This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

The optimal learning strategy depends on learning goals and processes : retrieval practice versus worked examples

Yeo, Darren J.; Fazio, Lisa K.

2019

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes : retrieval practice versus worked examples. Journal of Educational Psychology, 111(1), 73–90. doi:10.1037/edu0000268

https://hdl.handle.net/10356/144054

https://doi.org/10.1037/edu0000268

© American Psychological Association, 2018. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: https://doi.org/10.1037/edu0000268

Downloaded on 06 Nov 2024 01:05:18 SGT

The Optimal Learning Strategy Depends on Learning Goals and Processes: Retrieval Practice versus Worked Examples

Darren J. Yeo
Vanderbilt University and Nanyang Technological University

Lisa K. Fazio
Vanderbilt University

Note: This is a post-print version of
Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology, 111*(1), 73–90. https://doi.org/10.1037/edu0000268

Author Note

Darren J. Yeo, Department of Psychology & Human Development, Vanderbilt University, Division of Psychology, Nanyang Technological University; Lisa K. Fazio, Department of Psychology & Human Development, Vanderbilt University.

This research was supported by a Humanities, Arts, and Social Sciences International PhD Scholarship from Nanyang Technological University and the Ministry of Education (Singapore) to DJY. We would like to thank Ken Koedinger for his help in inspiring and designing Experiment 1. In addition, we would like to thank Emily Conder, Hana Crandall, Mimi Zhang, Morgan Goddard, Samantha Betman, and Ye Rin Lee for their valuable assistance with data collection and coding. Earlier versions of the results were presented at the 56th annual meeting of the Psychonomic Society, Chicago, November 2015 (Experiment 1), the 29th annual convention of Association for Psychological Science, Boston, May 2017 (Experiments 1-2), and the 58th annual meeting of the Psychonomic Society, Vancouver, November 2017 (Experiments 1-3).

Correspondence concerning this article should be address to Lisa K. Fazio, Department of Psychology & Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN, 37203. Email: lisa.fazio@vanderbilt.edu

Abstract

Testing (having students recall material) and worked examples (having students study a completed problem) are both recommended as effective methods for improving learning. The two strategies rely on different underlying cognitive processes and thus may strengthen different types of learning in different ways. Across three experiments, we examine the efficacy of retrieval practice and worked examples for different learning goals and identify the factors that determine when each strategy is more effective. The optimal learning strategy depends on both the kind of knowledge being learned (stable facts vs. flexible procedures) and the learning processes involved (schema induction vs. memory and fluency-building). When students' goal was to remember the text of a worked example, repeated testing was more effective than repeated studying after a one-week delay. However, when students' goal was to learn a novel math procedure, the optimal learning strategy depended on the retention interval and nature of the materials. When long-term retention was not crucial (i.e., on an immediate test), repeated studying was more optimal than repeated testing, regardless of the nature of materials. When long-term retention was crucial (i.e., on a one-week delayed test), repeated testing was as effective as repeated studying with non-identical learning problems (that may enhance schema induction), but more effective than repeated studying with identical learning problems (that may enhance fluency building). Testing and worked examples are both effective ways to learn flexible procedures, but they do so through different mechanisms.

Keywords: Testing effect; Worked examples; Retrieval practice; Mathematics; Instruction

Educational Impact and Implications Statement

This study suggests that learning strategies should be flexible across and within domains. Consistent with recent frameworks, rigid dichotomies between domains and instructional sequences should be avoided. The optimal learning strategy depends on the kind of knowledge to be learned (e.g., stable facts versus flexible procedures) and the target learning processes (e.g., inducing an underlying principle versus memory and fluency building).

The Optimal Learning Strategy Depends on Learning Goals and Processes: Retrieval Practice versus Worked Examples

What are the best ways to improve students' learning and retention of new information? In a report commissioned by the US Department of Education, a collection of teachers, learning scientists and psychologists identified seven recommendations on how to organize instruction to improve student learning (Pashler et al., 2007). One recommendation was that teachers should "use quizzing to promote learning" (Pashler et al., 2007, p. 2). That is, teachers should promote learning by having students actively recall information from memory, rather than simply restudying the information. A second recommendation was that teachers should "interleave worked example solutions with problem-solving exercises" (Pashler et al., 2007, p. 2). That is, rather than solve twice as many problems, which has been the conventional instruction for problem solving, students should alternate between studying worked-out solutions and solving the problems themselves.

Both recommendations are strongly supported by previous research (see Adesope, Trevisan, & Sundararajan, 2017; Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2014; Rowland, 2014, for reviews and meta-analyses), but it is currently unclear when teachers should rely on each learning strategy. The two strategies rely on different underlying cognitive processes and thus strengthen different types of learning in different ways. Therefore, the ideal learning strategy for a given situation will depend upon the goals of the learner (i.e., are they attempting to remember new information, learn a new problem-solving strategy, or generalize a current strategy to new problems) and the materials being used. Across three experiments, we aim to examine the efficacy of retrieval practice and worked examples for different learning goals and identify the factors that determine which strategy is most effective in a given situation.

Evidence for the Benefits of Retrieval Practice

In one of the canonical studies demonstrating the advantage of retrieval practice (Roediger & Karpicke, 2006a), students began by studying a text passage. Some students then restudied the passage three more times, whereas others tried to recall the passage on three consecutive trials. After a delay, all students were asked to recall the material. The repeated studying group had slightly higher recall performance than the repeated testing group after a 5-minute delay. However, after one week, the repeated testing group recalled much more of the passage, despite considerably less exposure to the material. This pattern of similar recall at short delays, but large benefits for retrieval at longer delays, is commonly found in the literature (see Adesope et al., 2017; Rowland, 2014, for meta-analyses).

Retrieval practice is thought to be a desirable difficulty, which is an instructional manipulation that introduces difficulties during study, but promotes long-term retention (Bjork, 1994). This testing advantage, commonly referred to as the testing effect or retrieval practice effect, has been observed with various types of materials (see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, for a review) including foreign language vocabulary (Fazio, Huelser, Johnson, & Marsh, 2010), general knowledge facts (Roediger & Marsh, 2005), spatial maps (Carpenter & Pashler, 2007), resuscitation skills (Kromann, Jensen, & Ringsted, 2009), and inductive input-output function learning (Kang, McDaniel, & Pashler, 2011), and has been found in both laboratory studies and in science and social studies classrooms (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger, Agarwal, McDaniel, & McDermott, 2011).

Evidence for the Benefits of Worked Examples

In mathematics instruction, problem-solving practice is a common instructional approach seen in both math textbooks and math classrooms. Typically, a worked example is followed by problem-solving practice that involves the retrieval of learned procedures. While problem-solving practice has been found to be more effective than studying of worked examples in at least one study (Darabi, Nelson, & Palanki, 2007), there is growing evidence that retrieval practice is no more effective than repeated studying for learning flexible procedures, and under some circumstances may even be suboptimal (Leahy, Hanham, & Sweller, 2015; van Gog et al., 2015; van Gog & Kester, 2012; see van Gog & Sweller, 2015, for a review). In fact, research examining problem-solving instruction in math and science has found that replacing problem-solving tasks with worked examples is beneficial for learning (see Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2014; van Gog & Rummel, 2010, for reviews). For example, students learned more algebra when they alternated between studying worked examples and solving problems versus solving twice as many problems (Sweller & Cooper, 1985). In a classroom study, students learned a 3-year curriculum in only two years with equivalent or better performance by replacing some of the problem-solving practice with carefully selected worked examples (Zhu & Simon, 1987).

This worked example advantage is commonly referred to as the worked example effect (Sweller, 2010). Under the cognitive load theory, worked examples are thought to reduce extraneous cognitive load (e.g., performing computations) in novice learners, which frees up working memory resources to acquire the underlying schema and learn the procedure (Kalyuga, Renkl, & Paas, 2010; Sweller, 1988, 2010). In summary, these findings indicate that studying additional worked examples is often more effective than solving additional problems, at least in the problem-solving domain.

Is Testing Effect Absent for "Complex" Learning Materials?

Some researchers have suggested that retrieval practice is not beneficial for "complex" materials such as those typically used in the worked examples literature (e.g., van Gog & Sweller, 2015). Highly complex learning materials are defined as "high in element interactivity, containing various information elements that are related and must therefore be processed simultaneously in working memory" (van Gog & Sweller, 2015, p. 248). This hypothesis suggests that the benefits of retrieval practice are restricted to learning materials that are less complex, and contain fewer interacting elements among the to-be-learned ideas (see van Gog & Sweller, 2015).

The main criticisms against the material complexity hypothesis are the lack of an objective measure of complexity, and evidence demonstrating a small, but positive testing effect with complex materials, including studies that contrasted problem-solving practice with worked examples (Karpicke & Aue, 2015; Rawson, 2015). For instance, Darabi, Nelson, and Palanki (2007) had engineering majors learn to diagnose and repair malfunctions in a simulated water-alcohol distillation plant, and then either study four descriptive worked examples, or complete four problem-solving practice trials. Problem-solving practice yielded higher problem-solving performance than studying worked examples (Darabi et al., 2007). Moreover, over four experiments, van Gog and colleagues (2015) found repeated testing to be as effective as studying worked examples during procedural learning involving electrical circuits. Taken together, the complexity of materials does not seem to be a major factor underlying the lack of benefits of

retrieval practice in procedural learning. Nonetheless, it is crucial to note that complexity of materials has not been explicitly contrasted or held constant in prior studies.

The Optimal Strategy May Depend on Learning Goals and Processes

Rather than focusing on differences between the materials used in studies examining the benefits of retrieval practice and worked examples, we believe that the key difference lies in the students' learning goals. Recent theoretical frameworks (Kalyuga & Singh, 2016; Koedinger, Corbett, & Perfetti, 2012) emphasize that learning is multi-faceted with varying overall goals and sub-goals, and the optimal strategy depends on the specific goal of the instructional tasks and the cognitive processes involved.

Knowledge-Learning-Instruction framework

The Knowledge-Learning-Instruction (KLI) framework proposed by Koedinger, Corbett, and Perfetti (2012) emphasizes that cognitive science and educational research have often focused exclusively on *instructional events* (e.g., retrieval practice and studying worked examples) and *assessment events* (e.g., recall and transfer tests), both of which are observable and can easily be manipulated experimentally (Koedinger et al., 2012). Much neglected by researchers has been the *learning processes* (e.g., memory and fluency-building, induction, sense-making) and *knowledge components* (e.g., facts, associations, categories, schemas, rules, procedures, principles) that are often not observable. The framework further suggests that the fit between the nature of the knowledge components to be acquired and the learning processes needed to acquire them is critical for performance on the learning assessment. For instance, a common type of knowledge has *constant* application conditions or cues, and *constant* target responses (e.g., learning that the word "merci" [constant cue] in French means "thank you" [constant response] in English, or that the area of a circle [constant cue], is $A = \pi r^2$ [constant response]). These are known as constant-constant knowledge components in the KLI framework (Koedinger et al., 2012). Koedinger and colleagues (2012) propose that constant-constant knowledge components necessitate predominantly memory processes, especially when assessed using long-term recall tests. Given that most of the materials used to examine the benefits of retrieval practice involve such constant-constant mappings, the authors hypothesize that retrieval practice is the optimal instructional event for constant-constant mappings.

In contrast, many domains require students to acquire knowledge that may have *variable* application conditions or cues and *variable* responses (e.g., learning that the modal verb "would" can be used as a past tense of "will", to talk about hypothetical situations, or for politeness, and that any number of fractions can be added using this general "make the denominator the same" formula $\frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{bc}{bd} = \frac{ad+bc}{bd}$). These are known as variable-variable knowledge components in the KLI framework (Koedinger et al., 2012). The authors propose that variable-variable knowledge components necessitate predominantly induction or compilation processes, especially when assessed using transfer or application tests. Given that most of the materials used to examine the benefits of studying worked examples involve variable-variable mappings, Koedinger and colleagues (2012) hypothesize that studying worked examples is the optimal instructional event for variable-variable mappings.

Crucially, when the fit between the knowledge components and the learning process is poor, learning will be suboptimal. An example of such a mismatch between the kind of knowledge to be acquired (i.e., learning goal) and learning processes is observed in a study by van Gog and colleagues (2015). The researchers had students in one learning condition recall

two worked examples (i.e., example-recall-example-recall) as they would with a factual passage. However, their final test involved a transfer-based problem-solving test. Hence, the students' learning process and goal during the learning phase (i.e., encode and fluently recall stable facts) and the ultimate goal during the test phase (i.e., flexible problem solving) did not match. Consistent with the KLI framework, they did not observe a testing advantage relative to students who repeatedly studied worked examples with the goal of learning the procedure (van Gog et al., 2015).

Reconceptualized Cognitive Load Theory

Building upon the KLI framework and reconceptualizing cognitive load theory, Kalyuga and Singh (2016) propose that learning a procedure as an overall goal is a complex learning task that can be differentiated into multiple sub-goals involving different knowledge components and learning processes (Kalyuga & Singh, 2016). That is, in math and science procedural learning, not only do students need to induce the underlying principles from the worked example, but they also need to gain fluency in applying the procedures and commit them to memory.

