**BRIEF REPORT**

# The effects of retrieval versus study on analogical problem solving

Daniel Corral[1] · Shana K. Carpenter[2] · Kyle J. St. Hilaire[2]

## Abstract

We report four experiments, wherein subjects engaged in either problem-solving practice or example study. First, subjects studied an example problem. Subjects in the example study condition then studied two more analogous problems, whereas subjects in the problem-solving practice conditions solved two such problems, each followed by correct-answer feedback. In Experiment 1, subjects returned 1 week later and completed a posttest on an analogous problem; in Experiments 2–4, subjects completed this posttest immediately after the learning phase. Additionally, Experiment 3 consisted of a control condition, wherein subjects solved these same problems, but did not receive feedback. Experiments 3 and 4 also included a mixed study condition, wherein subjects studied two examples and then solved one with feedback during the learning phase. Across four experiments, we found that the training conditions (i.e., problem-solving practice, mixed, and example study) performed equally well on the posttest. Moreover, subjects in the training conditions outperformed control subjects on the posttest, indicating that the null findings were due to the training conditions learning and transferring their knowledge equally well. After the posttest in Experiment 4, subjects were asked to solve repeated problems from the learning phase. Subjects in the problem-solving practice and mixed study conditions performed better on repeated problems than subjects in the example study condition, indicating that they better learned the solution strategies for these problems than subjects in the example study condition. Nevertheless, this benefit was insufficient to produce differential transfer of learning among the training conditions on the posttest.

**Keywords** Analogical problem solving · Retrieval practice · Example study · Concept learning · Transfer of learning

## Introduction

One of the most robust findings in cognitive psychology is that the practice of retrieving information leads to better learning than re-studying it (Bjork & Bjork, 1992; Carpenter, 2009, 2011, 2012; Carrier & Pashler, 1992; Dunlosky et al., 2013; Roediger & Butler, 2011). However, much of this work has involved assessments of memory performance. Recent work has attempted to extend the benefits of retrieval practice to tasks that involve problem-solving transfer, which requires more than memorization. Consider the solution strategy to the problem in Fig. 1A. Although applying this solution strategy to analogous problems (see Fig. 1B–D) involves memorizing the solution strategy, it also involves learning to recognize *when* and *how* to use it (Corral et al., 2021). If either of these latter two processes fail, even if a solution strategy is memorized, the learner will not be able to transfer it to new situations (Gick & Holyoak, 1987).

Accordingly, learners often fail to apply solutions from previous problems to novel, analogous scenarios that differ superficially (e.g., Butler, 2010; Gick & Holyoak, 1980, 1983). However, learners can solve such problems when they are reminded to think about how a novel problem relates to previous examples. These findings suggest that learners can successfully acquire and apply to-be-learned solution strategies but struggle to recognize when to use them (formally known as the *inert knowledge problem*; Whitehead, 1929). Thus, for retrieval practice to aid problem-solving transfer, it must help learners (a) acquire the corresponding solution strategy and (b) recognize when to apply it.

Present theories on the benefits of retrieval practice focus on how retrieval strengthens memory of the information that is retrieved, but do not directly explain how

---

✉ Daniel Corral
dcorral@syr.edu

[1] Department of Psychology, Syracuse University, Syracuse, NY, USA

[2] Department of Psychology, Iowa State University, Ames, IA, USA

A.

Suppose you are a doctor faced with a patient who has a malignant tumor in his stomach. To operate on the patient is impossible, but unless the tumor is destroyed, the patient will die. A kind of ray, at a sufficiently high intensity, can destroy the tumor. Unfortunately, at this intensity the healthy tissue that the rays pass through on the way to the tumor will also be destroyed. At lower intensities the rays are harmless to healthy tissue, but will not affect the tumor.

How can the rays be used to destroy the tumor without injuring the healthy tissue?

**Solution:**
**The doctor can use many weak beams of radiation with each beam taking a different route to the tumor, with all of the beams converging on the tumor simultaneously. Combined, the beams are strong enough to destroy the tumor, but individually are not strong enough to harm the healthy tissue they pass through.**

B.

