition failed to fill out the mental effort rating on two acquisition phase tasks, and these were replaced with the average of her other two ratings. On the immediate test, two participants in the EEEE-identical condition, two participants in the EEEE-isomorphic condition, and one participant in the EPEP-isomorphic condition failed to fill out one of the two mental effort ratings on the delayed test and were therefore excluded from this analysis. Next to the one participant who did not show up for the delayed test, four other participants' mental effort data were excluded from the analysis because they failed to fill out one of the two mental effort ratings on the delayed test (two in the EEEE-identical-delayed condition, one in the EEEE-isomorphic-delayed condition, and one in the EPEP-isomorphic-delayed condition).

An ANOVA showed no significant difference across the eight conditions in performance on the conceptual prior knowledge test, $F(7,116)=1.09$, $p=0.374$, $\eta_p^2=0.062$.

Final Test Performance A 2×2 ANOVA with factors instruction condition (EEEE vs. EPEP) and final test format (identical vs. isomorphic) on performance on the immediate final test showed no significant main or interaction effects (instruction condition: $F(1,56)<1$, $p=0.498$, $\eta_p^2=0.008$; final test format: $F(1,56)=2.60$, $p=0.113$, $\eta_p^2=0.044$; interaction: $F(1,56)<1$, $p=0.807$, $\eta_p^2=0.001$).

A 2×2×2 ANOVA on performance on the delayed final test, with factors instruction condition (EEEE vs. EPEP), final test format (identical vs. isomorphic), and final test moment (delayed vs. immediate+delayed), showed no main effect of instruction, $F(1,115)=1.35$, $p=0.247$, $\eta_p^2=0.012$ and a significant main effect of final test moment, $F(1,115)=14.13$, $p<0.001$, $\eta_p^2=0.109$, with participants in the immediate+delayed final test condition (M=63.96 %, SD=30.59) outperforming participants in the delayed final test condition (M=43.88 %, SD=29.16). There was no main effect of final test format, $F(1,115)=3.15$, $p=0.079$, $\eta_p^2=0.027$. None of the interaction effects were significant (all $F<1$, all $\eta_p^2<0.01$).

Mental Effort As in experiment 1, the average mental effort invested in the acquisition phase seems to suggest that participants in the EEEE condition (M=4.29, SD=1.88) invested less effort than participants in the EPEP condition (M=4.86, SD=1.57), but this difference was not significant, $F(1,122)=3.25$, $p=0.074$, $\eta_p^2=0.026$.

A 2×2 ANOVA on mental effort invested in the immediate final test, with factors instruction condition (EEEE vs. EPEP) and final test format (identical vs. isomorphic), showed no significant main or interaction effects (instruction condition: $F(1,51)<1$, $p=0.926$, $\eta_p^2<0.001$; final test format: $F(1,51)=2.66$, $p=0.109$, $\eta_p^2=0.050$; interaction: $F(1,51)=1.17$, $p=0.284$, $\eta_p^2=0.023$).

A 2×2×2 ANOVA on mental effort invested in the delayed final test, with factors instruction condition (EEEE vs. EPEP), final test format (identical vs. isomorphic), and final test moment (delayed vs. immediate+delayed), only showed a significant main effect of final test moment, $F(1,111)=14.35$, $p<0.001$, $\eta_p^2=0.114$, with participants in the immediate+delayed final test condition investing less effort in the delayed test (M=4.45, SD=2.05) than participants in the delayed final test condition (M=5.95, SD=2.24). No other main or interaction effects were significant (instruction: $F(1,111)=1.69$, $p=0.197$, $\eta_p^2=0.015$; final test format: $F(1,111)<1$, $p=6.39$, $\eta_p^2=0.002$; all two-way interactions: $F<1$, $\eta_p^2<0.01$; three-way interaction: $F(1,111)=1.177$, $p=0.280$, $\eta_p^2=0.010$).

Table 2 Mean (SD) performance and mental effort scores in experiment 2

	Immediate+delayed test				Delayed test			
	EEEE		EPEP		EEEE		EPEP	
	Isomorph	Identical	Isomorph	Identical	Isomorph	Identical	Isomorph	Identical
Prior knowledge (0–10)	1.90 (1.27)	2.23 (1.62)	2.77 (1.28)	2.80 (1.86)	2.28 (1.31)	1.94 (1.57)	2.07 (1.37)	2.84 (1.51)
Effort acquisition phase (1–9)	4.12 (1.94)	3.75 (1.65)	4.98 (1.44)	5.12 (1.69)	4.31 (1.98)	4.91 (1.91)	4.58 (1.75)	4.75 (1.51)
Performance immediate test (%)	62.50 (38.59)	76.67 (25.63)	69.58 (28.33)	80.00 (23.31)				
Performance delayed test (%)	57.50 (38.68)	63.89 (29.83)	63.33 (28.33)	71.11 (25.56)	38.67 (25.02)	42.65 (30.03)	36.61 (28.68)	56.77 (30.91)
Effort immediate test (1–9)	4.96 (2.17)	3.50 (2.12)	4.43 (1.47)	4.13 (2.13)				
Effort delayed test (1–9)	4.60 (2.56)	4.50 (1.91)	3.83 (1.44)	4.87 (2.21)	6.27 (2.38)	6.47 (1.93)	5.73 (2.59)	5.34 (2.10)

Discussion

This second experiment again showed no evidence of a testing effect. Even though in this experiment we applied yet another testing condition consisting of identical problem solving, used a restudy condition more comparable to other testing effect research (i.e., consisting of identical example study), and applied two different final test conditions consisting of either identical or isomorphic problem solving, there was no evidence that testing had a beneficial effect on delayed test performance compared to restudy.