Furthermore, even the same instructional event can be associated with different goals depending on its context, resulting in vastly different outcomes. For instance, on one hand, there has been substantial evidence that support the efficacy of providing a worked example before problem-solving practice (e.g., van Gog, Kester, & Paas, 2011), which is consistent with the cognitive load theory. On the other hand, there has also been substantial evidence supporting the efficacy of the reverse learning sequence (i.e., problem-example) under the "productive failure" and "invention learning" frameworks, which posit that problem solving before explicit instruction may serve to help students attend to relevant and critical aspects of the solutions when they become available (e.g., Kapur, 2008, 2010, 2014; Schwartz & Martin, 2004). Kalyuga and Singh (2016) suggest that this apparent contradiction can be resolved if we were to consider the same problem-solving activity as having distinct goals depending on its position in the learning sequence. Specifically, the goal of the problem-solving phase in the example-problem instruction is to fluently apply and reinforce the underlying principles of a procedure, whereas the goal of the problem-solving phase in the problem-example instruction is to activate prior knowledge so that students are more aware of their knowledge gaps, and more likely to focus on the deep features rather than the surface features of the ensuing worked example (Likourezos & Kalyuga, 2017).

In other words, while learning flexible procedures (variable-variable knowledge components) may be the overall goal of problem-solving learning, both the induction of principles (variable-variable knowledge components), and automatizing the application of the procedure and committing it to memory as stable-stable knowledge components may be crucial sub-goals. Thus, both retrieval practice and worked examples may improve problem-solving learning through different mechanisms. Worked examples improve induction of the procedural principles, while retrieval practice improves memory for the procedure.

Current Research

The previous studies differ in the materials used, the kind of knowledge to be acquired, and the student populations studied. Studies examining the benefits of retrieval practice typically have subjects learn specific facts with the goal of recalling them later under the same application conditions (i.e., a recall test). Studies examining the benefits of worked examples typically have subjects learn flexible science or math problem-solving procedures with the goal of applying the

learned procedures later under variable application conditions (i.e., a novel problem-solving test). To the best of our knowledge, no study has examined the relative efficacies of repeated studying and retrieval practice as a function of the learning goals – remembering stable facts and learning flexible procedures – using the same materials (i.e., multi-step math worked examples with high element interactivity), within the same study.

Using the *same materials*, but with *different learning goals*, the current research aims to hold the complexity of materials constant and examine when retrieval practice and repeated studying are beneficial. Specifically, we seek to show that one of the key differences lies in the kind of knowledge being learned (overall learning goal).

Additionally, we are interested in whether the relationship between students' judgments of their learning and their test performance also depends on the learning goal. Few studies have examined metacognitive monitoring in problem solving. They either used a different type of problem solving (e.g., chess; de Bruin, Rikers, & Schmidt, 2007; de Bruin, Rikers, & Schmidt, 2005), or examined procedural problem learning in children and adolescents (Baars, van Gog, de Bruin, & Paas, 2014, 2017) – all with a focus on improving students' monitoring accuracy, rather than contrasting the impact of learning strategies on judgments of learning. To our knowledge, little is known about judgments of learning for procedural problem solving within adults.

In sum, we hypothesized that when students are attempting to remember a passage, repeated retrieval would be more effective than repeated studying, especially after a 1-week delay. In contrast, when students are attempting to learn a flexible procedure, repeatedly studying worked examples would be more effective than repeated problem solving. For judgments of learning, we hypothesized that students' judgments would be biased toward repeated studying when their goal was to remember a passage, as has been found previously (e.g., Roediger and Karpicke, 2006a). When their goal was to learn a procedure, our predictions were less clear. Students' judgments could be biased toward repeated testing, as it would be more apparent to them if they could solve the practice problems as compared to repeated studying (e.g., Baars et al., 2014, 2017). Alternatively, students' judgments of learning could be biased toward repeated studying regardless of the learning goal, due to the fluency of processing (Karpicke, Butler, & Roediger, 2009) or an illusory understanding during repeated studying (Renkl, 2002).

Overview of Experiments

The current research was modeled off of the paradigm used in Roediger and Karpicke (2006a). In Experiment 1, undergraduates were asked to study a math worked example with different goals. Some were asked to remember a passage (i.e., the text of the worked example) to be recalled at a later time (either after a 5-minute or 1-week delay), and others were asked to learn a procedure to be applied to novel problems. Within each group, some students engaged in repeated studying, and others engaged in repeated testing. Judgments of learning were measured at the end of the learning phase. Recall performance was then measured for the remember passage group and problem-solving performance for the learn procedure group. In Experiments 2 and 3, we focused on the overall goal of learning a flexible procedure and examined the effects of increasing retrieval success during learning and using identical or variable learning problems on the usefulness of worked examples and retrieval practice. These subsequent set of experiments provide further evidence for the need to conceptualize procedural learning as involving multiple *sub-goals* (i.e., learning processes and knowledge components), instead of the traditional focus on its *overall goal* (Kalyuga & Singh, 2016; Likourezos & Kalyuga, 2017).

Experiment 1

Method

Participants. One hundred and sixty adults from a highly selective university's human subject pool (48 males; M_{age} = 19.9 years, SD = 2.1, range: 18 – 29), participated in exchange for course credit or monetary incentive ($5 per half-hour of participation). Participants completed the experiment individually or in small groups of up to seven people. Participants had the choice to sign up for either the single-session (5-minute retention interval) or the two-session experiment (1-week retention interval). Within each retention interval, participants were then randomly assigned to one of the four conditions. There were no differences in prior knowledge across the conditions (see online supplement for analyses).

Design. A 2 (learning goal: remember passage vs. learn procedure) × 2 (learning strategy: repeated studying, SSSS vs. repeated testing, STTT) × 2 (retention interval: 5 minutes vs. 1 week) between-subjects design was used, resulting in 20 participants per cell.

Materials.

Learning Problems. Four probability word problems involving the Poisson distribution were created by the researchers for the learning task. Each problem presented during the learning phase involved a four-step solution (see Figure 1 for the key problem that all participants were exposed to on the first learning trial; hereafter referred to as the 'Airport Problem'), and contained explicit sub-goals designed to facilitate the generalization of the solution procedure (Catrambone, 1996, 1998). To be consistent with the problem-solving literature and with real-world problem-solving learning, all other learning problems were isomorphic to the Airport Problem (i.e., they had the same basic problem structure and required the same four-step procedure, but differed in their cover stories).

The Poisson distribution was chosen as a focal learning topic as it is not typically covered in regular high-school curricula or in advanced placement statistics classes, but is still accessible to undergraduate and graduate students without prior knowledge.

Suppose that the arrival and departure of airplanes at a domestic airport follow two independent Poisson distributions. In a one-hour period, it is expected on average that there are 4 arrivals and 3 departures. Find the probability that, in a randomly selected two-hour period, the airport handles 10 or more, but less than 13 arrivals and departures.

Step 1:
Let A be the number of arrivals in a two-hour period.
1 hour $\rightarrow$ 4 arrivals
2 hours $\rightarrow 4 \times 2 = 8$ arrivals
So, A ~ $P_o(8)$.

Step 2:
Let D be the number of departures in a two-hour period.
1 hour $\rightarrow$ 3 departures
2 hours $\rightarrow 3 \times 2 = 6$ departures
So, D ~ $P_o(6)$.

Step 3:
Let T be the total number of arrivals and departures in a two-hour period, i.e., A + D.
T~ $P_o(8 + 6)$ i.e., T ~ $P_o(14)$.

Step 4:
$P(10 \leq T < 13) = P(T = 10) + P(T = 11) + P(T = 12)$

$$= \frac{e^{-14}.14^{10}}{10!} + \frac{e^{-14}.14^{11}}{11!} + \frac{e^{-14}.14^{12}}{12!} = .249$$

Figure 1. The Airport Problem and its solution. This problem was presented to all participants during the first learning trial.

Test Problems. Eight test problems were administered, two of which were isomorphic to the Airport Problem, and six of which utilized a subset or a variation of the four-step procedure to assess transfer of learning (see Table S1 in the online supplement for examples of isomorphic and transfer test problems). The problem types were presented in a fixed sequence across all participants, with the isomorphic problems presented first. This allowed the participants to immediately see how the learned procedure was applicable for solving the test problems (e.g., Catrambone, 1996).

Procedure. The experiment included both a learning phase and a test phase that occurred either 5 minutes or 1 week apart.

Learning phase. During the learning phase, participants were told that they would study how to solve a particular type of probability problem, and that they would be tested on the material later. All participants, regardless of condition, first studied a printed cover sheet (see online supplement) that provided orienting information about Poisson distribution and the relevant formula so that they would understand the procedure in the worked examples. Participants were given three minutes to read the cover sheet and they were also allowed to refer to it throughout the experiment. Thus, participants did not have to memorize any formulas. This allowed us to focus solely on participants' difficulties in learning the procedure, rather than difficulties in remembering the formulas. Calculators were provided throughout the entire study, and the cover sheet also contained instructions on how to utilize the calculator.

After reading the orienting information, participants were told their assigned learning

goal. Those in the remember passage group were instructed to remember as many details as they could about the Airport Problem, along with its solution. Those in the learn procedure group were instructed to learn the procedure to solve probability problems using the Poisson distribution. All participants were then presented with the Airport Problem and its solution on a computer screen. This first learning trial (S_1) was four minutes long, followed by a one-minute filler task of solving a visuo-spatial puzzle. After this first problem, the experimental procedure and materials differed slightly depending on condition.

Remember passage group. Participants in the repeated studying condition studied the Airport Problem and its solution three more times ($S_1S_1S_1S_1$). They were given four minutes to study the problem and solution each time. Those in the repeated testing condition recalled the Airport Problem and its solution three consecutive times ($S_1T_1T_1T_1$). For each test trial, participants were given a blank sheet of paper and four minutes to recall as much as possible. In both conditions, participants solved a visuo-spatial puzzle as a distractor task for one minute in between each trial.

Learn procedure group. Participants in the repeated studying condition studied the worked solutions to three new problems, which were isomorphic to the Airport Problem, but contained different cover stories ($S_1S_2S_3S_4$). The repeated testing condition solved the same three novel problems in a printed booklet ($S_1T_2T_3T_4$). Again, each study or test trial lasted four minutes and the participants solved visuo-spatial puzzles for one minute in between each trial.

Both groups. At the end of the learning phase, a brief computerized questionnaire was administered. Participants first indicated if they had learned about the Poisson distribution prior to their participation in the study, and if so, how much they remembered on a 7-point Likert scale (1 = not at all, 7 = very much). We also asked about high school and college-level math and statistics classes that they had taken or were currently taking. Finally, participants were asked to make a judgment of their learning. Specifically, we asked them to rate how well they thought they would be able to recall the worked example (for the remember passage group), or be able to solve that particular type of problem (for the learn procedure group) one week later on a 7-point Likert scale (1 = not at all, 7 = very well).

Test phase. After either the 5-minute or 1-week delay, participants completed the final test. During the 5-minute delay, participants completed five visuo-spatial puzzles, each for one minute. The final test varied depending on the participant's learning goal. For participants who were tasked with remembering the passage, they were first asked to freely recall the Airport Problem and its solution for four minutes, followed by a 35-minute problem-solving test featuring the eight test problems. The participants who were asked to learn the procedure completed the same two tasks, but in the opposite order. They first completed the problem-solving section, followed by a free recall of the Airport Problem. While we were primarily interested in students' performance on the task that matched their learning goal, we also measured their performance on the mismatched task (i.e., problem solving for the remember passage group, and recall Airport Problem for the learn procedure group) to assess if the mismatched tasks would show similar patterns of learning strategy efficacies as the explicit goals.

After the final test, participants were asked for demographic information before being debriefed about the study.

Scoring. Although both recall and problem-solving performance were assessed for both groups, the outcomes matching the explicit learning goals were of greater interest. Specifically, the main dependent variable for the remember passage group was the proportion of idea units

recalled from the Airport Problem and its solution. In contrast, the main dependent variable for the learn procedure group was the proportion of problems solved correctly. As the dependent variables differed depending on the explicit learning goals, separate 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) Analysis of Variance (ANOVA) were conducted for each learning goal. Similar ANOVAs were conducted for the judgments of learning. For completeness, ANOVAs for the mismatched tasks (i.e., problem-solving performance for the remember passage group, and recall performance for the Airport Problem for the learn procedure group) are presented in the online supplement. To preview, there were no effects of learning strategy (repeated studying vs. repeated testing) on the mismatched tasks. In contrast, as detailed below, there were large differences in performance on the matched tasks depending on participants' learning strategy.

