The Testing Effect for Learning Principles and Procedures from Texts

Running head: TESTING-EFFECT FOR PRINCIPLES AND PROCEDURES

THIS ARTICLE HAS BEEN PUBLISHED IN THE JOURNAL OF EDUCATIONAL RESEARCH UNDER http://dx.doi.org/10.1080/00220671.2013.823370

The Testing-Effect for Learning Principles and Procedures from Texts

Kim. J. H. Dirkx, Liesbeth Kester, and Paul. A. Kirschner

CELSTEC / Netherlands Laboratory for Lifelong Learning, Open University of the Netherlands, Heerlen, the Netherlands

Author Note

Correspondence concerning this article should be addressed to Kim J. H. Dirkx, Open Universiteit, CELSTEC, NL - P.O. Box 2960, 6401 DL, Heerlen, the Netherlands. E-mail: kim.dirkx@ou.nl.

The authors would like to extend a word of thanks to the teachers and students for participating in this study, and to the CELSTEC pub.group members for their comments on a previous draft of this article.

Abstract

In this study, it was explored whether a testing-effect occurs not only retention of facts but also for application of principles and procedures. For that purpose, thirty-eight high-school students either repeatedly studied a text on probability calculations (SSSS), or studied the text, took a test on the content, restudied the text, and finally took the test a second time (STST). Results show that testing not only leads to better retention of facts than restudying, but also to better application of acquired knowledge (i.e., principles and procedures) in high-school statistics. In other words, testing seems not only to benefit fact retention, but also positively affects deeper learning.

Keywords: testing-effect, retention, application, mathematics.

The Testing-Effect for Learning Principles and Procedures

Although most students prefer re-reading, taking notes, or writing a summary when preparing for an exam (Karpicke, Butler, & Roediger, 2009; Palmatier & Bennet, 1974), research has shown that retrieving information through testing can be a far more effective study technique (see Carpenter & DeLosh, 2006; Carrier & Pashler, 1992; Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006a). This effect is known as the testing-effect, which refers to the finding that retrieval by means of testing increases long-term (generally one week) retention more as compared to re-studying, constructing a concept map, or re-reading notes (Karpicke & Blunt, 2011; McDaniel, Howard, & Einstein, 2009; Roediger & Karpicke, 2006b).

Testing and Deep Learning

Although the testing-effect has been well researched with respect to vocabulary words and facts from expository texts, there has been relatively little interest in the effects of testing on ´deep learning´ such as comprehension, application, or transfer. There are, however, some recent studies which have investigated whether testing can enhance learning beyond verbatim retention of facts and found that testing can also enhance deep learning (e.g., Agarwal, McDaniel, Thomas, McDermott, & Roediger, 2011; Butler, 2010; Marsh, Roediger, Bjork, & Bjork, 2007). From an educational perspective, this is an encouraging finding since educational learning goals usually require deeper learning instead of simple retention of facts or making inferences. Yet, there is still relatively little known about the testing-effect for application of principles and procedures in domains such as statistics, mathematics, or the sciences even though it is more important that students acquire the ability to apply *principles and procedures* (i.e., problem solving) in these domains rather than simply retrieve singular facts from memory. Therefore, the present study investigated whether a testing-effect could be

found in the domain of statistics using initial test questions requiring the application of principles and procedures (e.g., using the probability formula to calculate the probability that you throw five with a die). A principle is defined here as a basic rule of a theory (e.g., when the temperature in a closed space is increased, the pressure will also increase) and a procedure as a set of steps or rules that describe how something should be done to achieve a desired result (e.g., to distill alcohol from water, one should heat the liquid to just above to boiling point of the alcohol but below that of water, collect the evaporatation, and cool it down to a liquid state).

In the rest of this introduction, the processes underlying the testing-effect are explained shortly, and then the expected effects of application tests and results of prior studies are described. The introduction ends with a description of the research questions and hypotheses.