Interestingly, however, experiment 2 showed that participants who had taken an immediate test after four acquisition tasks (either four examples or two example–problem pairs) had higher delayed test performance than participants who only took a delayed test. Moreover, participants who had also taken an immediate test reached this higher level of delayed test performance while investing less mental effort in completing this delayed test. These findings suggest that a testing effect might possibly occur when acquiring problem-solving skills from worked examples, but only once has a relatively comprehensive schema been acquired that allows for effective problem solving on the test. The acquisition phase performance data from the example–problem pairs conditions in experiments 1 and 2 indeed suggests that participants were still in the process of acquiring a schema of how to solve these problems (experiment 1—isomorphic: problem 1: M=34.58 %, SD=27.83, problem 2: M=66.25 %, SD=28.11; experiment 2—identical: problem 1: M=47.54 %, SD=31.30, problem 2: M=82.24 %, SD=24.88).

In other words, maybe testing during the acquisition phase was not effective for boosting longer-term retention because students were not yet able to successfully solve the problems (even when they were identical to the example they just studied). On the immediate test in experiment 2, after studying four examples or two example–problem pairs, they were better able to solve the problems, and this might have contributed to the superior longer-term retention.

However, in experiment 2, the delayed test only condition did not restudy examples while the other condition took the immediate test; in other words, there was no adequate control condition. Therefore, experiment 3 investigated whether taking a test consisting of either identical or isomorphic problems after having studied four examples (i.e., EEEE-PP) would be more beneficial for delayed test performance than continued study of isomorphic examples or restudy of identical examples (i.e., EEEE-EE).

Experiment 3

Experiment 3 compared the effects of testing after having studied four worked examples to continued example study (restudy). If the fact that the condition that took an immediate as well as delayed test outperformed the delayed test only condition in experiment 2 is due to a testing effect, then students in a condition taking a test after studying four examples should perform better on the delayed test than students engaging in restudy in experiment 3. If, on the other hand, the superiority of the condition that took both tests was only due to increased time spent on the learning tasks, then one would not expect to find a difference between the testing and restudy conditions in experiment 3.

In this experiment, we also varied whether testing tasks and restudy examples consisted of new, isomorphic (worked-out) problems or identical problems that are exactly the same as in the examples that were studied, to explore again whether this would make a difference in final test performance; the final test after 1 week consisted of isomorphic problem solving. As in experiments 1 and 2, because testing can be assumed to be more effortful than restudying, potential differences in invested mental effort were explored.

Method

Participants and Design Participants were 129 Belgian (Dutch-speaking) university students from a Department of Pedagogical and Educational Sciences (age M=18.79, SD=1.20; 8 male, 121 female). They were from the same university department as participants in experiment 2, but this was a new cohort of students who had not participated in the prior experiment. Participants were randomly assigned to one of the four instruction conditions resulting from a 2×2 design with factors testing/restudy and isomorphic/identical tasks: 1) immediate test with isomorphic problems (EEEE-$P_{iso}P_{iso}$; n=29); 2) immediate test with identical problems (EEEE-$P_{id1}P_{id3}$; n=33); 3) continued example study with isomorphic examples (EEEE-$E_{iso}E_{iso}$; n=34); and 4) continued example study with identical examples (EEEE-$E_{id1}E_{id3}$; n=33).

Materials The first four examples were the same as in experiment 1. The test problems in the immediate test conditions were either new, isomorphic problems (these differed from the examples as well as from the final delayed retention test) or identical problems that consisted simply of two examples (the first and third) in problem format. The examples in the continued study/restudy conditions were either new, isomorphic ones (the same problems that the isomorphic immediate test condition received, but with a fully worked-out solution) or identical ones (i.e., the first and third learning phase examples). The final delayed retention test (isomorphic) was the same as in experiments 1 and 2.

Procedure The experiment was run in four group sessions in two consecutive weeks, with maximally 36 participants per session. The procedure was identical to the procedure of experiments 1 and 2, with the exception of the immediate test phase: while half of the participants took a test (consisting of solving two identical or isomorphic problems depending on their assigned condition), the other half continued to study two examples (either new ones isomorphic to those of the learning phase or identical ones they had already seen in the learning phase). After 1 week, all participants completed the delayed test.

Data Analysis For scoring the final delayed test, the same method was used as in experiment 1 (and 2, in the isomorphic final test condition), resulting in a maximum score of eight points. On the immediate test, a maximum score of six points could be gained in both the isomorphic and the identical problems conditions because each task contained only one fault. Again, a model answer sheet was used and as in experiment 1, scoring was done by a single rater. Scores were converted to percentages.