Participants' free-recall of the Airport Problem was scored by awarding one point for each correctly recalled idea unit. General idea units were identified by the four main ideas in the cover story (i.e., "arrival", "departure", "flights", and "airport"). Specific idea units provided further details about the worked example, and were defined as keywords (e.g., "expected on average", "two-hour period"), key numbers or number ranges (e.g., "4 (arrivals)", "3 (departures)", "10 or more"), and equations (e.g., "P (T = 10) + P(T = 11) + P(T = 12)"). Synonymous ideas (e.g., "mean" instead of "average", or "more than 9" instead of "10 or more") were awarded full credit. In some instances, half a point was awarded for a partially recalled idea unit (e.g., "3 arrivals and 4 departures" instead of "4 arrivals and 3 departures").

Responses for the problem-solving items were scored by awarding one point for each correctly solved problem, disregarding apparent computation errors. For example, "$\frac{e^{-10.}\,10^{5}}{5!} + \frac{e^{-10.}\,10^{6}}{6!} + \frac{e^{-10.}\,10^{7}}{7!}$ (correct solution) = .101 (incorrect answer)", and "2 × 3 = 8", which was carried over to subsequent steps that were otherwise accurate, were given full credit. No partial credit was given for the problem-solving items. In addition, problem-solving responses were coded according to error types using the scheme outlined in Table 1 that was adapted from Koedinger, Alibali, and Nathan (2008). Errors made in each of the four steps were coded independently such that the errors were not carried forward in subsequent steps (e.g., correct substitution of variables in the formula using incorrectly computed means was not coded as an error in formula application). Because each step was coded independently, an incorrect solution could have more than one type of error.

Twenty percent of the free-recall responses were independently scored by three raters, and interrater reliability was high (Fleiss' κ = .89). Twenty percent of the problem-solving responses were independently scored by two raters, and interrater reliability was also high (Cohen's κ = .90 for accuracy and κ = .83 for error type). Any discrepancies in coding were resolved through discussion. Given the high interrater agreement, the remaining test booklets were scored by one rater.

Table 1

Error codes and definitions for problem-solving items

Error Type	Definition
No attempt	Student leaves the problem blank, other than copying down information from the text
Answer only	Student writes an incorrect answer without showing any work
Incomplete	Student performs some work, but does not provide a final numerical answer, or an indication of a final equation with numerical values substituted
Conceptual/Procedural	
Steps 1 and 2 (means)	Student incorrectly finds the mean(s) of the event(s)
Step 3 (sum of means)	Student incorrectly finds the sum of the means
Step 4 (inequality)	Student incorrectly interprets the required inequality
Formula application	Student does not apply the formula correctly in itself and/or in an equation, such as multiplying instead of adding the probabilities, or incorrect value substitution
Technical	
Arithmetic	Student makes an apparent computational error, but solution is otherwise correct
Copy slip	Students possibly miscopies a value given in the problem, or from own work

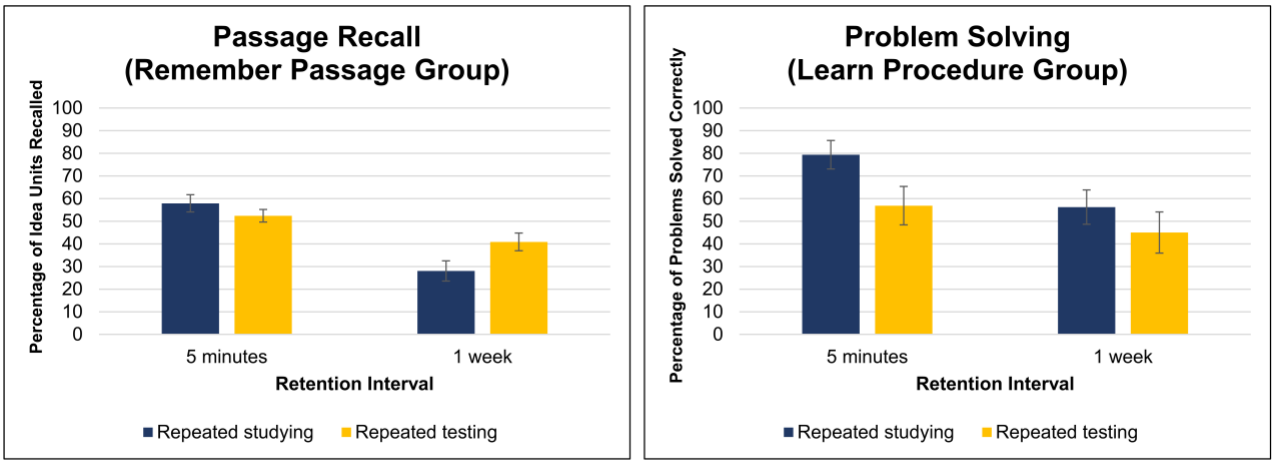
Results

Learning goal: Remember passage

Recall performance.

During learning. Participants in the 5-minute and 1-week retention groups were equally accurate in their initial recall. Participants in the 5-minute retention condition successfully recalled an average of 53.3% (*SD* = 13.1) of the idea units across the three learning trials, whereas those in the 1-week retention condition recalled 52.3% (*SD* = 15.2). This difference was not significant, $t < 1$, supporting our belief that there were no pre-existing differences between participants in the 5-minute and 1-week retention conditions.

Final test. We conducted a 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA on recall performance. As expected, participants recalled more after a delay of 5 minutes (*M* = 55.2%, *SD* = 14.9), than after 1 week (*M* = 34.5%, *SD* = 19.7), $F(1, 76) = 29.91$, $p < .001$, $\eta_p^2 = .282$. There was no overall difference in recall between repeated studying (*M* = 43.0%, *SD* = 23.7) and repeated testing (*M* = 46.6%, *SD* = 16.0), $F < 1$. However, as shown in Figure 2a, there was a significant interaction between learning strategy and retention interval, replicating the classic testing effect, $F(1, 76) = 5.83$, $p = .018$, $\eta_p^2 = .071$. Specifically, although retrieval practice (*M* = 52.4%, *SD* = 12.3) and repeated studying (*M* = 57.9%, *SD* = 17.0) did not differ after a 5-minute delay, $t(38) = 1.17$, $p = .251$, $d = .37$, one week later, retrieval practice (*M* = 40.9%, *SD* = 17.5) led to greater recall than repeated studying (*M* = 28.0%, *SD* = 20.0), $t(38) = 2.16$, $p = .038$, $d = .68$.



(a) (b)

Figure 2. Percentage of idea units recalled or problems solved correctly on the final test for the repeated studying and repeated testing groups after a 5-minute or 1-week delay for Experiment 1. (a) When the learning goal was remembering a passage, the repeated testing group outperformed the repeated studying group after a one week delay. (b) When the learning goal was learning a procedure, the repeated studying group outperformed the repeated testing group, regardless of retention interval. Error bars denote standard errors of the means.

Judgments of learning. A 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA revealed that individuals in the repeated studying condition (*M* = 4.53, *SD* = 1.36) predicted that they would be able to recall more details of the passage than those in the repeated testing condition (*M* = 3.58, *SD* = 1.47), $F(1, 76) =$

8.89, $p = .004$, $\eta_p^2 = .105$. There was no difference in judgments of learning between participants in the 5-minute ($M = 4.18$, $SD = 1.62$) and 1-week delay conditions ($M = 3.93$, $SD = 1.35$), $F < 1$, nor an interaction between learning strategy and retention interval, $F < 1$.

To directly examine if participants' judgments of learning were associated with their actual recall performance one week later, we examined the relation between the two variables for participants in the one-week delay condition (collapsed across learning strategies). Spearman's rank-order correlation (r_s) was used because the variables were not normally distributed. Final recall performance was not correlated with participants' judgments of learning, $r_s(38) = -.150$, $p = .355$. The participants were not very accurate in predicting how much they would remember.

Learning goal: Learn procedure

Problem-solving performance.

During learning. There was again no difference in initial test performance between the 5–minute and 1-week retention conditions. Participants in the 5-minute retention condition successfully solved 35.0% ($SD = 36.6$) of the problems across three learning trials, whereas those in the 1-week retention condition solved 28.3% ($SD = 32.9$), $t < 1$.

Final test. A 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA on problem-solving performance revealed that repeated studying of the worked examples led to higher problem-solving accuracy ($M = 67.8\%$, $SD = 32.9$) than repeated testing ($M = 50.9\%$, $SD = 39.3$), $F(1, 76) = 4.53$, $p = .037$, $\eta_p^2 = .056$ (Figure 2b). This finding is consistent with the classic worked example effect. Problem-solving performance was also higher in the 5-minute delay condition ($M = 68.1\%$, $SD = 34.9$) than in the 1-week delay condition ($M = 50.6\%$, $SD = 37.3$), $F(1, 76) = 4.87$, $p = .030$, $\eta_p^2 = .060$. There was no interaction between learning strategy and retention interval, $F < 1$.

Unexpectedly, we saw extremely low problem-solving performance during the learning phase (28 – 35%), but higher problem-solving performance during the final test (45 – 57%). Smith and Karpicke (2014) and van Gog and Kester (2012) both report similar counterintuitive learning and test performances for prose and problem-solving materials respectively.

Error analysis. Table 2 shows the frequency of the different error types across all the learn procedure participants and all problems. One clear result is a higher frequency of incomplete solutions during learning than on the final test. On the final test, the repeated testing conditions made more conceptual and procedural errors than the repeated studying conditions, particularly within the last two steps. This could be because they had poor memory for the procedure or because they failed to induce the logic of the procedure.

Table 2

Frequencies (percentage of trials) of errors in Experiment 1

	During Learning		Final Test			
	Repeated testing		Repeated studying		Repeated testing	
Error Type	5 minutes	1 week	5 minutes	1 week	5 minutes	1 week
No attempt	0	0	1.3	5.0	5.0	5.0
Answer only	0	1.7	0	0	0	0
Incomplete	33.3	31.7	5.6	3.8	2.5	5.6
Conceptual/Procedural						
Steps 1 and 2 (means)	8.3	1.7	5.3	10	9.7	8.8
Step 3 (sum of means)	11.7	8.3	2.5	3.8	4.4	6.3
Step 4 (inequality)	35.0	41.7	6.3	24.4	26.3	31.3
Formula application	25.0	31.7	0.6	1.9	13.8	21.9
Technical						
Arithmetic	3.3 (1.7)	3.3 (3.3)	0 (1.3)	0.6	0.6 (0.6)	0.6
Copy slip	0	0	1.9	0	1.3	1.3

Note. Percentages in parentheses refer to arithmetic errors made in solutions that were ultimately coded as correct.

Memory of procedure versus logic induction. To tease apart the above two alternatives, we examined if there was a higher-order interaction between the test problem types (i.e., isomorphic, or similar to the Airport Problem vs. transfer test problems) and learning strategy and/or retention interval. If participants in the repeated testing conditions performed poorly due to a lack of memory of the procedure, they should perform equally poorly on the isomorphic and transfer test problems. However, if they had remembered the procedure, but had not induced its underlying logic, we would expect them to perform better on the isomorphic problems than on the transfer problems. Note that this is an exploratory analysis and the experiment is not well-powered to detect a 3-way interaction.

A 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) × 2 (test problem type: isomorphic vs. transfer) mixed ANOVA on problem-solving performance revealed that isomorphic problem-solving performance (M = 68.8%, SD = 40.9) was higher than transfer problem-solving performance (M = 56.3%, SD = 38.4), $F(1, 76) = 14.86$, $p < .001$, $\eta_p^2 = .164$. Test problem type did not interact with either learning strategy ($F < 1$), or retention interval ($F < 1$), nor was there a three-way interaction, $F(1, 76) = 1.34$, $p = .251$, $\eta_p^2 = .017$. These findings suggest that inadequate memory of the procedure was unlikely to account for participants' poor performance in the repeated testing condition.

Judgments of learning. Similar to the remember passage group, individuals in the repeated studying condition (M = 4.88, SD = 1.22) predicted that they would be better able to solve similar problems than those in the repeated testing condition (M = 3.45, SD = 1.78), $F(1, 76) = 17.04$, $p < .001$, $\eta_p^2 = .183$. There was no difference in judgments of learning between participants in the 5-minute (M = 4.13, SD = 1.83) and 1-week delay conditions (M = 4.20, SD = 1.54), $F < 1$, and no interaction, $F < 1$.

Collapsed across the two 1-week retention conditions, final problem-solving performance was not correlated with participants' judgments of learning, $r_s(38) = .158$, $p = .330$. As with the recall test, participants were not very accurate in predicting how well they would do on the delayed problem-solving test.

Discussion

Experiment 1 demonstrates that the relative efficacies of repeated testing and repeated studying depend on the overall learning goal. Specifically, when students' goal was to remember the stable facts in the text of a worked example, those who engaged in retrieval practice recalled more idea units a week later than those who repeatedly studied the text. On the other hand, when students' goal was to learn a flexible problem-solving procedure, those who studied worked examples generally had better problem-solving performance than those who practiced solving the problems. Although the worked example effect was stronger on an immediate problem-solving test, a similar, but attenuated, effect was still observed a week later.