Why Is Testing Beneficial for Learning? A Testing-effect Perspective

Reading involves encoding processes and generally leads to storage of information in one's long-term memory (LTM). When a text or other material is re-read a couple of times, the memory traces that were created during initial reading are strengthened only slightly because no additional encoding processes are involved. However, when a test is made after initial reading, the information is not only encoded during initial study, but it is also retrieved from LTM. This retrieval is beneficial for learning, as it initiates the search for target information stored in LTM (i.e., information re-activation). This information re-activation is seen to strengthen existing memory traces to a greater degree than re-reading. Consequently, it facilitates later retrieval by the same or associated cues because cue (i.e., question) and target (i.e., answer) have become associated during testing (Bjork & Bjork, 1992). Thus, by retrieving target information from memory, it becomes easier to access. In contrast, when

re-reading a text -without retrieval- no connections are made between the cue and its target, making subsequent access to the information more difficult. As a result, testing is more beneficial for retention of that information than re-reading.

Although this effect of testing on retention is already fairly strong with a single test phase, research has shown that including multiple tests and restudy opportunities can increase the testing-effect to a large extent (e.g., McDaniel & Fisher,1991; Pashler, Cepeda, Wixted, & Rohrer, 2005; Roediger & Karpicke, 2006a). Whereas re-reading after testing provides immediate feedback and helps to correct errors (Foos & Fisher, 1988; Hamaker, 1986), repeated testing bolsters the testing-effect because information is repeatedly retrieved, and with every successful retrieval these memory traces are strengthened. Including a restudy phase also mirrors real classroom practice because students will mostly use tests as formative assessment instruments, using the formative information to guide restudy behavior or correct errors during renewed self-testing (e.g., Karpicke, Butler, & Roediger, 2009).

Testing-effect for Deep Learning

Although there has been a great deal of research on testing to enhance retention of facts (see Roediger & Karpicke, 2006a for an overview) and some recent interest in testing for deep learning (e.g., inference making, comprehension), there is still rather little known about the effects of testing on application of principles and procedures. An exception is a recent study by Agarwal et al. (2011). Their study investigated whether testing in the classroom prompts application of constructs and principles (experiment 2a and 2b) in 7^{th} and 8^{th} grade students. In those experiments, one-third of the items were initially tested in a concept-term format (e.g., “What is the struggle between organisms to survive in a habitat with limited resources? (A) Parasitism, (B) Competition, (C) Limited Factors, or (D) Predation”, answer = B), one-third in an application format (e.g., “Both foxes and raccoons on Long

Island eat pheasant, which has been in decline in recent years. The foxes and raccoons´ situation is an example of what ecological process? (A) Parasitism, (B) Competition, (C) Limited Factors, or (D) Predation", answer = B), and one-third was not tested. Concept questions tested conceptual knowledge, whereas application questions required *discovering what principle or construct was being illustrated* in a particular scenario or situation (Note: Many might consider this not to be real application, but rather recognition of a principle). The initial test and posttest items tested the same principles, but used different cover stories to cue principle recognition. For example, one of the initial application test questions was: "When Sally is at home by the fireplace, smoke rises up the chimney because hot air rises, and partly because it is pushed by the wind blowing across the top of the chimney. This lowers the overall air pressure causing the high pressure at the bottom to push up the smoke. What principle keeps smoke from filling up the room? (A) Mead's principle; (B) Bernoulli's principle, (C) Piaget's principle, (D) Erikson's principle" (answer = B). The related posttest question was: "When a pitcher throws a curve ball, the spin of the ball created high pressure on the top of the ball, which pulls the ball downward. What principle is being illustrated in this example?" (answer: Bernoulli's principle). Agarwal et al. found that previously tested content was learned better than untested content and that discovering what concept or principle was being illustrated led to better delayed retention of definitional information and application (i.e., discovering what concept was illustrated in new contexts). Although these results are very interesting and promising, it is debatable whether the application questions used by Agarwal et al. required the real application of a principle, or whether they instead required the recognition of the principle. Additionally, the study of Agarwal et al. did not look into the effects of testing on the application of procedures, but only at the application of constructs and principles.