A small country was ruled from a strong fortress by a dictator. The fortress was situated in the middle of a country, surrounded by farms and villages. Many roads led to the fortress through the countryside. A rebel general vowed to capture the fortress. The general knew that an attack by his entire army would capture the fortress. He gathered his army at the head of one of the roads, ready to launch a full-sale direct attack. However, the general then learned that the dictator had planted mines on each of the roads. The mines were set so that small bodies of men could pass over them safely, since the dictator needed to move his troops and workers to and from the fortress. However, any large force would detonate the mines. Not only would this blow up the road, but it would also destroy many neighboring villages. It therefore seemed impossible to capture the fortress.

How should the general solve this problem?

**Solution:**
**However, the general devised a simple plan. He divided his army into small groups and dispatched each group to the head of a different road. When all was ready, he gave the signal and each group marched down a different road. Each group continued down its road to the fortress, so that the entire army arrived together at the fortress at the same time. In this way, the general captured the fortress and overthrew the dictator.**

**Fig. 1** Each problem used in Experiments 1–4 (taken from Catrambone & Holyoak, 1989), along with the corresponding correct solutions. Panel **A** shows Dunker's radiation problem (1945), Panel **B** shows the invading general problem, Panel **C** shows the fire chief problem, and Panel **D** shows the aquarium problem

retrieval aids the transfer of learning (Carpenter et al., 2022; Pan & Rickard, 2018). One possibility is that when learners attempt to solve a problem, it allows them the opportunity to retrieve and apply the solution strategy. This opportunity is not afforded to learners when they study worked examples, as worked examples already contain the solution strategy and its application. Based on the voluminous literature on retrieval practice, problem-solving practice should therefore produce better learning and memory of to-be-learned solution strategies, leading to superior problem solving.

It is also possible that problem-solving practice enhances memory of previous problems, which might help learners recognize the similarity among old problems and novel analogs. Critically, this recognition can help learners figure out when to use the correct solution strategy.

These hypotheses offer a mechanistic account of how retrieval practice (via problem-solving practice) can facilitate problem-solving transfer. Nevertheless, these ideas have not been empirically examined and are presently open questions.

On the other hand, research on the *worked example effect* has found evidence against the benefits of retrieval practice on problem-solving tasks, as subjects who study worked examples often outperform subjects who engage in problem-solving practice (Cooper & Sweller, 1987; Sweller & Cooper, 1985; Van Gog & Kester, 2012; Van Gog et al., 2011). However, these studies often do not provide feedback

C.

> An oil well in Saudi Arabia exploded and caught fire. The result was a blazing inferno that consumed an enormous quantity of oil each day. After initial efforts to extinguish it failed, famed firefighter Red Adair was called in. Red knew that the fire could be put out if a huge amount of fire retardant foam could be dumped on the base of the well. There was enough foam available at the site to do the job. However, there was no hose large enough to put all the foam on the fire fast enough. The small hoses that were available could not shoot the foam quickly enough to do any good. Adding to the problem, the firefighters did not have enough fire retardant to waste any of it and still put out the fire. It looked like there would have to be a costly delay before a serious attempt could be made.
>
> How should Red solve this problem?
>
> **Solution:**
>
> **However, Red Adair knew just what to do. He stationed men in a circle all around the fire, with all of the available small hoses. When everyone was ready, all the hoses were opened up and the foam was directed at the fire from all directions. In this way a huge amount of foam quickly struck the source of the fire. The blaze was extinguished, and the Saudis were satisfied that Red had earned his three-million-dollar fee.**

D.