Results

Data are presented in Table 3. Average performance on the practice test problems in the EEEE-PP isomorphic condition was M=77.00 % (SD=24.33 %), and in the EEEE-PP identical condition, it was: M=53.87 % (SD=31.13). Three participants were not present at the delayed test and were therefore excluded from all analyses (one from the isomorphic testing condition, one in the identical testing condition, and one in the identical restudy condition).

Unexpectedly, an ANOVA showed a significant difference across the four conditions in performance on the conceptual prior knowledge test, $F(3,122)=2.92$, $p=0.037$, $\eta_p^2=0.067$. Bonferroni post hoc tests showed that performance in the isomorphic testing condition was significantly higher than performance in the isomorphic restudy condition ($p=0.036$). No other differences were significant. Note though, that in general, prior knowledge was very low, and

Table 3 Mean (SD) performance and mental effort scores in experiment 3

	Testing (PP)		Restudy (EE)	
	Isomorphic	Identical	Isomorphic	Identical
Prior knowledge (0–10)	2.27 (1.15)	1.58 (1.38)	1.42 (1.12)	1.65 (1.02)
Final test performance (%)	68.30 (30.28)	52.93 (32.92)	59.66 (30.02)	51.70 (31.22)
Effort testing/restudy (1–9)	4.60 (1.46)	6.10 (2.16)	3.05 (1.65)	2.83 (1.55)
Effort final test (1–9)	4.52 (1.84)	5.28 (2.06)	5.67 (1.63)	5.69 (1.36)

that this would only be problematic if a testing effect would be found for the isomorphic restudy condition.

Final Test Performance A 2×2 ANOVA on performance on the final delayed test showed no main effect of testing/restudy, $F(1,122)<1$, $p=0.377$, $\eta_p^2=0.006$. There was a significant main effect of task format, $F(1,122)=4.40$, $p=0.038$, $\eta_p^2=0.035$, indicating that participants who tested/restudied isomorphic tasks performed better ($M=63.63$ %, SD=30.20) on the delayed test than participants who tested/restudied identical tasks ($M=52.31$ %, SD=31.82). There was no significant interaction, $F(1,122)<1$, $p=0.506$, $\eta_p^2=0.004$.

Mental Effort In the testing/restudy phase, 11 participants in the testing conditions failed to fill out one or both of the mental effort ratings and were not taken into account in the analysis. There was a significant main effect of testing/restudy, with the testing conditions investing more mental effort ($M=5.34$, SD=1.97) than the restudy conditions ($M=2.94$, SD=1.59) on the two tasks, $F(1,111)=56.50$, $p<0.001$, $\eta_p^2=0.337$. There was also a main effect of task format, with isomorphic tasks ($M=3.72$, SD=1.74) requiring less effort than identical tasks ($M=4.21$, SD=2.44), $F(1,111)=4.05$, $p=0.047$, $\eta_p^2=0.035$, but this was qualified by an interaction effect, $F(1,111)=7.15$, $p=0.009$, $\eta_p^2=0.061$, with follow-up t tests showing that there was no significant difference between the restudy conditions in invested mental effort, $t(64)=0.539$, $p=0.592$, Cohen's $d=0.137$ but there was a significant difference between the testing conditions, with the isomorphic testing condition investing less effort than the identical testing condition $t(47)=2.871$, $p=0.006$, Cohen's $d=0.813$.

On the final test, two participants in the identical restudy condition failed to fill out one of the mental effort ratings and were not taken into account in the analysis. There was a significant main effect of testing/restudy, $F(1,120)=6.21$, $p=0.014$, $\eta_p^2=0.049$, indicating that participants in the testing conditions ($M=4.92$, SD=1.98) invested less effort than participants in the restudy conditions ($M=5.68$, SD=1.49). There was no main effect of task format, $F(1,120)=1.59$, $p=0.209$, $\eta_p^2=0.013$, nor an interaction effect $F(1,120)=1.38$, $p=0.242$, $\eta_p^2=0.011$.

Discussion

Again, we did not find evidence of a testing effect. The finding that in the testing/restudy phase, testing was more effortful than restudying examples is not surprising given prior findings from research on the worked example effect (see, e.g., Paas 1992; Van Gog et al. 2006), and this finding even holds despite the fact that the interaction showed that participants in the isomorphic testing condition invested less mental effort than those in the identical testing

condition (they still invested more effort than the restudy conditions). This suggests that in terms of instructional efficiency, (re)studying examples is more efficient than testing, in the sense that the same level of final test performance was attained with less investment of mental effort during acquisition (Hoffman and Schraw 2010; Van Gog and Paas 2008).

It should be noted that caution is needed in interpreting some of the findings from this study; for instance, regarding the interaction on invested mental effort in the testing/restudy phase, it is likely that the fact that participants in the isomorphic testing condition had to invest less effort in testing is due to their somewhat higher prior knowledge instead of the format of the tasks. The same applies to the final, delayed test, where we found that the testing conditions invested less effort; however, when looking at the means in Table 3, it seems likely that this difference was caused mainly by the lower effort invested by the participants in the isomorphic testing condition.

Possibly, the somewhat higher prior knowledge of those participants is also what caused the difference in performance between participants who had tested/restudied with isomorphic tasks and participants who had tested/restudied with identical tasks on the delayed test, although the mean test score of the isomorphic restudy condition also seemed higher than the mean score of the identical restudy condition, so possibly, a variability effect is at play here (cf. Paas and Van Merriënboer, 1994). It is important to note that the somewhat higher prior knowledge of the isomorphic testing condition cannot explain the absence of a testing effect, as this would have worked in favor of finding such an effect.