These findings demonstrate that complexity of materials does not seem to be a bane to the benefits of retrieval practice as van Gog and Sweller (2015) have suggested. Our materials were identical in both the learn procedure and remember passage conditions, and thus were equally complex. However, studying worked examples was more beneficial when participants were attempting to solve problems, and repeated testing was more beneficial when attempting to remember the specific facts within the problem. While our findings suggest that retrieval practice might not be as efficacious for learning problem-solving procedures, retrieval practice is still beneficial for recalling stable facts, even for materials with high element interactivity.

The absence of a testing effect when students' goal was to learn a flexible procedure could be due to poor performance during the learning phase. Error analyses revealed a high frequency of incomplete solutions in the repeated testing conditions during learning.

Finally, our findings reveal that students' judgments of learning were biased towards repeated studying, regardless of their learning goal. Such a bias is possibly due to the fluency of processing during repeated studying (Karpicke et al., 2009). This is not surprising given that students frequently report using repeated studying as a favorite learning strategy (Hartwig & Dunlosky, 2012; Karpicke et al., 2009; Kornell & Bjork, 2007). Even when students endorse retrieval practice as a study strategy, few are aware of the benefits that retrieval practice has over repeated studying on long-term retention (Kornell & Bjork, 2007; McCabe, 2011). There was also no association between judgments of learning and test performance, regardless of the learning goal. Interestingly, when learning a novel procedure, repeated testing did not make it more apparent to students if they would be able to solve the novel problems, as compared to repeated studying.

By and large, our finding that repeated testing was ineffective in improving learning for the learn procedure group is surprising given the prior research on the testing effect, but expected given the prior research on the worked example effect. Prior to concluding that testing is not as beneficial when the learning goal is to learn a novel procedure, we wanted to deal with two current methodological limitations – the variability of learning problems, and low performance during the learning phase.

Variability of learning problems

Besides having different learning goals, the remember passage group and the learn procedure group also differed in the variability of the problems learned. The remember passage group required the use of four identical passages (the Airport Problem), given that the repeated testing participants read one passage and then recalled it three times. For the learn procedure group, we chose to use four isomorphic problems with different cover stories (i.e., Airport Problem and three other problem contexts) to be consistent with the problem-solving literature, and also with real-world problem-solving learning. However, recent theories suggest that retrieval practice is most beneficial when it involves the reinstatement of prior context, which provides more distinct cues to restrict the memory probe space during the final test (Karpicke & Aue, 2015; Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014). The non-identical problems used during the learning phase may have prevented this contextual reinstatement (see van Gog et al., 2015, for a similar hypothesis). Alternatively, according to the KLI framework, retrieval practice may be especially beneficial for memory and fluency-building processes (Koedinger et al., 2012), and non-identical problems may not be useful for increasing the automatization of the procedure.

Inadequate learning

Any potential benefits of testing in the learn procedure group may have been diminished by inadequate learning during the first learning trial, a lack of corrective feedback in subsequent practice problems, or both (Karpicke et al., 2014; Rawson, 2015). Previous studies have observed that the benefits of testing depended on initial retrieval success, with lower retrieval success during learning being associated with poorer subsequent test performance (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Agarwal, & Roediger, 2009; see Rowland, 2014, for a meta-analysis; Smith & Karpicke, 2014).

Even though retrieval success during learning is critical for subsequent performance, it is not the only factor. The advantage of repeated testing over repeated studying can still be observed with low initial retrieval success. For instance, Smith and Karpicke (2014) compared different testing conditions (multiple-choice, short-answer, or a hybrid) with a repeated studying control condition, and found large and consistent advantages of repeated testing over repeated studying on inferential questions across two experiments even with initial retrieval successes between 26 and 37%. Moreover, in a meta-analysis by Rowland (2014), low initial retrieval success did not reliably lead to a testing effect, but neither did it lead to a reversed testing effect. In other words, low initial retrieval success during repeated testing has not been demonstrated to be reliably detrimental relative to repeated studying. Nonetheless, we wanted to improve performance during the learning phase for the repeated testing group.

Experiment 2

In order to examine if variable problems or low performance during the practice phase may have attenuated the testing effect in the learn procedure group, Experiment 2 focused exclusively on the learn procedure goal. We made two key changes: 1) all four learning problems featured the same cover story, but different numbers, and 2) the addition of a feedback condition to increase learning in the testing group. Three learning strategies were contrasted – repeated studying of identical worked examples ($S_1S_1S_1S_1$), repeated testing of identical problems without feedback ($S_1T_1T_1T_1$), and repeated testing of identical problems with feedback ($S_1T_{1F}T_{1F}T_{1F}$).

If reinstatement of the episodic context is indeed a key mechanism underlying the testing effect, or if fluency building is enhanced by identical problems, then repeated testing with identical problems should lead to higher problem-solving performance than repeated studying of identical worked examples, especially after one week. Alternatively, testing may not be as useful for learning flexible procedures, and repeated studying would again to lead to higher problem-solving performance than repeated testing, similar to the results observed in Experiment 1. Additionally, feedback should boost problem-solving performance over and above the benefit of repeated testing, such that the repeated testing with feedback condition should yield the highest problem-solving success.

Method

Participants. One hundred and twenty adults from a highly selective university's human subject pool (36 males; M_{age} = 19.6 years, SD = 1.70, range: 18 – 25) participated in exchange for course credit or monetary incentive ($5 per half-hour of participation). As in Experiment 1, participants were randomly assigned to one of the three learning strategy conditions within each retention interval. There was no difference in prior knowledge across all conditions (see online supplement).

Design. A 3 (learning strategy: repeated studying, SSSS vs. repeated testing without feedback, STTT vs. repeated testing with feedback, $ST_FT_FT_F$) × 2 (retention interval: 5 minutes vs. 1 week) between-subjects design was used, resulting in 20 participants per cell. The dependent variable was the number of problems (out of eight) solved correctly.

Materials. The first worked example (S_1) was identical to that used in Experiment 1 – the Airport Problem. The subsequent three problems or worked examples had the same cover story, but different numerical values. We altered the numbers in the problem to ensure that participants actually retrieved the procedure and solved the problem, rather than simply remembering the solution.

The problem-solving test was identical to the one used in Experiment 1, except for a small change to one of the eight problems that prevented students from skipping the first two steps of the procedure. The solutions to both versions are nearly identical.

Procedure. The procedure was generally similar to that for the learn procedure group in Experiment 1, except for the duration of the learning trials and the addition of the repeated testing with feedback condition. Participants in the repeated testing with feedback condition solved a problem, and were then presented with the worked solution. Participants in this feedback condition had four minutes to solve the problem and two minutes to review the worked example. To match the overall presentation time in the retrieval practice with feedback condition, the other two conditions were given six minutes on each learning trial. Participants still completed a one-minute filler task between each problem. After the learning trials, participants again made judgements of learning for how well they would be able to solve the problems one week later.

After either the 5-minute or 1-week delay, participants completed the final problem-solving test. The duration of the test was reduced from 35 minutes in Experiment 1 to 30 minutes in Experiment 2. Thirty minutes was more than adequate for most, if not all, participants. Finally, the brief questionnaire that assessed how familiar the participant was with this particular type of probability problem and their relevant prior knowledge was administered after the final test instead of at the end of the learning phase.

Scoring. The scoring scheme and procedures were identical to that described in Experiment 1. Twenty percent of the problem-solving booklets were independently scored by two raters, and interrater reliability was again high (Cohen's κ = .97 for accuracy and κ = .90 for error type).

Results

Problem-solving performance.

During learning. A 2 (learning strategy: repeated testing without feedback vs. repeated testing with feedback) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA was conducted on problem-solving performance during the learning phase. There were no pre-existing differences between participants in the 5-minute (M = 83.3%, SD = 31.1) and 1-week retention conditions (M = 90.0%, SD = 25.3), $F(1, 76) = 1.09$, $p = .301$, $\eta_p^2 = .014$. In addition, there was no difference in the proportion of problems solved successfully across the three learning trials for participants who did (M = 88.3%, SD = 26.7) and did not receive feedback (M = 85.0%, SD = 30.1), $F < 1$. Participants were able to successfully solve the identical problems, even without feedback. There was also no interaction between learning strategy and retention interval, $F < 1$.

Final test. A 3 (learning strategy: repeated studying vs. repeated testing without feedback vs. repeated testing with feedback) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA revealed no overall difference in problem-solving performance among repeated studying (M = 57.8%, SD = 34.5), repeated testing without feedback (M = 64.1%, SD = 30.6), and repeated testing with feedback (M = 71.6%, SD = 28.3), $F(2, 114) = 2.14$, $p = .123$, $\eta_p^2 = .036$. Problem-solving performance also did not differ between the 5-minute (M = 68.1%, SD = 27.9) and 1-week delays (M = 60.8%, SD = 34.6), $F(1, 114) = 1.80$, $p = .183$, $\eta_p^2 = .016$. However, there was an interaction between learning strategy and retention interval, $F(2, 114) = 6.50$, $p = .002$, $\eta_p^2 = .102$ (Figure 3).

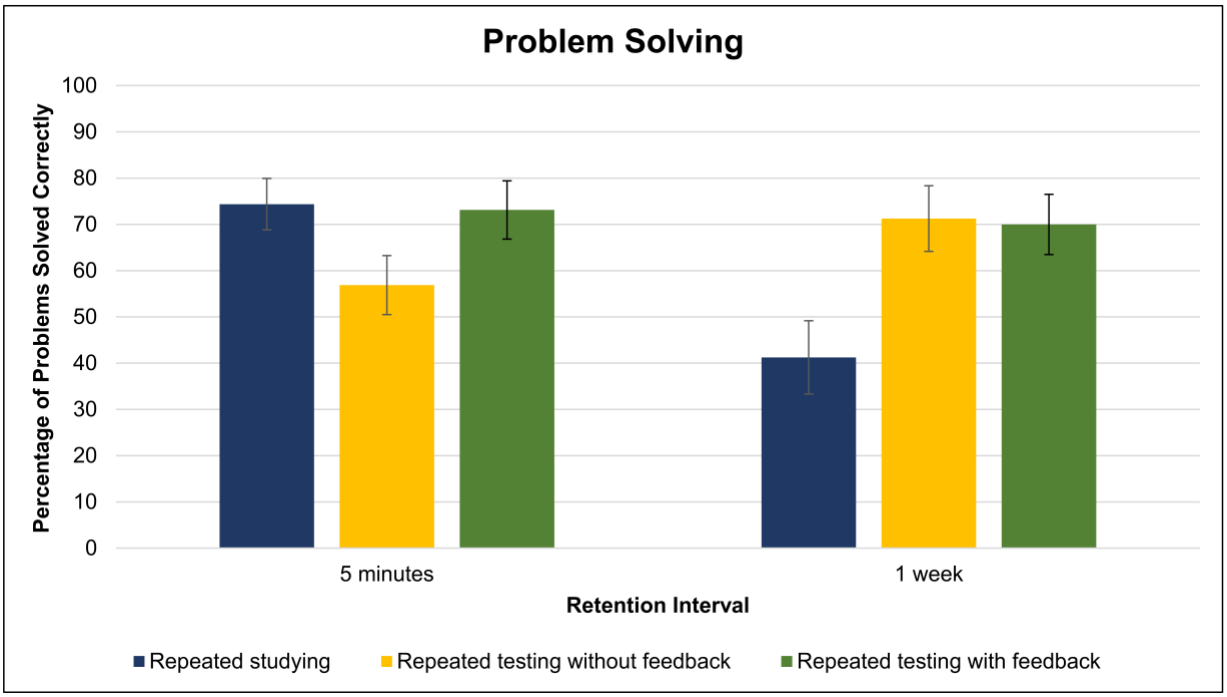


Figure 3. Percentage of problems solved correctly on the final test for the repeated studying, repeated testing without feedback, and repeated testing with feedback groups, after a 5-minute or 1-week delay for Experiment 2. Both repeated testing groups outperformed the repeated studying group after a one week delay. Error bars denote standard errors of the means.

To clarify the interaction, a one-way (learning strategy: repeated studying vs. repeated testing without feedback vs. repeated testing with feedback) ANOVA was conducted separately for the 5-minute and 1-week delay conditions. With a 5-minute delay, there was no difference between repeated studying (M = 74.4%, SD = 24.8), repeated testing without feedback (M = 56.9%, SD = 28.5), and repeated testing with feedback (M = 73.1%, SD = 28.2), $F(2, 57) = 2.57$, $p = .085$, $\eta_p^2 = .083$. However, after a 1-week delay, a main effect of learning strategy was observed (repeated studying: M = 41.3%, SD = 35.4; repeated testing without feedback: M = 71.3%, SD = 31.7; repeated testing with feedback: M = 70.0%, SD = 29.1), $F(2, 57) = 5.57$, $p = .006$, $\eta_p^2 = .164$. Post-hoc comparisons using Tukey's HSD test revealed that individuals in the repeated studying condition solved fewer problems correctly than those in the repeated testing conditions without feedback, $t(38) = 2.82$, $p = .013$, $d = .89$, and with feedback, $t(38) = 2.81$, $p = .018$, $d = .89$. There was no difference in the proportion of problems solved successfully between the two repeated testing conditions, $t < 1$.