> A major aquarium in a city on the East Coast decided to create a large aquarium display containing a replica of the sunken ocean liner the Titanic amid the sea environment of its resting place, which is deep in the Atlantic Ocean off the coast of Newfoundland. A professional aquarium designer was assigned to the project. She placed a small replica of the vessel in the center of a large tank, with a realistic seabed. Then she added to the tank sea plants and fish of the sort that live in the Atlantic at the depth of the sunken Titanic. The display was virtually finished when the designer was confronted with a major problem she had failed to anticipate. In order to maintain the deep-water environment required by the fish and plants, the tank had to be kept quite dark, as the deep-water organisms were not adapted to light. However, if the tank was kept completely dark, people would not be able to see the small replica of the Titanic in the center of the tank, which, after all, was the main point of the exhibit. Putting lights inside the model of the wreck looked too artificial. The designer considered shining a powerful spotlight on the model of the vessel. However, if the spotlight was located inside the tank, it would raise the temperature of the water too high; and if it was located outside the tank, the bright beam seriously disrupted the feeding habits of some of the fish.
>
> How should the designer solve this problem?
>
> **Solution:**
>
> **Just before the display was to open, the designer hit upon a new idea. She had several low-powered spotlights placed at different locations around the outside of the tank, all focused on the replica of the ship. Each of the lights was quite dim, so the light-averse fish were not disturbed as they swam around the Titanic. But since all of the lights were focused on the ship, their beams added up so as to illuminate it enough that its realistic details could be seen by viewers. When the display opened, everyone thought it was both realistic and esthetically striking.**

**Fig. 1** (continued)

with problem-solving practice. Feedback is critical for learning (Benassi et al., 2014) and is necessary for complex concept acquisition (Corral & Carpenter, 2020). Thus, not presenting feedback during problem-solving practice might undermine learning. In such cases, it is unclear whether the results are due to the advantage of studying worked examples over problem-solving practice, or if those results are driven by differences in feedback presentation.

Recent studies have applied retrieval practice to problem solving by having subjects read a problem scenario and retrieve its surface details (instead of solving the problem; Hostetter et al., 2018; Peterson & Wissman, 2018). Although

this approach has not yielded a clear benefit of retrieval over study, retrieving a problem's surface features can lead to forming incorrect representations, which can inhibit learning and transfer (Anderson, 1993; Corral & Jones, 2014; Holyoak & Koh, 1987; Ross, 1987, 1989; Sweller et al., 1983).

In the present paper, we explore the effects of retrieval practice and example study on problem-solving transfer through a design that addresses the aforementioned limitations. We report four experiments that compare problem-solving practice and example study. If problem-solving practice enhances memory for a solution strategy and when to apply it, then problem-solving practice should lead to best performance on a posttest involving application of the solution strategy to a new problem.

## Experiment 1

Subjects were presented four analogous problems from Catrambone and Holyoak (1989). First, subjects were asked to study a single problem scenario with the solution strategy. Subjects in the example condition were then presented two more example problems to study, whereas subjects in the problem-solving practice condition were asked to solve two problems, each followed by correct-answer feedback. Subjects returned 1 week later and completed a posttest that consisted of a novel problem.

### Methods

#### Subjects

One hundred thirty-eight undergraduate students from Iowa State University (ISU) participated in this experiment for course credit in an introductory psychology course. Approximately 67% of students who attend ISU are 21 years of age or younger, approximately 43% identify as female, and approximately 75% are White.

Thirty-one subjects did not return for the second part of the experiment (i.e., the posttest) and were not included in any of the analyses. All reported analyses were thus based on the remaining 106 subjects: (a) problem-solving practice ($n = 54$) and (b) example study ($n = 52$).

Experiments 1–3 were approved by the Institutional Review Board at ISU. The sample size for Experiments 1–3 were approximated based on an a priori power analysis with 80% power to detect a medium effect size ($f = .25$; $\alpha = .05$).

#### Design and materials

Subjects were randomly assigned to one of two conditions: (a) problem-solving practice and (b) example study. All instructions and materials were presented on a computer

monitor on a white background at the center of the screen. Subjects entered all responses using a computer keyboard and mouse.