In sum, across three experiments, using the same materials as Van Gog and Kester (2012) but different testing conditions, we found no evidence of a testing effect. Although it is highly unlikely that this lack of testing effects is related to the materials (which have been used, and showed significant effects compared to other instructional conditions in other studies, e.g., Van Gog et al. 2011), it cannot be entirely ruled out. Moreover, next to using the same materials, all three experiments were conducted with university students. Therefore, we conducted a fourth experiment, comparing the effects of testing after studying four examples (i.e., EEEE-EE vs. EEEE-PP) on delayed test performance in the domain of probability calculation with vocational education students.

Experiment 4

This experiment conceptually replicated the testing and restudy conditions with isomorphic problems from experiment 3, but with different materials and a different participant population.

Method

Participants and Design

Participants were 75 vocational education students (M_{age}= 17.46, SD=1.32; 12 male, 53 female) enrolled in a teaching assistant study program at a Dutch institute for vocational education. All participants were in their first year and they did not have prior knowledge of probability calculation, as this was not part of their curriculum.

Participants were randomly assigned to either the restudy or testing condition. We had to exclude ten participants because they were absent during the delayed test. Also, there were eight students who were present the second week, but not the first week, and their data was compared to performance of the participants who were present the week before, as a check that the latter had indeed learned from the example study during the acquisition phase 1 week

earlier. As a consequence, data from 57 participants was available for the main analysis (restudy: n=27; testing: n=30). Participants were rewarded with a small present and a certificate acknowledging their participation.

Materials

The materials were designed in close collaboration with participants' math teacher.

Acquisition Phase Tasks The first four examples demonstrated how to solve probability calculation problems with two sequential events. The order was important in all the examples (e.g., first drawing a yellow ball, then a blue ball out of a vase), and two examples were with replacement (e.g., the yellow ball going back into the vase before drawing a second time), the other two without replacement (e.g., the yellow ball staying out of the vase). The examples were presented as a video on a projection screen. The video consisted of an animated demonstration of the solution procedure for each of the four examples, explained by a female voice-over, and had a duration of 5 min and 44 s. The examples had different cover stories (i.e., drawing balls out of a vase, choosing soft drinks, choosing candy bars, and throwing dice), and students' attention was explicitly drawn to the difference between examples with and without replacement.

After the video, the restudy condition received two further worked examples to study, isomorphic to the ones in the video; one was with and one without replacement. The examples were presented on two separate pages of A4 paper, with one example per page. The test condition also received two separate pages of A4, but these contained two problems to solve, on the first page with, on the second page without replacement.

Final Test The final test consisted of eight problems, presented on separate pages. The first four problems were isomorphic to the ones in the acquisition phase; the last four were isomorphic in terms of the procedure but were made slightly more difficult by using either higher numbers in the calculations or a more complex cover story.

Mental Effort Rating Scale The same mental effort rating scale was used as in experiments 1, 2, and 3, and it was applied after each restudy/test problem and after each problem on the final test.

Procedure

The study was conducted with four existing classes of students, and within each class, students were randomly assigned to either the restudy or the testing condition. The first session started with a short introduction by the experimenter about the general procedures during the session (e.g., that students would not be allowed to ask questions once the procedure started). Then, the video with the four examples was shown on a projection screen at the front of the room that was visible for all students. Participants then received a booklet with either worked examples or test problems depending on their assigned condition; they were verbally instructed that they were not allowed to turn pages unless the experimenter said so, and they were given a written reminder of this on each page. Participants were given 5 min per page (containing either one example or two test problems), timed by the experimenter, who indicated that they could proceed to the next page after time was up. After each restudy or test task, participants rated how much effort they invested in studying the example or solving the problem.

During the second session exactly 1 week later, another booklet was handed out, containing the eight final test problems (one per page). Again, participants were verbally instructed that they were not allowed to turn pages unless the experimenter said so, and they were given a written reminder of this on each page. Participants were given 3 min per page, timed by the experimenter, who indicated that they could proceed to the next page after time was up. After each problem, participants rated how much effort they invested in solving it. Students were allowed to use calculators during both the initial and final tests.

Data Analysis

Performance on the final test problems was scored by assigning one point if the correct solution was provided, and therefore, the final test score ranged from 0 to 8. Five participants failed to fill out one of the mental effort ratings on the final test, so their average effort investment was computed based on their responses on seven problems.

Results

Average performance on the practice test problems in the testing condition was M=69.12 % (SD=39.44).

Manipulation Check Eight students completed only the final test but not the acquisition phase because they were absent during the first session but present during the second session. Therefore, their data was compared to performance of the participants who were present the week before, as a check that they had indeed learned from example study during the acquisition phase 1 week earlier. A Mann–Whitney U test showed that the eight students who were absent during the acquisition phase scored significantly lower on the final test (Mdn= 1) than the 57 students who engaged in both the acquisition phase and the final test (Mdn=3), U= 334.00, p= 0.031, r=0.267.