Within the repeated studying group, participants solved more problems correctly after a 5-minute delay (M = 74.6%, SD = 24.8) than after a 1-week delay (M = 42.1%, SD = 34.8), $t(34.4) = 3.40$, $p = .002$, $d = 1.08$. However, there was no difference in problem-solving performance across the two delays (5-minute: M = 57.8%, SD = 26.7, 1-week: M = 71.5%, SD = 31.6) within the repeated testing group without feedback, $t(38) = 1.48$, $p = .147$, $d = .47$, and within the repeated testing group with feedback (5-minute: M = 73.4%, SD = 28.3, 1-week: M = 70.2%, SD = 29.0), $t < 1$.

Error analysis. In contrast with Experiment 1, there were substantially fewer incomplete solutions during learning (0 – 3%, see Table S4 in online supplement). The most conceptual and

procedural errors were seen within the repeated studying with a 1-week delay condition. All other conditions were comparable in the distribution and frequencies of errors made.

Memory of procedure versus logic induction. As with Experiment 1, we conducted an exploratory analysis to examine if there was a higher-order interaction with the test problem types. A 3 (learning strategy: repeated studying vs. repeated testing without feedback vs. repeated testing with feedback) × 2 (retention interval: 5 minutes vs. 1 week) × 2 (test problem type: isomorphic vs. transfer) mixed ANOVA on problem-solving performance revealed that isomorphic problem-solving performance (M = 71.7%, SD = 37.1) was higher than transfer problem-solving performance (M = 62.1%, SD = 33.3), $F(1, 114) = 11.28$, $p = .001$, $\eta_p^2 = .090$. Problem type did not interact with learning strategy, $F < 1$, or retention interval, $F(1, 114) = 1.73$, $p = .191$, $\eta_p^2 = .015$, and there was no three-way interaction, $F < 1$. However, the experiment is not well-powered to detect a 3-way interaction.

Judgments of learning.

A 3 (learning strategy: repeated studying vs. repeated testing without feedback vs. repeated testing with feedback) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA revealed no difference in participants' predictions of how well they would be able to solve similar problems among the three learning strategies (repeated studying: M = 6.03, SD = 1.03; repeated testing without feedback: M = 5.55, SD = 1.28; repeated testing with feedback: M = 5.65, SD = 1.31), $F(2, 114) = 1.73$, $p = .182$, $\eta_p^2 = .029$. There was also no difference between the 5-minute (M = 5.87, SD = 1.16) and 1-week delays (M = 5.62, SD = 1.28), $F(1, 114) = 1.29$, $p = .258$, $\eta_p^2 = .011$, nor an interaction, $F(2, 114) = 1.57$, $p = .213$, $\eta_p^2 = .027$.

Collapsed across the three 1-week retention conditions, problem-solving performance was not correlated with participants' judgments of learning, $r_s(58) = .038$, $p = .770$.

Experiment 1 vs. Experiment 2.

To clarify if the contrasting results between both experiments were due to the minor changes in duration of learning trials and duration of final test introduced in Experiment 2, we conducted a 2 (Experiment: 1 vs. 2) × 2 (learning strategy: repeated studying vs. repeated testing without feedback) ANOVA on the overall problem-solving performance separately for the 5-minute and 1-week retention intervals. If those minor methodological changes mattered, we should expect to observe a global impact on problem-solving performance in all conditions, regardless of retention interval. However, if they did not matter, and the key change was instead the use of identical problems, the impact on performance should be specific to the 1-week retention group.

5-minute retention group. Repeated studying (M = 76.9%, SD = 26.3) yielded higher performance than repeated testing (M = 56.9%, SD = 33.1), across both experiments, $F(1, 76) = 8.74$, $p = .004$, $\eta_p^2 = .103$. There were no differences in overall accuracy across the two experiments, and no interaction between experiment and learning strategy, Fs < 1.

1-week retention group. An advantage of repeated testing was observed in Experiment 2, but not in Experiment 1, verified by an interaction between experiment and learning strategy, $F(1, 76) = 6.74$, $p = .011$, $\eta_p^2 = .081$ (see Figure 4). Specifically, while problem-solving performance did not differ between the repeated testing and repeated studying conditions in Experiment 1 after a 1-week delay, repeated testing was more effective than repeated studying in Experiment 2. There were no differences in overall performance between the two experiments, F

< 1, or between the repeated studying and repeated testing conditions, $F(1, 76) = 1.39$, $p = .242$, $\eta_p^2 = .018$.

Discussion

Experiment 2 aimed to investigate if a testing effect could be observed for learning flexible procedures when participants had higher accuracy during the learning phase, and when the learning problems shared an identical cover story. While we failed to find the testing effect after a 1-week delay when non-identical learning problems were used in Experiment 1, the testing effect was observed when identical learning problems were used in Experiment 2. For these identical problems, feedback did not confer any benefits on learning beyond what repeated testing alone affords, especially after a 1-week delay. Hence, when retrieval success was high, feedback was redundant. Taken together, our findings suggest that a testing effect can be observed with materials with high element interactivity, such as learning flexible procedures.

Importantly, problem-solving accuracy during learning was more than twice as high in Experiment 2 as compared to Experiment 1. If inadequate learning had accounted for the findings of Experiment 1, we would expect to see a global boost in problem-solving performance in all conditions in Experiment 2, regardless of retention interval. However, despite the higher performance during the learning trials, accuracy on the final test was similar across the two experiments after a 5-minute delay (see Figure 4). The differences across the two experiments were only seen after the 1-week delay. This finding speaks strongly against the inadequate learning hypothesis.

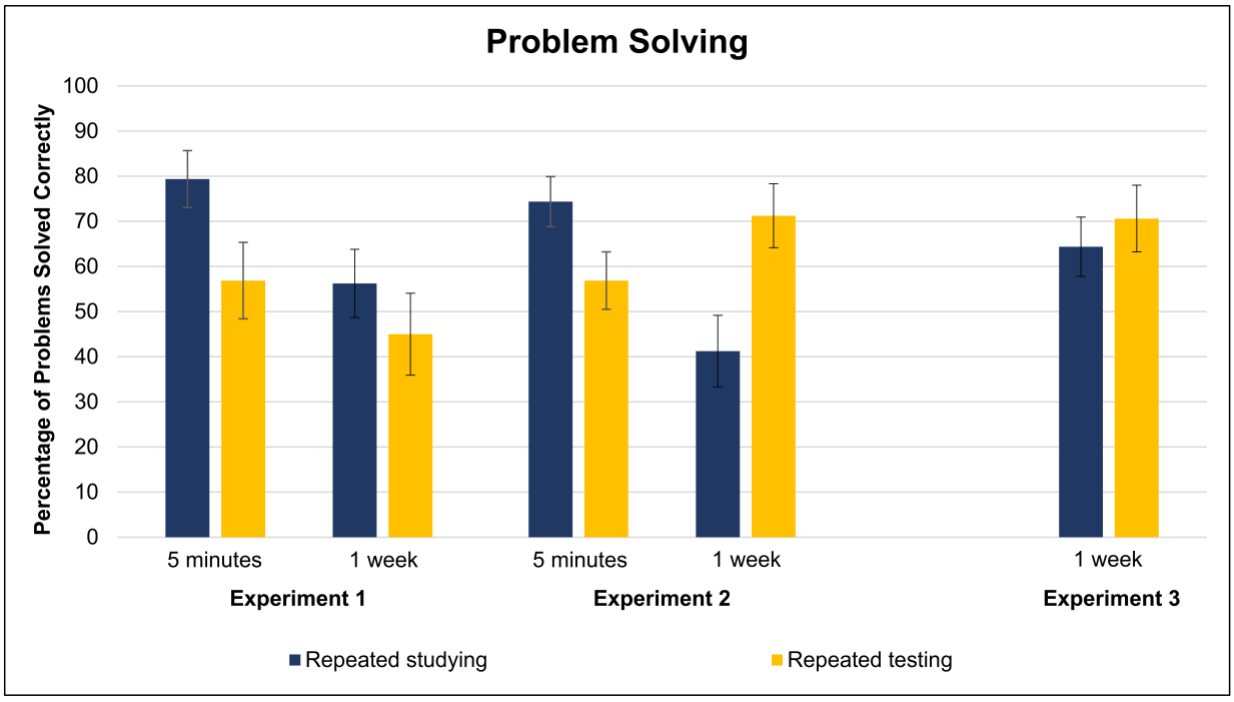


Figure 4. Percentage of problems solved correctly on the final test for the repeated studying and repeated testing without feedback groups after a 5-minute or 1-week delay across Experiments 1, 2, and 3. The performance patterns of the repeated studying and repeated testing groups were identical in the 5-minute delay condition across Experiments 1 and 2. In the 1-week delay

condition, the testing effect was observed when identical problems were used in Experiment 2, but not in Experiments 1 and 3. Error bars denote standard errors of the means.

Thus, the change from non-identical to identical learning problems likely accounts for the existence of a testing effect in Experiment 2. This is consistent with a previous study that found benefits of retrieval practice with identical learning problems on a delayed test. Darabi and colleagues (2007) had students learn to diagnose and repair malfunctions in a simulated water-alcohol distillation plant, and then either studied four descriptive worked examples, or completed four problem-solving practice trials – all within the same distillation plant context. Problem-solving practice within a constant context was found to be more effective for learning than studying worked examples within the same constant context (Darabi et al., 2007).

The identical learning problems could have enhanced memory and fluency-building processes by providing opportunities for the reinstatement of the Airport Problem context. Consistent with the episodic context hypothesis (Karpicke et al., 2014), the reinstatement opportunities may have strengthened the specificity of the contextual cues associated with the procedure, thereby narrowing the search space, and increased the chances of activating the correct target response. This is akin to teachers' reminders for students to recall how they solved a particular problem previously during problem-solving practice (Reeves & Weisberg, 1994; Ross & Kennedy, 1990). This context reinstatement may not have been offered by the non-identical problems in Experiment 1. It is possible that participants in Experiment 1 were not spontaneously reinstating the context of the Airport Problem and reasoning analogically. Experiment 3 was designed to test this hypothesis with variable learning problems.

Finally, with regards to students' judgments of their learning of the procedure, learning strategy did not consistently moderate students' judgments of learning and problem-solving performance across both experiments. Given the contrasting efficacies of the learning strategies across both experiments, it is not surprising to find incongruent findings for judgments of learning. It is, however, clear from the lack of association between judgments of learning and test performance that students' judgments of learning were often inaccurate, even when the goal was to learn procedures.

Experiment 3

To test whether episodic reinstatement of contextual cues underlay the testing effect found in Experiment 2, but not in Experiment 1, we performed a partial replication of Experiment 1 (with variable learning problems – $S_1S_2S_3S_4$ or $S_1T_2T_3T_4$). In addition to the repeated studying and repeated testing conditions from Experiment 1, we included a repeated testing condition that instructed participants to engage in episodic recall of the Airport Problem during the learning trials. To further resolve the issue between inadequate learning and incomplete solutions that led to the exceptionally low problem-solving performance during learning in Experiment 1, we also modified the reading and computational demands of the learning problems such that they would yield more complete solutions (see Materials section and Table 3 for details). Finally, we focused on the 1-week retention interval since that was where the results of the previous two experiments differed.

If episodic reinstatement of contextual cues was key to obtaining the testing effect, we would expect instructions to recall the Airport Problem when solving variable learning problems to yield higher problem-solving performance one week later than solving without explicit episodic recall instructions and repeated studying.