The materials consisted of four analogous problems (see Fig. 1), which were taken from Catrambone and Holyoak (1989; also see Gick & Holyoak, 1980, 1983). These problems consisted of Dunker's radiation problem (1945) and three analogs. These problems all differ superficially, but consist of the same structure, wherein there are enough resources/forces in the problem scenario to reach the problem's goal state, but to do so, those resources/forces must be fully concentrated on a given target, and there is an obstruction that prevents this from occurring. Due to their shared structure, these problems can be solved by using the same solution strategy, wherein the resources/forces in the problem scenario are divided along alternate paths and then converge simultaneously on the specified target (i.e., a divide and converge strategy).

The correct response for each problem was broken down into three components: (a) dividing the resources/forces in the problem, (b) sending resources/forces down different paths that surround the target location, and (c) the resources/forces converging on the target location simultaneously (for a similar approach to scoring these types of problems, see Snoddy & Kurtz, 2021). Responses were scored on a scale of 0–1. A response was scored as fully correct if all three of these solution components were included. Partial credit was awarded for responses that included one (1/3 partial credit) or two (2/3 partial credit) of these solution components. Posttest responses were scored blindly to condition; responses from the learning phase from subjects in the problem-solving practice condition were scored blindly to the order that the problems were presented. This same scoring scheme was used for Experiments 2–4.

#### Procedure

This experiment was two parts: The learning phase occurred in the first part and the posttest occurred in the second part, 1 week later. During the learning phase, all subjects were presented with three of the four problems (one at a time). At the beginning of the learning phase, all subjects were presented with one of these problems along with the corresponding solution (as shown in Fig. 1) and were asked to carefully read and study them.

For subjects in the example study condition, they were then shown two more of these example problems (one at a time), along with their corresponding solutions. For subjects in the problem-solving practice condition, after being shown the first example problem, they were asked to solve two problems (one at a time). Subjects were asked to type out their solution into a textbox, which was presented directly beneath the scenario. After responding, these subjects were

**Table 1** Mean time subjects in each training condition took to complete the learning phase in each experiment

|              | Example study | Problem-solving | Mixed        |
|--------------|---------------|-----------------|--------------|
| Experiment 1 | 6.98 (.675)   | 9.06 (.662)     | ----------   |
| Experiment 2 | 9.93 (.636)   | 11.31 (.630)    | ----------   |
| Experiment 3 | 5.91 (.429)   | 8.68 (.436)     | 7.78 (.429)  |
| Experiment 4 | 9.64 (.410)   | 10.96 (.423)    | 10.46 (.410) |

Standard errors of the mean are shown in parentheses. The results show a general pattern, wherein subjects in the example study conditions generally took less time to complete the learning phase than subjects in the other training conditions. This pattern was statically reliable in Experiments 1 and 3 (both $ps < .032$), but did not reach statistical significance in Experiments 2 and 4 (both $ps > .078$)

**Table 2** Mean performance on each posttest problem scenario in each experiment (collapsed across training conditions)

|              | Radiation   | Invading general | Fire chief  | Aquarium    |
|--------------|-------------|------------------|-------------|-------------|
| Experiment 1 | .453 (.068) | .578 (.056)      | .476 (.074) | .277 (.069) |
| Experiment 2 | .444 (.066) | .683 (.075)      | .562 (.058) | .526 (.060) |
| Experiment 3 | .408 (.056) | .557 (.053)      | .509 (.061) | .313 (.050) |
| Experiment 4 | .325 (.036) | .556 (.038)      | .453 (.039) | .334 (.037) |

Standard errors of the mean are shown in parentheses. Experiments 1, 3, and 4 revealed differential posttest performance among the problem scenarios (all $ps < .013$), as the aquarium (Experiments 1, 3, and 4) and radiation (Experiments 3 and 4) problems were the most difficult to solve. However, problem scenario did not interact with training condition in any of the experiments (all $ps > .269$). Thus, the reported results do not differ based on the problem scenario

presented with correct-answer feedback (as shown in Fig. 1), wherein they were shown the solution to the problem and were asked to carefully study the problem and its solution. Subjects in both conditions were therefore presented identical materials and feedback. Thus, the only difference between these conditions was whether subjects were asked to solve the problem prior to being shown its solution.