Final Test Performance There was no significant difference between the restudy (M=38.43 %, SD=32.32) and test (M=50.00 %, SD=35.51) condition in final test performance, $t(55)$= 1.282, p=0.205, Cohen's d=0.341.

Mental Effort There were no significant differences between conditions in mental effort invested during restudy (M=2.85, SD=2.06) or initial testing (M=3.08, SD=1.99), $t(55)$= 0.47, p=0.640, Cohen's d=0.11, and on mental effort invested in the final test (restudy: M= 3.47, SD=2.12; testing: M=3.20, SD=1.97), $t(55)$=0.51, p=0.611, Cohen's d=0.13.

Discussion

Experiment 4 was conducted with a different kind of problem-solving task and very different kind of participant population compared to experiments 1, 2, and 3, but again, we found no evidence of a testing effect. In contrast to experiment 3, we did not find that after having studied four examples, continuing to study required less effort than taking an initial test in this experiment. In general, students reported rather low effort investment on the test tasks compared to the other experiments, which may be related to differences in the student populations (i.e., vocational education is a lower track than university education, which is

the highest, and there is considerable difference in effort exertion of students at different tracks according to teacher reports, Carbonaro 2005).

Small-Scale Meta-analysis

The testing effect refers to the final test advantage of testing over restudying after a delay. The present study included eight delayed final test performance comparisons between restudy (in the form of repeatedly studying examples) and testing in the form of example–problem sequences (i.e., ignoring the recall condition in experiment 1). Figure 1 presents the 95 % confidence interval (CI) for each of these comparisons (green color when viewed online; this figure was generated using Cumming's 2012, ESCI software: www.thenewstatistics.com). The point estimates, denoted by the (green) squares, indicate the difference between the mean final test performance (expressed in percentage correct) between testing and restudying. Hence, a positive point estimate corresponds with a numerical testing effect. As can be seen in Fig. 1, seven comparisons yielded a numerical testing effect, whereas one comparison showed a numerical reversed testing effect. However, the CIs are very wide, which implies that each of the comparisons provides a rather imprecise estimate of the testing effect parameter. In order to get an indication of the combined effect across comparisons, we combined them in a random-effects small-scale meta-analysis, using Cumming's (2012) ESCI software (www.thenewstatistics.com). This resulted in a combined estimate based on 180 (restudy) and 173 (testing) participants. The 95 % CI of the combined effect is presented in red in Fig. 1. This combined effect shows that the mean performance (in terms of percentage points) was somewhat higher in the testing condition than in the restudy condition, mean testing–restudy difference=5.836, 95 % CI [−0.619; 12.292], d=0.19, but this difference was not statistically significant (p=0.076) in a two-tailed test. Also, the combined CI is much narrower than each of the CIs of the individual comparisons. Consequently, the combined estimate provides a more accurate estimate of the magnitude of the testing effect in worked examples than the single comparisons alone. Lastly, the heterogeneity index Q had a value of 2.393 with a p value of 0.935, suggesting there is no reason to assume that the samples in the eight comparisons were drawn from populations with different testing-effect parameter values. However, the heterogeneity index should be interpreted with caution because it is based on a small number of studies.

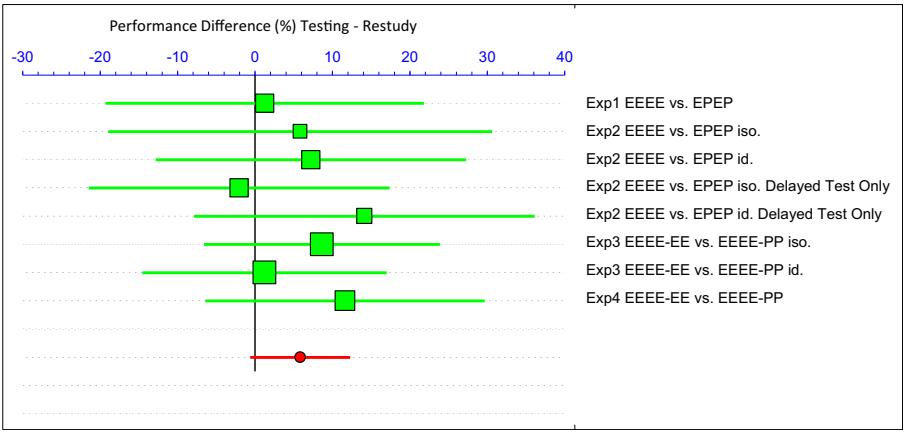


Fig. 1 Results of the meta-analysis

General Discussion

This study presented a total of four experiments comparing the effects of restudying vs. testing after studying worked examples, using different types of initial tests, different problem-solving domains, and different participant populations across the experiments. In none of these experiments, nor in an overall analysis, did we find evidence of a testing effect. In experiments 1 and 2, we investigated the effects of several different types of testing after studying one example, such as solving an isomorphic problem, recalling the worked example that was studied, or solving the exact same problem as in the worked example that was studied, and none of these showed a benefit over restudying an isomorphic or identical example in terms of immediate (after 5 min) or delayed (after 1 week) final test performance. Experiment 2 did show that participants taking an immediate test after four tasks had higher delayed test performance than participants who did not take an immediate test. However, when a restudy control condition was included (in experiments 3 and 4), in which participants either engaged in testing after having studied four examples or continued studying examples, no benefits of testing were found on delayed final test performance after 1 week. This suggests that the findings from experiment 2 were due to longer engagement with the learning tasks rather than to having engaged in testing.