Two studies focusing on worked examples (i.e., on the electrical circuit) might also be interesting here. In two separate studies, Van Gog and Kester (in press) and Van Gog, Kester, Dirkx and Hoogerheide (submitted) investigated if problem solving (i.e., testing) can increase performance on a posttest more than studying worked examples (i.e., study). Worked examples provide learners with a written, step-by-step solution procedure to a problem and are especially effective for novice learners because they contain very little distracting information. In the studies by Van Gog and Kester and Van Gog et al., it was found that *repeatedly* solving problems led to better test performance on an *immediate* posttest, including the same problems, when compared to restudying worked examples. However, no effects were found on a one week delayed test. Although these studies used worked examples instead of expository texts, the results are intriguing, as they show that repeated application of procedures can enhance performance on an *immediate* test.

The studies discussed here provide valuable information on the testing-effect for application of principles (Agarwal et al., 2011) and procedures (Van Gog & Kester, in press; Van Gog, et al., submitted). However, there are still a number of open questions, such as: Are the effects of testing for principles and procedures also found among different domains (e.g., statistics)? Are the effects of testing for principles and procedures also found on a delayed test? Are the effects of testing for principles and procedures also found when learning from an expository text? Although many of these questions have been resolved for learning facts, more research is needed to resolve these questions for learning principles and procedures.

Application of Principles and Procedures

In contrast to making inferences or recognizing facts and concepts, applying principles and procedures requires not only knowing what information is asked for and activating the information in memory (Barnett & Ceci, 2002; Bloom, 1956; Hamaker, 1986), but also

successful application of the principles and procedures. Therefore, the question must be translated into a set of learned concepts to allow for selecting relevant / required information (Barnett & Ceci called this the *recognition process*). When a student has learned, for example, the procedure for calculating the probability of picking a black marble out of a box containing 10 marbles in which only one is black, and is then asked to calculate the probability of buying the winning raffle ticket at a raffle in which a 100 tickets were sold, the student needs to translate the raffle question into a question with marbles. To do this the student, for example, might reformulate the problem as: Calculate the probability of picking the one black marble - which represents the winning ticket - out of all of the 100 marbles - which represents the raffle tickets. This step leads to a better organized knowledge base and helps the learner choose the correct principle or procedure (Hamaker). After translating the question, the necessary information (e.g., principle or procedure) must be retrieved from memory. In the example just given, the learner needs to retrieve the procedure 'dividing the number of possible targets by the total number of possibilities' ($P(G)$ = number of targets ÷ total number of possibilities). This retrieval ‘step’ strengthens the memory traces for these principles or procedures (i.e., testing-effect) making them more accessible during later tests (Barnett & Ceci; Butler, 2010). Finally, the learner must successfully apply the procedure (i.e., one winning ticket ÷ 100 sold tickets) and answer the question (i.e., the chance of buying the winning ticket is 1%). This step is also known as the *execution process* (Barnett & Ceci).

Research Questions and Hypotheses

The present study investigated whether repeated testing not only facilitates the retention of facts, but also the application of principles and procedures to solve problems on a delayed posttest. For that purpose, two types of questions (i.e., fact questions and problem solving questions requiring the application of principles and procedures) were used during

initial testing and it was investigated if answering these types of questions during initial testing would benefit performance on a delayed posttest.

Based on the testing-effect literature, it was expected that repeated testing (with restudy) for factual knowledge, as well as testing for application of principles and procedures, will benefit performance on a delayed factual knowledge posttest and a delayed application posttest because the initial tests will help students to build strong memory traces for the retrieved information.

Method

Participants

Participants were 38 Dutch-speaking secondary school students (60% male) in their 4th year of a 6-year, pre-university school trajectory (range = 15-16 years; *M* = 15.91; *SD* = .67). The experiment took place during regular lessons in groups of 15-20 students.