Table 3
Contrast of an example learning problem used in Experiment 1 and its modified version used in Experiment 3

Experiment 1	Experiment 3
In a restaurant, large number of cups and saucers are washed each day. The number of cups that are broken each day while washing averages 2.1, whereas the number of saucers broken each day averages 1.6, independently of the number of cups broken. Suppose that the number of broken cups and saucers follow two independent Poisson distributions. Find the probability that the total number of cups broken and saucers broken during a randomly chosen week of 7 days is at least 22 but no more than 26. **Step 1:** Let C be the number of cups broken in a week. 1 day $\rightarrow$ 2.1 cups 7 days $\rightarrow 2.1 \times 7 = 14.7$ cups So, C ~ $P_o(14.7)$. **Step 2:** Let S be the number of saucers broken in a week. 1 day $\rightarrow$ 1.6 saucers 7 days $\rightarrow 1.6 \times 7 = 11.2$ saucers So, S ~ $P_o(11.2)$. **Step 3:** Let T be the total number of cups and saucers broken in a week, i.e., C + S. T ~ $P_o(14.7 + 11.2)$ i.e., T ~ $P_o(25.9)$. **Step 4:** $P(22 \leq T \leq 26) = P(T = 22) + P(T = 23) + P(T = 24) + P(T = 25) + P(T = 26)$ $= \frac{e^{-25.9} \cdot 25.9^{22}}{22!} + \frac{e^{-25.9} \cdot 25.9^{23}}{23!} + \frac{e^{-25.9} \cdot 25.9^{24}}{24!} + \frac{e^{-25.9} \cdot 25.9^{25}}{25!} + \frac{e^{-25.9} \cdot 25.9^{26}}{26!} = .364$	Suppose that, in a restaurant, the number of cups and saucers being broken each day while washing follow two independent Poisson distributions. In one day, it is expected on average that there are 2 cups and 1 saucer broken. Find the probability that, in a randomly selected seven-day period, the total number of cups broken and saucers broken is at least 18 but no more than 20. **Step 1:** Let C be the number of cups broken in a seven-day period. 1 day $\rightarrow$ 2 cups 7 days $\rightarrow 2 \times 7 = 14$ cups So, C ~ $P_o(14)$. **Step 2:** Let S be the number of saucers broken in a seven-day period. 1 day $\rightarrow$ 1 saucer 7 days $\rightarrow 1 \times 7 = 7$ saucers So, S ~ $P_o(7)$. **Step 3:** Let T be the total number of cups and saucers broken in a seven-day period, i.e., C + S. T ~ $P_o(14 + 7)$ i.e., T ~ $P_o(21)$. **Step 4:** $P(18 \leq T \leq 20) = P(T = 18) + P(T = 19) + P(T = 20)$ $= \frac{e^{-21} \cdot 21^{18}}{18!} + \frac{e^{-21} \cdot 21^{19}}{19!} + \frac{e^{-21} \cdot 21^{20}}{20!} = .244$

Method

Participants. Sixty adults from the same university's human subject pool (14 males; M_{age} = 19.9 years, *SD* = 1.36, range: 18 – 23) participated in exchange for course credit or monetary incentive ($5 per half-hour of participation). Participants were randomly assigned to one of the three conditions. There was no difference in prior knowledge across all conditions (see online supplement).

Design. A single-factor between-subjects design (learning strategy: repeated studying, SSSS vs. repeated testing, STTT, without recall instruction, vs. repeated testing, STTT, with recall instructions) was used, with 20 participants per condition. The dependent variable was the number of problems (out of eight) solved correctly.

Materials. The learning problems used were modified versions of those used in Experiment 1. Keeping the scenarios identical, we altered the phrasing of the problems so that they were less lengthy on average, and more comparable to the phrasing, number of sentences, and words in the Airport Problem (see Table 3). The average number of characters in the worked examples was 806 in Experiment 1, and 740 in Experiment 3. In order to yield fewer incomplete solutions, we also modified the values and the required inequalities such that they were less computationally intensive (e.g., whole numbers instead of decimals). A comparison of an example problem used in Experiment 1 and its modified version used in Experiment 3 can be found in Table 3. The problem-solving test was identical to the one used in Experiment 2.

Procedure. The procedure was similar to Experiment 2. Each learning trial was six-minutes long, and all participants completed the problem-solving test one week later. No feedback was given. The only change was the addition of a repeated testing with recall instructions group. Before each learning problem, participants in this group were told "While the problem context is different from the previous one, the procedure to be used here is exactly the same as that used to solve the first problem (i.e., the 'Airport' problem). Many people find it helpful to recall how the Airport Problem was solved when trying to solve this new problem."

Scoring. The scoring scheme and procedures were identical to that described in Experiment 1. Twenty percent of the problem-solving booklets were independently scored by two raters, and interrater reliability was again high (Cohen's κ = .96 for accuracy and κ = .87 for error type).

Results

Problem-solving performance.

During learning. An independent-samples *t*-test was conducted on the problem-solving performance between repeated testing with and without recall instruction. There was no difference in accuracy between participants who were instructed to recall the Airport Problem to help them solve a new problem (*M* = 73.3%, *SD* = 41.3) than those who did not receive the episodic recall instructions (*M* = 75.0%, *SD* = 37.3), $t < 1$.

Final test. A one-way (learning strategy: repeated studying vs. repeated testing without recall instructions vs. repeated testing with recall instructions) ANOVA revealed no overall difference in problem-solving performance among repeated studying (*M* = 64.4%, *SD* = 29.3), repeated testing without recall instructions (*M* = 70.6%, *SD* = 33.0), and repeated testing with recall instructions (*M* = 73.8%, *SD* = 36.5), $F < 1$.

Error analysis. As with Experiment 2, there were substantially fewer incomplete solutions during learning compared to Experiment 1 (0%, see Table S5 in online supplement).

The distribution and frequencies of errors made during the final test were relatively similar across conditions.

Memory of procedure versus logic induction. As with the previous experiments, we also conducted an exploratory analysis to examine if there was an interaction between the test problem types and learning strategy. In contrast with previous experiments, a 3 (learning strategy: repeated studying vs. repeated testing without recall instructions vs. repeated testing with recall instructions) × 2 (test problem type: isomorphic vs. transfer) mixed ANOVA on problem-solving performance revealed no difference between isomorphic problem-solving performance (M = 74.2%, SD = 37.4) and transfer problem-solving performance (M = 68.1%, SD = 34.3), $F(1, 57) = 2.53$, $p = .117$, $\eta_p^2 = .043$. There was also no interaction between test problem type and learning strategy, $F < 1$.

Judgments of learning.

A one-way (learning strategy: repeated studying vs. repeated testing without recall instructions vs. repeated testing with recall instructions) ANOVA revealed a significant difference in participants' predictions of how well they would be able to solve similar problems among the three conditions (repeated studying: M = 5.75, SD = 1.37; repeated testing without episodic recall: M = 5.15, SD = 1.14; repeated testing with episodic recall: M = 4.55, SD = 1.50), $F(2, 57) = 3.98$, $p = .024$, $\eta_p^2 = .122$. Post-hoc comparisons using Tukey's HSD test revealed that individuals in the repeated testing with recall instructions had lower judgments of learning than those in the repeated studying condition, $t(38) = 2.64$, $p = .012$, $d = 0.83$, but similar judgments of learning as those in the repeated testing without recall instructions, $t(38) = 1.42$, $p = .163$, $d = 0.45$. There was no difference in judgments of learning between the repeated studying and repeating testing without recall instructions, $t(38) = 1.51$, $p = .140$, $d = 0.48$.

Collapsed across the three conditions, problem-solving performance was not correlated with participants' judgments of learning, $r_s(58) = .227$, $p = .081$.

Discussion

Experiment 3 examined whether episodic reinstatement of the initial problem underlay the testing effect observed in Experiment 2. To this end, we used variable learning problems similar to those in Experiment 1, and included a condition that encouraged episodic recall of the Airport Problem when they solved subsequent practice problems with different cover stories. While we observed a numerical advantage of such episodic recall with repeated testing, it was not statistically different from repeated testing without episodic recall or from repeated studying. One possible explanation for the lack of benefit of recall instructions is that the instructions were ineffective in reinstating the contextual cues from the Airport Problem. Alternatively, students that did not receive the recall instructions could be spontaneously reinstating the contextual cues from the Airport Problem anyway, especially given the high surface correspondence between the target and source analogs (see Table 3 and Figure 1; Reeves & Weisberg, 1994).

Interestingly, Experiment 3 saw an overall boost in both learning and subsequent test performances compared to Experiment 1, but did not significantly modify the relative efficacies of repeated studying and repeated testing after a 1-week delay (Figure 4). Even with the increase in retrieval success during learning in Experiment 3, a testing effect was still not observed. In fact, our findings mirror those of van Gog and colleagues in that low problem-solving success during learning (< 50%) leads to a worked example effect (van Gog & Kester, 2012), but high

problem-solving success during learning (> 50%) leads to equivalent efficacies between repeated testing and repeated studying (van Gog et al., 2015).

With comparably high performance during the learning phase in Experiments 2 and 3, we found that repeated testing was as effective as repeated studying when variable learning problems were used, but more effective than repeated studying when identical learning problems were used. In addition, problem-solving performance on the transfer test problems was lower than that on the isomorphic test problems when identical learning problems were used, but comparable when variable learning problems were used. A possible explanation is that variable learning problems can facilitate the induction of the underlying logic of the procedure such that they can be applied to novel and structurally dissimilar problems. While variable learning problems were also used in Experiment 1, those used in Experiment 3 were more analogous in surface structure to the Airport Problem than in Experiment 1 and may have been more conductive to induction (see Table 3 and Figure 1 for comparison). This is consistent with Catrambone and Holyoak's (1989) finding that schema induction and transfer were better supported when target learning problems were more analogous in their surface structure to the source problem (see Reeves & Weisberg, 1994, for a review). Taken together, these suggest that problem-solving practice is no less effective than studying worked examples, and that variable learning problems may better support inductive processes, especially when surface correspondence between the target and source problems is maximized.

Finally, for judgments of learning, the findings are largely similar to those in Experiment 2, suggesting that the addition of feedback or episodic recall did not increase one's judgement of learning relative to the other learning conditions, as well as their accuracy relative to the actual performance one week later.

General Discussion

The knowledge-learning-instruction (KLI) framework (Koedinger et al., 2012) proposed that the relative efficacies of retrieval practice and repeated studying lie in the kind of knowledge being learned. By demonstrating a testing advantage when one's goal was to learn stable facts in a passage, and a worked example advantage when one's goal was to learn flexible procedures in Experiment 1, the current study provides empirical support for the KLI framework in clarifying the role of overall learning goals in the relative efficacies of retrieval practice and worked examples. Moreover, by keeping the overall learning goal the same, but altering the nature of the learning problems used in Experiments 2 and 3, we demonstrated that the testing advantage is possible when one's goal was to learn flexible procedures. Our findings are therefore also consistent with the reconceptualized cognitive load theory (Kalyuga & Singh, 2016; Likourezos & Kalyuga, 2017) in that the *same* instructional task (i.e., problem-solving practice) can target distinct sub-goals – in this case, with *different* materials (i.e., identical versus variable learning problems), leading to different relative efficacies of repeated testing and worked examples. Hence, the current study clarifies and extends the specifics of both the KLI framework and the reconceptualized cognitive load theory, and generates new hypotheses to be tested.

While much research has been devoted to investigating instructional events and assessment events, the learning processes and knowledge components often reside in a black box that is unobservable. As the KLI framework proposed, the effectiveness of an instructional recommendation depends on the fit between the knowledge components and the learning processes in a conceptual black box. On one hand, the acquisition of stable knowledge components with constant application condition and constant response during recall-based

assessment events (e.g., Area of a circle, $A = \pi r^2$) is thought to be supported primarily by memory and fluency-building processes, for which retrieval practice is most appropriate, especially when long-term retention is concerned. On the other hand, the acquisition of flexible knowledge components with variable application conditions and variable responses during transfer-based problem-solving tests (e.g., $\frac{1}{5} + \frac{3}{5} = \frac{4}{5}$, but $\frac{1}{5} + \frac{3}{10} = \frac{2}{10} + \frac{3}{10} = \frac{5}{10}$), is thought to be supported primarily by induction or compilation processes. Besides the importance of the fit between knowledge components and learning processes, we propose that instructional recommendations should be qualified by the retention interval and nature of learning materials used. Inductive processes are not the sole learning event during problem-solving instruction. All types of long-term knowledge retention involve some degree of memory processes, and should be supported in any instructional strategy (Koedinger et al., 2012). The dominance of one learning process over another may be related to the retention interval and the nature of the learning materials, such as the variability of problems used. In turn, the dominant learning process may determine the optimal learning strategy.

Retention interval and variability of learning materials as moderators

Our findings suggest that long-term retention of flexible knowledge components does not rely only on induction or compilation processes, but also on memory processes (Reeves & Weisberg, 1994; Renkl, 2014). However, memory processes are likely to be less critical than inductive processes when students are assessed immediately, whereas memory processes may be more critical when students are assessed after a substantial delay. This is supported by (1) the consistent worked example effects in the 5-minute delay conditions in Experiments 1 and 2, and (2) that worked examples were no more beneficial and were sometimes detrimental when students were tested one week later across the three experiments. In addition, the lower relevance of memory processes on an immediate than on a delayed problem-solving assessment may account for the fact that retrieval success during learning did not matter after a 5-minute delay, but did after a 1-week delay.

This retention interval hypothesis is consistent with the fact that the benefits of worked examples have more consistently been found with immediate problem-solving tests, compared with delayed tests (e.g., Leahy, Hanham, & Sweller, 2015; van Gog et al., 2015; van Gog, Paas, & Van Merriënboer, 2006; van Gog & Kester, 2012). It is also important to note that while the testing effect can be observed even in the short-term (less than one day), the magnitude of the effect tends to be larger for longer retention intervals (one day or longer) (see Adesope et al., 2017; Rowland, 2014, for meta-analyses).