With regard to research on example-based learning, the lack of difference between engaging in problem solving after example study or continuing example study is very interesting. Many early studies on the worked example effect used example–problem pairs because these were assumed to be more motivating for students (see Sweller and Cooper 1985), and with a few exceptions (e.g., Nievelstein et al. 2013; Van Gerven et al. 2002; Van Gog et al. 2006), this somehow became common practice. It would be worthwhile to investigate this motivational component in future research, because seen solely from the perspective of learning outcomes, the results from this study as well as others (Leppink et al. 2014; Van Gog et al. 2011) suggest that there is no need to alternate examples and problems.

The mental effort data are also interesting in this respect. In experiments 1 and 2, mental effort investment—although it seemed somewhat lower in the restudy (examples only) condition—did not differ significantly between the restudy and testing (example–problem pair) conditions (a finding that replicates effort data from these conditions in Van Gog et al. 2011). So after having studied one example, participants spend only a little (and not significantly) more effort on solving a practice problem than on studying another (experiment 1) or the same (experiment 2) example. However, in experiment 3, when participants had to engage in restudy or testing after having studied *four* examples, less effort was invested in restudy, compared to testing as well as compared to experiments 1 and 2 it seems (see Tables 1, 2, and 3). Although this should be interpreted with caution as it was not replicated with vocational education students in experiment 4, it may suggest that initially, students are still motivated to invest effort in worked example study, as they realize they have not yet learned the procedure, whereas once they become more familiar with the procedure, continuing to study examples becomes less motivating than engaging in problem solving. Again, with regard to learning outcomes, however, the higher effort invested in practice problem solving in experiment 3 did not translate into higher final test performance.

Indeed, our findings suggest that when students are still in the process of acquiring a cognitive schema of a solution procedure, they seem to gain very little from engaging in problem-solving practice in terms of learning outcomes. A recent study even found such a lack of learning gains when problem-solving practice was *additional* to example study (i.e., instead of partially replacing it): students who received three example–problem pairs did not outperform students who only received the three examples on a test (Baars et al. 2014a). What that

study did show, though, is that the practice problems helped students to make more accurate judgments of learning: they overestimated their future test performance less than students who made these judgments only on the basis of studying examples. In other words, although practice (i.e., testing) may not add much in terms of schema acquisition, it does not seem to hinder either, and it might have beneficial metacognitive or motivational effects for students, which could affect their persistence and therefore their learning outcomes in real educational settings.

Although performance on the isomorphic practice test problems was substantially higher in experiment 3 than in experiment 1, we still found no indication of a testing effect. Nevertheless, it might be interesting for future research to explore what would happen if practice problem solving performance would be increased further, for instance, by means of feedback on a practice test, restudy after practice testing, or successive relearning. In research on effects of taking practice tests, it has been shown, for instance, that practice tests can sometimes have an adverse effect on memory in the sense that incorrect answers selected on a multiple choice practice test may persist and be given as answers on a final cued recall test (e.g., Roediger and Marsh 2005), unless feedback is provided after a multiple-choice practice test (e.g., Butler and Roediger 2008). In another study, feedback was a prerequisite for a testing effect to even occur (i.e., the results regarding the short answer practice test by Kang et al. 2007).

Providing effective feedback on practice problems that goes beyond merely stating that a step was solved (in)correctly would essentially amount to giving students a worked example to study (i.e., the solution to each step) of the problem they just practiced. That is, it seems to resemble restudy after testing, which has been shown to increase the testing effect in several studies (see Rawson and Dunlosky 2012). There are indications from recent worked examples research that feedback/restudy might have beneficial effects on learning (Baars et al. 2014b). Baars et al. had students work with example–problem pairs and provided some of them with standards (i.e., correct answer feedback for each step) to which they could compare their practice problem performance. This not only helped them to make more accurate self-assessments of their performance but it also improved their learning outcomes compared to not having such standards available.

Note though that there was no restudy only comparison condition included in the study by Baars et al. (2014b). It is therefore an open question whether feedback or restudy after testing would be able to boost performance of the testing condition compared to continued restudying of examples. A recent study suggests that feedback after problem solving in the form of worked examples might only increase learning time but not learning outcomes compared to studying examples only (McLaren et al. 2014). Perhaps, however, very different results would be obtained when students first engage in example study before they solve practice problems with feedback or restudy opportunities, which future research should establish (e.g., by comparing example–practice problem–feedback example or practice problem vs. example–example–example).