Materials

The complete experiment was paper-and-pencil based. The following materials were used:

Expository text. An expository text of 899 words on probability calculations was written based on a mathematics textbook used throughout the Netherlands (Reichard, 2003). It was rewritten in such a way that all participants would be able to comprehend the text without the aid of a teacher. The Flesch-Douma readability score - the Dutch equivalent of the Flesch reading ease score - was 71 which indicates that the text was suitable for the nominal reading level of the participants (Hoogteijling, 1967).

The text contained five sections with an average length of 180 words per section (range = 131-266 words). Section one described and explained the formula used in probability calculations (i.e., the Laplace probability definition). Section two introduced a special kind of

probability calculation (i.e., the vase model). Sections three, four, and five explained how to solve different types of probability calculation problems, including an example problem for each type of problem. The participants had not studied this particular topic yet at school.

Factual knowledge test. The factual knowledge test consisted of five short-answer questions, one for each text section. The factual knowledge questions asked for facts that were not covered by the application questions (see appendix). An example of a factual knowledge question is: What does the letter *G* stand for in the formula? (Answer: *G* stands for the situation for which you want to calculate the probability). The answers to the factual knowledge questions could be found literally in the text. A pilot study showed that, on average, participants were able to answer about 50% of the questions correctly indicating that the test was not too easy but also not too difficult for them.

Application test. The application test consisted of five short-answer questions, one for each section. The questions required participants to apply a principle/procedure that was explained and illustrated in the original text to a new situation. An example of an application question is: Denise bought a ticket from a raffle. From the 100 tickets sold, there are two winning tickets. Simulate this situation with the vase model (Answer: There are 2 green marbles and 98 black marbles in the vase. The green marbles are the prize-winning raffle tickets.). In the original text, two examples (i.e., different examples than in the initial test) of such simulations were given (thus the study-only group read these examples four times). A pilot study showed that, on average, the participants were able to answer about 50% of the questions correctly indicating that the test was not too easy but also not too difficult.

Postttest. The posttest contained the 10 short-answer questions (five factual and five application) that were used in the initial test.

Distracter task. A Sudoku puzzle was used as a distracter task. A Sudoku puzzle is a partially completed 9X9 number grid that must be completed in such a way that each column, each row, and each of the 9 3X3 boxes contain all of the numbers from 1 to 9 only once.

Design and Procedure

A between-subject design with "Learning Sequence" (Study-Study-Study-Study (SSSS), Study-Test-Study-Test (STST)) as between subject factor was used. Participants either studied the expository text repeatedly (SSSS), or studied the complete text, then took the initial tests, restudied the text and took the tests a second time (STST). The participants were randomly distributed across the experimental conditions and tested during regular school hours in groups of 15-20 students.

The participants received a condition-dependent envelope with the assignments. They were told to read the instructions in the envelope carefully, so that they knew exactly what was expected from them. The instructions explained the experimental procedure so that the participants knew that they were preparing for a delayed posttest consisting of factual knowledge questions and application questions. After the participants read the instructions, the experimenter orally repeated the instructions and emphasized that they were to read the text carefully. Every new task/learning activity was introduced in the same way (in written and oral form).

All participants, regardless of experimental condition, studied the text in an initial 8-minute study-period. The study time was based on the reading fluency scores reported by Hasbrouck and Tindal (2006) and a pilot study - which confirmed that 8 minutes was sufficient for initial study. Participants in the SSSS condition then restudied the text in three consecutive study phases of 8 minutes separated by a 2-minute distracter task. For participants in the STST condition the initial 8 minute study phase was followed by the initial tests (i.e., the factual knowledge and application test (8 minutes)). Then, they restudied the text (8

minutes), and took the tests a second time (8 minutes). Here too, the phases were followed by a distracter task. One week after this learning phase, all participants took the posttest (10 minutes). The learning phase and posttest were thus equally long for both conditions (4X8 minutes reading and/or testing plus 3X2 minutes on a distracter task and 10 minutes posttest).