After a 1-week delay, students may forget the procedure, its underlying logic, or both. While worked examples are useful for inducing the underlying logic of the procedure, students are likely to forget the induced logic one week later. This can explain why in Experiment 2, worked examples were more effective on an immediate test and testing was more effective on a delayed test. However, the extent of forgetting may be attenuated either by studying variable learning problems (Experiments 1 and 3), or with retrieval practice (Experiments 2 and 3). Variable learning problems may support more durable induction of the underlying logic of the procedure. Past studies have shown similar benefits of variable worked examples (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Paas & Van Merriënboer, 1994; Quilici & Mayer, 1996; Reed & Bolstad, 1991; van Gog et al., 2015), although some did not (e.g., Renkl, Stark, Gruber, & Mandl, 1998). It has been proposed that presenting two or more non-identical worked examples support transfer by allowing students to link analogous solutions to problems (Cooper

& Sweller, 1987; Gick & Holyoak, 1983).Retrieval practice likely reduced forgetting such that students could retrieve traces of either the procedure and/or its logic to solve the problems successfully.

Judgments of procedural learning

Across the three experiments, we did not find a consistent moderation of judgments of learning and problem-solving performance by learning strategy. Students' judgments of learning also did not correlate with their actual performance one week later. Generally, students' judgments tended to be biased toward repeated studying regardless of the learning goal, possibly due to the fluency of processing (Karpicke et al., 2009), or an illusion of understanding during repeated studying (Renkl, 2002). As observed in Experiments 2 and 3, this bias could be attenuated when one's goal was to learn a procedure, as repeated testing may possibly provide students with more concrete clues regarding their ability to solve the problems successfully after a week's delay. Even adults in a highly-selective college were often inaccurate with their judgments of learning of problem-solving procedures. Future studies should therefore examine how students' metacognition can be improved during procedural learning.

Educational implications

Mathematical and science problem solving involves a myriad of knowledge components, including concepts (e.g., proportions, atomic structure), definitions (e.g., irrational numbers, gravity), formulas (e.g., area of a circle, relation between energy and mass), theorems and laws (e.g., Pythagoras' Theorem, Newton's laws of motion), and procedures (e.g., performing a *t*-test, deducing stoichiometric relations). These knowledge components may be constant-constant, variable-constant, or variable-variable mappings, and thus require different learning processes (Koedinger et al., 2012). Furthermore, the more complex the learning material is, the more likely it entails multiple sub-goals comprising different knowledge components and their corresponding learning processes (Kalyuga & Singh, 2016). So, teachers and students need to be aware of them, and be flexible with their instructional and learning strategies respectively, within and across domains.

Some goals relate to flexible knowledge components and induction processes (e.g., schema acquisition of a procedure), for which studying worked examples may be more effective than retrieval practice. Other goals pertain to stable knowledge components, and memory and fluency-building processes (e.g., memorizing definitions, or a fixed sequence of steps in a procedure), for which retrieval practice is more effective than restudying. The use of identical learning problems versus variable learning problems is an example of how teachers can enhance the intended learning process or sub-goal.

Teachers and students should also consider the intended retention period. For short-term retention, memory and fluency-building processes may be less crucial than schema induction, hence problem-solving practice may be suboptimal than studying worked examples. For long-term retention, both schema induction, and memory and fluency-building processes may play a role in mitigating forgetting, hence it may be more effective for students to incorporate problem-solving practice during learning, be it pure retrieval practice or interleaved with worked examples.

Conclusion

When students' goal was to remember the text of a worked example, repeated testing resulted in higher recall performance than repeated studying one week later. However, when students' goal was to learn a novel math procedure, the optimal learning strategy depended on the learning processes or sub-goals associated with the retention interval and the nature of the materials. When long-term retention was not crucial (i.e., on an immediate test), repeated studying was more optimal than repeated testing, regardless of the nature of materials. When long-term retention was crucial (i.e., on a one-week delayed test), repeated testing was as effective as, if not more effective than, repeated studying. Hence, our findings suggest that a testing effect is possible for flexible procedures. They also suggest that multiple learning processes, such as memory and inductive processes, are involved in procedural learning. The dominance of one learning process over another may be related to the retention interval and nature of the learning materials, such as the variability of problems used. In summary, the optimal learning strategy depends on both the learning goal and the learning processes activated during practice.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of Problem Solving after Worked Example Study on Primary School Children's Monitoring Accuracy. *Applied Cognitive Psychology*, *28*(3), 382–391. https://doi.org/10.1002/acp.3008

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology*, *37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about Knowing.*

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941–956. https://doi.org/10.1002/acp.1239

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474–478. https://doi.org/10.3758/BF03194092

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 1020–1031. https://doi.org/10.1037/0278-7393.22.4.1020

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, *127*(4), 355–376. https://doi.org/10.1037/0096-3445.127.4.355

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1147–1156. https://doi.org/10.1037/0278-7393.15.6.1147

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347–362. https://doi.org/10.1037/0022-0663.79.4.347

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, *23*(4), 1809–1819. https://doi.org/10.1016/j.chb.2005.11.001

de Bruin, A B H, Rikers, R., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology*, *19*(4–5), 671–688. https://doi.org/10.1080/09541440701326204

de Bruin, Anique B H, Rikers, R. M. J. P., & Schmidt, H. G. (2005). Monitoring accuracy and self-regulation when learning to play a chess endgame. *Applied Cognitive Psychology*, *19*(2), 167–181. https://doi.org/10.1002/acp.1109

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions

From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266
Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*(3), 335–350. https://doi.org/10.1080/09658211003652491
Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6
Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y
Kalyuga, S., Renkl, A., & Paas, F. (2010). Facilitating flexible problem solving: A cognitive load perspective. *Educational Psychology Review*, *22*(2), 175–186. https://doi.org/10.1007/s10648-010-9132-9
Kalyuga, S., & Singh, A.-M. (2016). Rethinking the Boundaries of Cognitive Load Theory in Complex Learning. *Educational Psychology Review*, *28*(4), 831–852. https://doi.org/10.1007/s10648-015-9352-0
Kang, S. H. K., McDaniel, M. a., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998–1005. https://doi.org/10.3758/s13423-011-0113-x
Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669
Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science*, *38*(6), 523–550. https://doi.org/10.1007/s11251-009-9093-x
Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107
Karpicke, J. D., & Aue, W. R. (2015). The Testing Effect Is Alive and Well with Complex Materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3
Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009
Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). *Retrieval-Based Learning. An Episodic Context Account. Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 61). https://doi.org/10.1016/B978-0-12-800283-4.00007-1
Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-Offs Between Grounded and Abstract Representations: Evidence From Algebra Problem Solving. *Cognitive Science: A Multidisciplinary Journal Cognitive Science*, *32*(32), 366–397. https://doi.org/10.1080/03640210701863933
Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x
Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/BF03194055
Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21–27. https://doi.org/10.1111/j.1365-2923.2008.03245.x
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*, *27*(2), 291–304. https://doi.org/10.1007/s10648-015-9296-4

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1787–1794. https://doi.org/10.1037/xlm0000012

Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: alternative pathways to learning complex tasks. *Instructional Science*, *45*(2), 195–219. https://doi.org/10.1007/s11251-016-9399-4

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, *15*(1), 1–11. https://doi.org/10.1037/a0014721

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

Paas, F. G. W. C., & 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*(1), 122–133. https://doi.org/10.1037/0022-0663.86.1.122

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A. C., Koedinger, K. R., McDaniel, M. A., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144–161. https://doi.org/10.1037//0022-0663.88.1.144

Rawson, K. A. (2015). The Status of the Testing Effect for Complex Materials: Still a Winner. *Educational Psychology Review*, *27*(2), 327–331. https://doi.org/10.1007/s10648-015-9308-4

Reed, S. K., & Bolstad, C. a. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 753–766. https://doi.org/10.1037/0278-7393.17.4.753

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin*, *115*(3), 381–400. https://doi.org/10.1037/0033-2909.115.3.381

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction*, *12*(5), 529–556. https://doi.org/10.1016/S0959-4752(01)00030-5

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

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*(1), 90–108. https://doi.org/10.1006/ceps.1997.0959

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental*

Psychology: Applied, *17*(4), 382–395. https://doi.org/10.1037/a0026252
Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/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*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155
Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 42–55. https://doi.org/10.1037/0278-7393.16.1.42
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559
Schwartz, D. L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, *22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1
Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7
Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5
Sweller, J., & Cooper, G. A. (1985). The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition and Instruction*, *2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3
van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*, *36*(8), 1532–1541. https://doi.org/10.1111/cogs.12002
van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing After Worked Example Study Does Not Enhance Delayed Problem-Solving Performance Compared to Restudy. *Educational Psychology Review*, *27*(2), 265–289. https://doi.org/10.1007/s10648-015-9297-3
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*(3), 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004
van Gog, T., Paas, F. G. W. C., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction*, *16*(2 SPEC. ISS.), 154–164. https://doi.org/10.1016/j.learninstruc.2006.02.003
van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*(2), 155–174. https://doi.org/10.1007/s10648-010-9134-7
van Gog, T., & Sweller, J. (2015). Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. *Educational Psychology Review*, *27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x
Zhu, X., & Simon, H. A. (1987). Learning Mathematics From Examples and by Doing.

Cognition and Instruction, *4*(3), 137–166. https://doi.org/10.1207/s1532690xci0403_1

Online Supplement

Table S1

Examples of isomorphic learning/test and transfer test problems

Problem Type	Worked-out Solutions
Isomorphic Given: • Two events • Different means within the same given interval for each event To find: • Same larger required interval (Note: Inequality type varied across the problems: "greater than", "less than", "at most"/"no more than", and "at least"/"no less than")	A car salesperson sells, on average, 3 new cars and 2 used cars in two weeks. The number of new cars she sells is independent of the number of old car she sells, and they each follow independent Poisson distributions. Find the probability that she sells at least 5 but at most 7 cars in a randomly chosen four-week period. **Step 1: Find the mean of event X for the required interval or space.** Let N be the number of new cars she sells in a four-week period. 2 weeks $\rightarrow$ 3 new cars 4 weeks $\rightarrow 3 \times 2 = 6$ new cars So, $N \sim P_o(6)$. **Step 2: Find the mean of event Y for the required interval or space.** Let U be the number of used cars she sells in a four-week period. 2 weeks $\rightarrow$ 2 used cars 4 weeks $\rightarrow 2 \times 2 = 4$ used cars So, $U \sim P_o(4)$. **Step 3: Find the combined mean of events X and Y for the required interval or space.** Let T be the total number of new and used cars she sells in a four-week period, i.e., N + U. $T \sim P_o(6 + 4)$ i.e., $T \sim P_o(10)$. **Step 4: Find the required probability.** $P(5 \leq T \leq 7) = P(T = 5) + P(T = 6) + P(T = 7)$ $= \frac{e^{-10} \cdot 10^5}{5!} + \frac{e^{-10} \cdot 10^6}{6!} + \frac{e^{-10} \cdot 10^7}{7!} = .191$
Transfer Given: • One event To find: • Smaller required interval • Inequality that includes zero	At a newly opened bistro, the number of orders for clam chowder received in a randomly chosen one-hour period follows a Poisson distribution with mean 4.6. Find the probability that there are less than 2 orders received in a randomly chosen 30-minute interval. **Step 1: Find the mean of event X for the required interval or space.** Let C be the number of orders for clam chowder received in a 30-minute interval. 1 hour $\rightarrow$ 4.6 orders

30 minutes $\rightarrow$ 4.6 ÷ 2 = 2.3 orders
So, C ~ $P_o(2.3)$.

Step 2: Find the probability.
$P(0 \leq C < 2) = P(C = 0) + P(C = 1)$

$$= \frac{e^{-2.3} \cdot 2.3^0}{0!} + \frac{e^{-2.3} \cdot 2.3^1}{1!} = .331$$

Transfer
Given:
- Two events
- Different means and different given intervals for each event

To find:
- larger required interval for one event and a smaller required interval for the other event

The two most common types of disciplinary offenses in a particular boys' school in England is keeping long hair and failure to wear the school badge. Assuming that each school week consists of five school days and each school month consists of 20 school days, the mean number of disciplinary offenses recorded per day involving long hair is 1.35, and the mean number of disciplinary offenses recorded per school month involving failure to wear the school badge is 5. The number of cases for each disciplinary offense is assumed to have an independent Poisson distribution. Find the probability that more than 8 and less than 11 cases of disciplinary offenses are recorded in a randomly chosen week.

Step 1: Find the mean of event X for the required interval or space.
Let H be the number of disciplinary offences recorded for long hair per week.
1 day $\rightarrow$ 1.35 cases
5 days $\rightarrow$ 1.35 × 5 = 6.75 cases
So, H ~ $P_o(6.75)$.

Step 2: Find the mean of event Y for the required interval or space.
Let B be the number of disciplinary offences recorded for failure to wear the school badge per week.
20 days $\rightarrow$ 5 cases
5 day $\rightarrow \frac{5}{20} \times 5$ = 1.25 cases
So, B ~ $P_o(1.25)$.