The acquisition phases in the present study were relatively short. When investigating effects of testing compared to restudy in longer acquisition phases, it is possible that completing practice test problems may become more effective than restudy. However, it should be kept in mind that if a benefit of testing (by means of intermediate problem solving) would be found compared to restudy in longer sequences of tasks, the question would be whether this would be due to the beneficial effects of testing or to the negative effects of high levels of instructional guidance at a point where learners no longer need it. That is, it is known that example study loses its effectiveness for learning once a certain level of knowledge is reached (Kalyuga et al. 2001; see also Kalyuga 2007; Kalyuga et al. 2003). One possibility in longer acquisition phases might be to move focus from studying effects of testing at the whole task level to effects of testing at part-task level. Asking learners to complete solution steps (Paas 1992) or increasingly more solution steps (i.e., a fading strategy; see Renkl and Atkinson 2003) could

be regarded as a form of testing at a part-task level and is known to be an effective strategy for acquiring problem-solving skills from worked examples (Renkl 2014), although effects of fading on delayed retention (after one or more days/weeks) compared to example study only or example–problem pairs have not yet been systematically investigated.

If we forget about studying "the testing effect" as such for a moment and change our focus towards finding out what optimal sequences of examples and practice problems are for long-term durable learning, then another interesting avenue for worked examples research would be to investigate effects of initial learning to criterion and successive relearning opportunities (see Rawson and Dunlosky 2012). Although learning to criterion may be substantially harder for complex materials than for key terms or word pairs, especially given that problem solving does not depend on literal recall and learners should ideally be able to solve isomorphic problems, there may be benefits for long-term retention of taking practice tests until a learner completes a problem correctly at least once. Successive relearning opportunities adaptive to students' level of knowledge (i.e., having learners solve problems after several days and allowing them to restudy worked examples of the problems they could not yet complete) may be especially helpful for strengthening the third and fourth phase in example-based learning, "declarative rule formation" and "automation and flexibilization" (Renkl 2014).

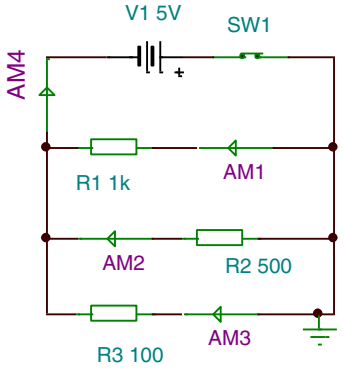
A potential limitation of our study, in particular of experiment 2, is that the sample sizes in some of the studies may not have given sufficient power to detect medium-to-small effects. However, although the power of the individual studies may have been low, this is definitely not the case for the small-scale meta-analysis, which was based on data from 180 (restudy) and 173 (testing) participants across the experiments. With this sample size, we had enough power to detect a medium-to-small effect. Hence, a lack of power argument does not apply to the combined findings from the small-scale meta-analysis in our study. Furthermore, the effect sizes associated with the testing–restudy difference in each of the individual studies as well as in the small-scale meta-analysis appear to be considerably smaller than the effects typically reported in the literature. So, even if the effect had been statistically significant in the meta-analysis (which it was not), that would not change the central conclusion of this study, namely that the testing effect—if existent at all—is small at best for materials with high element interactivity and much smaller than for materials with low element interactivity.

To conclude, the four experiments from this study along with the study by Van Gog and Kester (2012) and Leahy et al. (2015) suggest that the testing effect may not apply to the acquisition of complex problem-solving skills from worked examples. The testing effect has mainly been demonstrated with low element interactivity materials, and these findings along with others in this special issue suggest that it may not necessarily apply to learning more complex, high element interactivity materials (Van Gog and Sweller, this issue). However, most meaningful educational materials, including many problem-solving tasks, are high in element interactivity. As such, the finding that testing does not seem to foster longer-term retention of complex, high element interactivity materials is not only theoretically relevant, because it sheds light on a potential boundary condition of the testing effect, but also of practical interest, given the relevance of the testing effect for education (see, e.g., Agarwal et al. 2012; Karpicke and Grimaldi 2012; McDaniel et al. 2007; Roediger and Karpicke 2006a). For educators, it is very relevant to know for which types of learning tasks the testing effect does and does not apply. That being said, as stated above, solving practice problems after worked example study does not seem to hurt learning either, and future research might perhaps uncover metacognitive or motivational benefits of engaging in practice problem solving.

Acknowledgments Experiments 1 and 2 were funded by a Veni grant (# 451-08-003) from the Netherlands Organization for Scientific Research (NWO) awarded to Tamara Van Gog. During the realization of experiments 3 and 4, she was supported by a NWO Vidi grant (# 452-11-006). The authors would like to thank Jan Elen and Nick Stassens of KU Leuven, Belgium, for facilitating experiments 2 and 3, Dianne Kaho of Albeda College Rotterdam for facilitating experiment 4, and Pauline Reijners and Myrthe Knetemann for assistance with data collection and scoring for experiment 3.

Appendix

An example of a worked example (translated from Dutch)



1. Using Ohm's law, calculate how the circuit depicted above *should* function, that is, what you should measure at ammeters AM1 to AM4. AM1 to AM3 indicate the current in the parallel branches; AM4 indicates the total current.

   In a parallel circuit, the total current (I_t) equals the sum of the currents in the parallel branches (I_1, I_2, etc.)