Scoring

Test questions. The factual and application questions were awarded a full point when answered correctly. Some questions contained two partial answers. In such cases, a half-point was awarded for each partially correct answer. For example: There are different situations for which you can calculate probabilities. Each situation has a different procedure that needs to be followed. Describe the first situation as explained in the text (i.e., this is an example of a factual question). Answer: Without putting the marbles back in the vase (.5 points) and taking the marbles one by one (.5 points). No points were awarded for missing or incorrect answers. The proportion of correct answers was calculated for each test.

Analysis

Initial tests. The percentage correct on the initial tests was calculated. Then, in order to give insight in the effects of the restudy phase, the number of items answered correct on both initial tests (c–c) or answered correct only on the second initial test (i–c) was calculated and frequency statistics were calculated for the number of items remembered and the number of items that were forgotten. Finally, a repeated measures analysis tested if participants in the STST condition significantly improved when they took the initial test a second time.

Posttest. With two independent sample t-tests it was investigated is the participants in the STST condition significantly outperformed participants in the SSSS condition on the factual knowledge and application posttest questions.

Results

In all analyses reported, a significance level of $p \leq .05$ was used. When significant effects were reported, Eta square (Pearson, 1911) was used as a measure of effect size for analyses of variance (ANOVA´s) and Cohen´s *d* (Cohen, 1988) for *t*-test analyses.

Initial test. On the first administration of the initial test, participants correctly answered 33.50% (*SD* = 19.81) of the factual knowledge and 35% (*SD* = 22.36) of the application questions. On the second administration, participants answered 56% of the factual knowledge (*SD* = 22.34) and 56.50% application questions (*SD* = 17.55) correct. Frequency statistics show that only 7% of the factual questions were answered correctly on both tests. However, the restudy phase seemed to affect correction of incorrect answers to a large extent since 23% of the items - which were answered incorrectly on the first administration - was correctly answered on the second administration. For the application questions, 25% of the items was correctly answered on both administrations and the restudy phase seemed to affect the correction of errors for the application questions to a large extent since 32% of the items - which were incorrectly answered the first time - was correctly answered the second time (see Figure 1).

Insert Figure 1 about here

A repeated measures analysis confirmed that participants significantly improved with the second administration of the initial test for factual and application questions respectively ($F(1,19) = 20.97$, $MSE = .60$, $p < 0.001$, $\eta p^2 = .53$ and $F(1,19) = 20.97$, $MSE = .60$, $p < 0.001$, $\eta p^2 = .51$).

Posttest. An independent sample *t*-test showed a significant testing-effect of Learning Sequence on the proportion of factual knowledge questions that were correctly answered ($t(36) = 4.38$, $p < .001$, $d = 1.44$). Participants in the STST group outperformed participants in the SSSS group on the factual knowledge questions ($M = 49.50\%$; $SD = 22.59$; $M = 21.67\%$; $SD = 15.44$ respectively; see Figure 2).

A significant effect of Learning Sequence was also found for the application questions ($t(36) = 2.97$, $p = .005$, $d = .96$). Participants in the STST group performed better ($M = 60\%$; $SD = 23.40$) on the application questions than participants in the SSSS group ($M = 37.78\%$; $SD = 22.64$).

Insert Figure 2 about here

Discussion

While students and teachers hold that restudying is the best strategy for learning, testing-effect studies show that testing may provide larger benefits. However, these studies mostly focused on the learning of vocabulary, facts, or concepts and not on deeper levels of learning (i.e., knowledge application). The study reported on here investigated whether answering questions requiring the application of principles and procedures during learning benefits problem solving in mathematics. In line with the expectations, testing benefitted not only the retention of facts from a mathematics text, but also the application of the principles and procedures contained in that text. More specifically, answering factual questions - while learning - enhanced retention of that information after one week (in accordance with earlier testing-effect studies) and, more importantly, answering questions requiring the application of principles and procedures during learning led to better performance on a delayed posttest requiring application of principles and procedures. In that regard, the research reported on here contributes to the sparse body of literature focusing on the testing-effect for application of principles and procedures.