Step 3: Find the combined mean of events X and Y for the required interval or space.
Let T be the total number of cases of disciplinary offences recorded in a week, i.e., H + B.
T ~ $P_o(6.75 + 1.25)$ i.e., T ~ $P_o(8)$.

Step 4: Find the required probability.
$P(8 < T < 11) = P(T = 9) + P(T = 10)$

$$= \frac{e^{-8} \cdot 8^9}{9!} + \frac{e^{-8} \cdot 8^{10}}{10!} = .223$$

Transfer

Given:

- Two events
- Same given mean within the same given interval/space (defined in terms of quantity instead of temporal intervals)

To find:

- Larger, but different required intervals/spaces

In the production of cellphone screen protectors, scratches occur at random and independently, and they follow a Poisson distribution with a mean of 0.15 scratches per screen protector. In a quality control inspection, 100 screen protectors produced by manufacturer A and 200 screen protectors produced by manufacturer B were selected randomly. Find the probability that there are more than 51 but no more than 54 scratches in a randomly selected quality control inspection.

Step 1: Find the mean of event X for the required interval or space.

Let A be the number of scratches found on 100 screen protectors produced by manufacturer A in an inspection.

1 screen protectors $\rightarrow$ 0.15 scratches

100 screen protectors $\rightarrow 0.15 \times 100 = 15$ scratches

So, $A \sim P_o(15)$.

Step 2: Find the mean of event Y for the required interval or space.

Let B be the number of scratches found on 200 screen protectors produced by manufacturer B in an inspection.

1 screen protectors $\rightarrow$ 0.15 scratches

200 screen protectors $\rightarrow 0.15 \times 200 = 30$ scratches

So, $B \sim P_o(30)$.

Step 3: Find the combined mean of events X and Y for the required interval or space.

Let T be the total number of scratches found on the 300 screen protectors produced by manufacturers A and B in a randomly selected inspection i.e., A + B.

$T \sim P_o(15 + 30)$ i.e., $T \sim P_o(45)$.

Step 4: Find the required probability.

$$P(51 < T \leq 54) = P(T = 52) + P(T = 53) + P(T = 54)$$

$$= \frac{e^{-45.}\,45^{52}}{52!} + \frac{e^{-45.}\,45^{53}}{53!} + \frac{e^{-45.}\,45^{54}}{54!} = .084$$

Note. While each step was labelled "Step 1/2/3/4", the sub-goals were not seen by participants during the learning phase.

Introductory cover sheet

The Poisson Distribution

A Bernoulli trial is one in which the outcome is either a success or a failure. An example would be flipping a coin where heads is considered a "success" and tails is considered a "failure." Often, over the course of a series of Bernoulli trials, the most important information is not which trials ended in success and which in failure, but rather, **how many** ended in success or failure. Let $\boldsymbol{X}$ denote the number of successes in $\boldsymbol{n}$ Bernoulli trials, when the probability of a success on any particular trial is $\boldsymbol{p}$. Then $\boldsymbol{X}$ is said to have a Binomial distribution, $\boldsymbol{X \sim B(n, p)}$, and the probability of getting $\boldsymbol{x}$ successes in $\boldsymbol{n}$ trials is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , x = 0, 1, 2, \dots, n.$$

The equation above can sometimes get quite messy when $\boldsymbol{n}$ and $\boldsymbol{x}$ get large. For certain events that occur **singly**, **independently** and **randomly**, with the **probability** $\boldsymbol{p}$ of one **event occurring** within a small fixed interval of time (or space) is the **same** and **fairly low** at all points in time (or space), we can often use the Poisson distribution, $\boldsymbol{X \sim P_o(\lambda)}$, as a replacement for the Binomial distribution to model the frequency of the occurrence of the events. We can replace the Binomial equation with the Poisson equation:

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad , x = 0, 1, 2, \dots, n, and\ \lambda > 0$$

where $\boldsymbol{e} \approx 2.718$, and where $\boldsymbol{\lambda}$ is the expected value (that is, the **average** or **mean** value) of the random variable $\boldsymbol{X}$. This equation is much easier to calculate for the various values of $\boldsymbol{X}$ than the Binomial equation.

The distribution has a mean number (or expected number) of occurrences, $\boldsymbol{\lambda}$, in a given time (or space) that is proportional to the time (or space) interval. For example, if $\boldsymbol{\lambda}$ is the mean number of phone calls received in a 1-minute interval, then the mean number of phone calls received in a 2-minute interval will be equal to $2\boldsymbol{\lambda}$.

If $X \sim P_o(\lambda_1)$ and $Y \sim P_o(\lambda_2)$, where X and Y are **independent**, and $W = X + Y$ (sum of Poisson random variables), then W is also a Poisson random variable, $W \sim P_o(\lambda_1 + \lambda_2)$.

***

Instructions on how to use the Texas Instruments TI-30XS MultiView calculator

To compute the value of a number raised to a power n:
For example, to find $10^{-3.2}$, press: **1 0** [^] [(–)] **3 . 2** [enter]

To compute the factorial of integer n:
For example, to find 7!, press: **7** [prb] **3** [enter]

Additional Analyses for Experiment 1

Subjective prior knowledge

Subjective prior knowledge was measured both categorically (yes or no), and as a continuous variable on a recall scale of 1 (not at all) to 7 (very much), if answered participants answered 'yes'. Participants with absolutely no prior exposure to Poisson distribution were coded as '1' even though they did not explicitly answer the question. Participants with prior knowledge (see Table S2) were not found to be disproportionately distributed among the four conditions for the remember passage group (p = .441, two-tailed Fisher's exact test). Similarly, participants with prior knowledge were not found to be disproportionately distributed among the four conditions for the learn procedure group (p = .391, two-tailed Fisher's exact test).

Within the remember passage group, a 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA on subjective ratings of prior knowledge also revealed no main effects of learning strategy (repeated studying: M = 1.38, SD = 0.81; repeated testing: M = 1.30, SD = 0.69; F < 1), retention interval (5-minute: M = 1.20, SD = 0.56; 1-week: M = 1.48, SD = 0.88; $F(1, 76) = 2.72$, p = .103, η_p^2 = .035), or an interaction (F < 1). Within the learn procedure group, there were also no main effects of learning strategy (repeated studying: M = 1.40, SD = 0.78; repeated testing: M = 1.35, SD = 1.33; F < 1), retention interval (5-minute: M = 1.58, SD = 1.43; 1-week: M = 1.18, SD = 0.50; $F(1, 76) = 2.74$, p = .102, η_p^2 = .035), or an interaction (F < 1). These suggest that our findings were not influenced by an unequal distribution of participants with prior knowledge of Poisson distribution.

Table S2

Proportion of participants with prior knowledge of Poisson distribution within each condition

	Remember passage		Learn procedure	
Retention interval	Repeated studying	Repeated testing	Repeated studying	Repeated testing
5 minutes	15%	10%	25%	20%
1 week	30%	25%	20%	5%

Problem-solving performance for remember passage group

Although irrelevant to our main hypotheses, for completeness, we also analyzed participants' ability to learn the procedure when their goal was to recall the Airport Problem. A 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA on problem-solving performance revealed that problem-solving performance was higher after a 5-minute delay (M = 65.6%, SD = 32.0) than after a 1-week delay (M = 46.3%, SD = 42.1), $F(1, 76) = 5.82$, $p = .024$, $\eta_p^2 = .065$. However, there was no difference in problem-solving performance between repeated studying (M = 59.4%, SD = 38.8) and repeated testing (M = 52.5%, SD = 38.2), $F < 1$, and no interaction, $F < 1$. When problem solving was not an explicit learning goal, the efficacies of repeated studying and retrieval practice did not differ significantly in either delay condition.

Recall performance for learn procedure group

Similarly, for completeness, we analyzed participants' ability to recall the Airport Problem when their goal was to learn the procedure. It is crucial to note that all participants in the learn procedure group were only exposed to the Airport Problem once regardless of learning strategy assignment. Recall performance was higher after 5 minutes (M = 17.9%, SD = 18.2), than after 1 week (M = 4.8%, SD = 8.1), $F(1, 76) = 17.21$, $p < .001$, $\eta_p^2 = .185$. However, there was no difference in recall performance between repeated studying (M = 9.0%, SD = 13.2) and repeated testing (M = 13.6%, SD = 17.3), $F(1, 76) = 2.15$, $p = .147$, $\eta_p^2 = .027$, and no interaction, $F < 1$. As with the remember passage group, the efficacies of repeated studying and retrieval practice did not differ in either delay condition.

Additional Analyses for Experiment 2

Subjective prior knowledge

Participants with prior knowledge (Table S3) were not disproportionately distributed among the six conditions ($p = .631$, two-tailed Fisher's exact test). A 2 (learning strategy: repeated studying vs. repeated testing) × 2 (retention interval: 5 minutes vs. 1 week) ANOVA on subjective ratings of prior knowledge also revealed no main effects of learning strategy (repeated studying: M = 1.73, SD = 1.26; repeated testing without feedback: M = 1.43, SD = 1.03; repeated testing with feedback: M = 1.70, SD = 1.68; $F < 1$), retention interval (5-minute: M = 1.58, SD = 1.34; 1-week: M = 1.65, SD = 1.36; $F < 1$), or an interaction ($F < 1$). These suggest that our findings were not influenced by an unequal distribution of participants with prior knowledge of Poisson distribution.

Table S3

Proportion of participants with prior knowledge of Poisson distribution within each condition

Retention interval	Repeated studying	Repeated testing without feedback	Repeated testing with feedback
5 minutes	35%	20%	15%
1 week	35%	25%	20%

Table S4

Frequencies (percentage of trials) of errors in Experiment 2

	During Learning				Final Test					
	Repeated testing without feedback		Repeated testing with feedback		Repeated studying		Repeated testing without feedback		Repeated testing with feedback	
Error Type	5 minutes	1 week	5 minutes	1 week	5 minutes	1 week	5 minutes	1 week	5 minutes	1 week
No attempt	3.3	0	1.7	0	1.9	10	3.8	1.9	1.9	2.5
Answer only	1.7	0	0	0	0	0	1.9	0	0	0
Incomplete	0	1.7	3.3	0	1.9	2.5	1.3	0	1.3	0
Conceptual/Procedural										
Steps 1 and 2 (means)	0.8	0	3.3	0	3.8	17.8	8.8	7.8	5.6	4.1
Step 3 (sum of means)	0	0	0	1.7	0	8.8	4.4	2.5	0.6	1.3
Step 4 (inequality)	8.3	5.0	0	3.3	18.1	25.0	18.8	18.8	18.1	20.6
Formula application	5.0	0	5.0	6.7	3.1	21.3	5.6	3.1	3.1	5.0
Technical										
Arithmetic	0	1.7 (1.7)	0 (1.7)	0 (1.7)	0	2.5 (0.6)	1.9 (0.6)	0	0 (1.3)	0
Copy slip	1.7	1.7	0	0	0	0.6	0.6	0	0	0.6

Note. Percentages in parentheses refer to arithmetic errors made in solutions that were ultimately coded as correct.

Additional Analyses for Experiment 3

Subjective prior knowledge

Participants with prior knowledge were not disproportionately distributed among the three conditions (repeated studying: 35%; repeated testing without recall instructions: 10%; repeated testing with recall instructions: 25%; $p = .207$, two-tailed Fisher's exact test). A one-way (learning strategy: repeated studying vs. repeated testing without recall instructions vs. repeated testing with recall instructions) ANOVA on subjective ratings of prior knowledge also revealed no main effects of learning strategy (repeated studying: $M = 1.85$, $SD = 1.35$; repeated testing without recall instructions: $M = 1.20$, $SD = 0.70$; repeated testing with recall instructions: $M = 1.55$, $SD = 1.15$, $F(2, 57) = 1.76$, $p = .182$, $\eta_p^2 = .058$. These suggest that our findings were not influenced by an unequal distribution of participants with prior knowledge of Poisson distribution.

Table S5

Frequencies (percentage of trials) of errors in Experiment 3

Error Type	During Learning		Final Test		
	Repeated testing without recall instructions	Repeated testing with recall instructions	Repeated studying	Repeated testing without recall instructions	Repeated testing with recall instructions
No attempt	0	0	5.6	5.0	0.6
Answer only	0	0	0	0	0
Incomplete	0	0	3.8	2.5	0.6
Conceptual/Procedural					
Steps 1 and 2 (means)	5.0	5.0	9.4	6.6	10.6
Step 3 (sum of means)	0	5.0	6.3	3.1	7.5
Step 4 (inequality)	13.3	18.3	13.8	15.6	17.5
Formula application	8.3	20.0	9.4	4.4	14.4
Technical					
Arithmetic	1.7	0 (1.7)	0	0 (1.3)	0.6 (1.3)
Copy slip	0	0	0.6	0.6	1.3

Note. Percentages in parentheses refer to arithmetic errors made in solutions that were ultimately coded as correct.