   The total current should be $I_t=I_1+I_2+I_3$ (what is measured at AM4=AM1+AM2+AM3) or

$$I_t = \frac{U}{R_1} + \frac{U}{R_2} + \frac{U}{R_3} = \frac{5\,\mathrm{V}}{1\,\mathrm{k}\Omega} + \frac{5\,\mathrm{V}}{500\,\Omega} + \frac{5\,\mathrm{V}}{100\,\Omega}$$
$$= 5\,\mathrm{mA} + 10\,\mathrm{mA} + 50\,\mathrm{mA} = 65\,\mathrm{mA}$$

   Thus, if the circuit would function correctly, you should measure:

AM1=5 mA	AM2=10 mA	AM3=50 mA	AM4=65 mA

2. Suppose you would measure the following:

AM1=5 mA	AM2=7.14 mA	AM3=50 mA	AM4=62.14 mA

   Your calculation and the measurement do not match, so there is a fault in the circuit.

3. What is the fault and in which component is it?

   If the *current* in a branch is *lower* than it should be, the *resistance* in that branch is *higher* (because equal voltage U divided by a larger resistance R means a lower current I).

The current in the second branch is lower than it should be if the circuit were functioning correctly: I_2=7.14 mA. Thus, R_2 has a higher resistance than the indicated 500Ω. How high that resistance is can be calculated using the current that was measured in I_2:

$$R_2 = \frac{U}{I_2} = \frac{5\,\mathrm{V}}{7.14\,\mathrm{mA}} = 0.7\,\mathrm{k}\Omega = 700\,\Omega$$

References

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: retrieval practice improves learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24*, 437–448. doi:10.1007/s10648-012-9210-2.

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. doi:10.3102/00346543070002181.

Baars, M., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2014a). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*, 382–391. doi:10.1002/acp.3008.

Baars, M., Vink, S., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2014b). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. doi:10.1016/j.learninstruc.2014.04.004.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: knowing about knowing* (pp. 185–205). Cambridge: MIT Press.

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. doi:10.1080/09541440701326097.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616. doi:10.3758/MC.36.3.604.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education, 78*, 27–49. doi:10.1177/003804070507800102.

Carpenter, S., & Pashler, H. (2007). Testing beyond words: using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474–478. doi:10.3758/BF03194092.

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. doi:10.3758/MC.36.2.438.

Carroll, W. M. (1994). Using worked out examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*, 360–367. doi:10.1037/0022-0663.86.3.360.

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. doi:10.1037/0022-0663.79.4.347.

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. doi:10.1080/20445911.2011.507188.

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: applications in learning and problem-solving. *Educational Psychologist, 45*, 1–14. doi:10.1080/00461520903213618.

Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 117–142). Cambridge: Cambridge University Press.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629. doi:10.1037/a0015183.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. doi:10.1007/s10648-007-9054-3.

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. doi:10.1037/0022-0663.93.3.579.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31. doi:10.1207/S15326985EP3801_4.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19*, 528–558. doi:10.1080/09541440601056620.

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin and Review, 18*, 998–1005. doi:10.3758/s13423-011-0113-x.

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: a perspective for enhancing meaningful learning. *Educational Psychology Review, 24*, 401–418. doi:10.1007/s10648-012-9202-2.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*, 239–264. doi:10.1007/s10648-007-9049-0.

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*. doi:10.1007/s10648-015-9296-4.

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., & Van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32–42. doi:10.1016/j.learninstruc.2013.12.001.

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. doi:10.3758/BF03194052.

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: Cognitive Science Society.

McLaren, B., Van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2014). Exploring the assistance dilemma: comparing instructional support in examples and problems. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Lecture notes in Computer Science, 8474: intelligent tutoring systems* (pp. 354–361). Berlin: Springer. doi:10.1007/978-3-319-07221-0_44.

Nievelstein, F., Van Gog, T., Van Dijck, G., & Boshuizen, H. P. A. (2013). The worked example and expertise reversal effect in less structured tasks: learning to reason about legal cases. *Contemporary Educational Psychology, 38*, 118–125. doi:10.1016/j.cedpsych.2012.12.004.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach. *Journal of Educational Psychology, 84*, 429–434. doi:10.1037/0022-0663.84.4.429.

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: a cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133. doi:10.1037/0022-0663.86.1.122.

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24*, 419–435. doi:10.1007/s10648-012-9203-1.

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.

Renkl, A. (2014). Towards an instructionally-oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. doi:10.1111/cogs.12086.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: a cognitive load perspective. *Educational Psychologist, 38*, 15–22. doi:10.1207/S15326985EP3801_3.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*, 90–108. doi:10.1006/ceps.1997.0959.

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. doi:10.1111/j.1745-6916.2006.00012.x.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. doi:10.1111/j.1467-9280.2006.01693.x.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1155–1159. doi:10.1037/0278-7393.31.5.1155.

Salden, R., Koedinger, K. R., Renkl, A., Aleven, V., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review, 22*, 379–392. doi:10.1007/s10648-010-9143-6.

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. doi:10.1016/j.chb.2008.12.011.

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. doi:10.1207/s1532690xci0201_3.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–295. doi:10.1023/A:1022193728205.

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. doi:10.1016/S0959-4752(01)00017-2.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1532–1541. doi:10.1111/cogs.12002.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. doi:10.1080/00461520701756248.

Van Gog, T., & Rummel, N. (2010). Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. doi:10.1007/s10648-010-9134-7.

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. doi:10.1016/j.learninstruc.2006.02.003.

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. doi:10.1016/j.cedpsych.2010.10.004.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580. doi:10.1080/09658210244000414.