The present study supports the findings of Agarwal et al. (2011), Van Gog and Kester (in press), and Van Gog, et al. (submitted) that testing can be beneficial for application of principles and procedures. However, in contrast to Agarwal et al., the application questions used here required the explicit *application of principles or procedures* to solve probability calculations instead of the identification of a construct or principle being illustrated as in the

Agarwal et al. study. In other words, the current study shows that testing can foster application of principles and procedures to solve problems; here probability calculations. These findings fit the results of Van Gog and Kester (in press) and Van Gog et al., (submitted), but, although Van Gog and Kester and Van Gog et al. only found results on the short term, the present study shows that the beneficial effects of testing for application can also be found after a one week delay.

The study presented here included a restudy phase and a subsequent test opportunity. This design imitates not only educational practice where students often restudy the learning material after taking an initial (self-)test and often repeat the (self-)test after restudy. Prior studies have also shown that for successful application and transfer (i.e., application of knowledge to a new situation) repeated testing is an important prerequisite (e.g., Butler, 2010; Van Gog and Kester, submitted) since students need a solid memory representation in order to be able to apply or transfer their knowledge. It might however be critiqued that it is difficult to disentangle the benefits of testing versus restudy with this design. It would therefore be interesting to further investigate this issue and see if testing for principles and procedure also enhances learning when only one test moment is included - with or without restudy -, or what the effects are of repeated testing without restudy. This would give more insight in the processes underlying the testing-effect. An advantage of the design (i.e., testing with restudy) used in the present study is however that it represents real-world classroom practice in the sense that in real-classroom practice students will generally look at the materials after making a initial test or self-test (Karpicke, Butler, & Roediger, 200) and then try to solve the test questions a second time (i.e., use tests as formative assessment instruments). A second limitation (see for example Butler & Roediger, 2008; Carrier & Pashler, 1992; Karpicke & Roediger, 2008) is that the items in the initial and posttest were identical, and thus nothing can be reported on the application of the acquired knowledge in slightly (i.e., near transfer) to

greatly (i.e., far transfer) different problem situations. Now that there is a certain degree of proof that the testing-effect is valid for application of principles and procedures to solve probability calculations, future research will look into this transfer issue.

Implications

In spite of these limitations, the results might be regarded as interesting, both from a theoretical and an applied perspective. The results show that initial testing for application of principles and procedures can be beneficial for learning. More specifically, that a testing-effect can be found not only when using factual or concept questions, but also when application questions are used during the learning phase. Applying knowledge to answer a question or solve a problem requires strong traces in one's memory of the principles and procedures involved in order to successfully apply them. The learner needs *to translate the question or problem* into learned concepts, *retrieve relevant principles or procedure*, *select the correct principle or procedure* to answer the question or solve the problem, and finally *apply them correctly* to the question in order to answer it. As this study shows, initial testing can help to create strong memory traces for such complex learning tasks and enhance delayed test performance.

From a more applied perspective, the results of the present study suggest that teachers should rely more on tests to *support* learning, instead of only using tests to *assess* learning. The advantage of real-classroom practice is that there is more time to integrate testing into the curriculum (as compared to the limited learning-time available in this experiment). Teachers could for example use daily tests at the end of the course hour (i.e., quizzing), use it in online courses, or support students to use self-tests during learning (see Carpenter, Pashler, & Cepeda, 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Anderson, Derbisch, & Morisette, 2007, for research on the testing-effect in real classroom practice for learning facts). This could heavily enhance learning. Research (for facts,

definitions and concepts), has namely shown that frequent testing, more variation in test questions/problems, and more time for schema construction bolsters the testing-effect - even without taking to much time away from classroom instruction (e.g., Butler, 2010; Roediger & Karpicke, 2006a).

Thus, although most teachers and students perceive testing as an assessment tool and restudy as the primary learning strategy, this study shows that testing is a much more effective study strategy, not only for the retention of facts, but also for application of principles and procedures.

References

Agarwal, P. K., McDaniel, M. A., Thomas, R. C., McDermott, R. C., & Roediger, H. L., III (2011, March). *Quizzing promotes deeper acquisition in middle school science. Transfer of quizzed content to summative exams*. Paper presented at the SREE Conference, Washington D.C., United States.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612-637. doi: 10.1037/0033-2909.128.4.612.

Bjork, R. A. & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp, 35-67). Hillsdale, NK: Erlbaum.

Bloom, B. S. (1956*). Taxonomy of educational objectives: The classification of educational goals.* Essex, England: Harlow.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology, 36,* 1118-1133. doi: 10.1037/a0019902.

Butler, A. C. & Roediger, H. L., III (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.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268-276. doi: 10.3758/BF03193405.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. Applied Cognitive Psychology, 23, 760 –771. doi:10.1002/acp.1507.

Carrier, S. K. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633-642. doi: 10.3758/BF03202713.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.) Hilsdale, NJ: Erlbaum.

Foos, P. W. & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80,* 179-183.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*, 212-242.

Hasbrouck, J., & Tindal, G.A. (2006). Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers *The Reading Teacher, 59,* 636–644. doi: 10.1598/RT.59.7.3.

Hoogteijling, J. (1967). Toetsing van schoolboeken en examenopgaven door vaststelling van de leesbaarheid volgens Flesch-Douma. [Testing of schoolbooks and exams through determining the readability using Flesch-Douma]. *Pedagogische Studiën, 44,* 366-399.

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

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772-775. doi: 10.1126/science.1199327.

Karpicke, J. D., Butler, A.C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471-479. doi:10.1080/09658210802647009.

Karpicke, J. D. & Roediger, H. L., III (2008, February 15). The critical importance of retrieval for learning. *Science, 15,* 966-968. doi: 10.1126/science.1152408.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*, 194-199. doi: 10.3758/BF03194051.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III (2011). Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement. *Journal of Educational Psychology*, *103,* 399-414. doi: 10.1037/a0021782.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. doi:10.1080/09541440701326154.

McDaniel, M. A. Howard, D. C. & Einstein, G. O. (2009). The read-recite-review strategy: Effective and portable. *Psychology Sciences, 20,* 516-522. doi: 10.1111/j.1467-9280.2009.02325.x.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192-201. doi: 10.1016/0361-476X(91)90037-L.

Palmatier, R. A., & Bennet, J. M. (1974). Note-taking habits of college students. *Journal of Reading*, *18,* 215-218.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3-8. doi: 10.1037/0278-7393.31.1.3.

Pearson, K. (1911). On a correction needful in the case of the correlation ratio. *Biometrika, 8,* 254-256.

Reichard, L. A. (2003). *Getal en ruimte. VWO A/B [Numbers and space. Pre-university education A/B]*. Netherlands, Houten: EPN.

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,* 3, 249-255. doi: 10.1111/j.1467-9280.2006.01693.x.

Van Gog, T., & Kester, L. (in press). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*. doi: 10.1111/cogs.12002.

Van Gog, T., Kester., L., Dirkx, K. J. H., & Hoogerheide, V. (submitted). Does Testing after Worked Example Study Improve Delayed Problem-Solving Performance?

Appendix 1

The probability definition contains different symbols. The symbol *P* for example stands for probability and *G* stands for the situation for which you want to calculate the probability, such as the situation of throwing two (sum is two) with two dice. How do you calculate this probability? When you throw two dice, there are 6X6 = 36 possible outcomes (sum scores), because every dice has six sides, each with a different number of dots. For the situation *(G)* 'sum is two' there is only one possibility, namely that you throw one with the first dice and one with the second. Thus, there is only one possible outcome for 'sum is two' and 36 possible outcomes of throwing two dice. According to the formula *P*(sum is two) = 1/36.

Factual question

1) What does the symbol G in the formula stand for?

Application question

1) Robin throws two dice. Calculate the probability that he throws a 12.