All aspects of the experiment (i.e., example and feedback study, problem solving) were self-paced. To move to the next problem, a prompt and continue button were presented near the right side of the bottom of the screen, which notified subjects that they could move on when they were ready by clicking on this button. For each problem that subjects were asked to solve (i.e., problem-solving practice and posttest problems), they were required to type a response into the textbox before they were allowed to see the correct solution (applicable for problem-solving practice problems) or move on (applicable for posttest problems).

After completing the learning phase, all subjects were thanked for their participation and were presented with a prompt reminding them to return 1 week later for the second part of the experiment. Table 1 shows mean completion times on the learning phase partitioned by training condition for each experiment. Subjects took approximately 8 min to complete the learning phase ($M = 8.03$ min).

Upon returning 1 week later, all subjects were given a posttest in which they were asked to solve one problem. For all subjects, the posttest problem that they were presented was the problem that was withheld during the learning phase. The order in which the problems were presented was randomized for each subject but was constrained by pseudo counterbalancing, wherein the same presentation orders were used in each condition equally.[1] Thus, the order

in which the problem scenarios were presented was identical across training conditions. Table 2 shows mean posttest performance for each problem scenario (collapsing across training conditions).

## Results and discussion

Table 3 shows mean performance for each problem that was solved in the learning phase and the posttest across each experiment. Table 4 shows mean performance on the posttest partitioned by condition in each experiment. Because time spent on the learning phase was allowed to vary across conditions, it was included as a covariate in all experiments.

An Analysis of Covariance (ANCOVA) (with time on the learning phase as a covariate) revealed no reliable posttest differences between conditions (see top row of Table 4), $F(1,103) = 0.028$, $p = .868$, $MSE = .119$, $\eta_p^2 = .000$.[2] Furthermore, a Bayesian version of this ANCOVA found support for the null hypothesis ($BF = .209$), as these results were 4.80 times more likely to occur under the null model.[3] Thus, when the problem-solving and example study conditions were carefully controlled and only differed on whether subjects attempted to solve the problems, subjects in both conditions produced similar levels of transfer on the posttest.

---

[1] It is important to note that, by and large, studies that compare problem-solving practice to example study (e.g., Cooper & Sweller, 1987; Sweller & Cooper, 1985; Van Gog & Kester, 2012; Van Gog et al., 2011; also see Hostetter et al., 2018; Peterson & Wissman, 2018) do not typically randomize the presentation order of the problems, which

Footnote 1 (continued)

may introduce order effects that differentially interact with the training conditions. Randomizing the presentation order of the problems thus affords a more rigorous form of experimental control.

[2] For Experiments 1–4, the pattern and interpretation of the results are the same when time on the learning phase is not included as a covariate.

[3] For the Bayesian results reported in Experiments 1–4, the null hypothesis is compared to the alternative hypothesis that differences exist between or among the training conditions.

**Table 3** Mean performance on learning phase and posttest problems for subjects in the problem-solving and mixed study conditions in each experiment

|  | First problem | Second problem | Posttest |
|---|---|---|---|
| **Experiment 1** |  |  |  |
| Problem-Solving | .345 (.040) | .420 (.051) | .438 (.049) |
| **Experiment 2** |  |  |  |
| Problem-Solving | .356 (.047) | .499 (.044) | .491 (.045) |
| **Experiment 3** |  |  |  |
| Problem-Solving | .362 (.040) | .368 (.049) | .404 (.048) |
| Mixed Study | ---------- | .373 (.048) | .494 (.050) |
| **Experiment 4** |  |  |  |
| Problem-Solving | .287 (.034) | .240 (.030) | .419 (.035) |
| Mixed Study | ---------- | .344 (.033) | .387 (.032) |
| **Experiments 1–4** |  |  |  |
| Problem-Solving | .327 (.02) | .356 (.021) | .436 (.021) |
| Mixed Study | ---------- | .354 (.032) | .423 (.032) |

Standard errors of the mean are shown in parentheses. Experiments 1–4 find suggestive evidence that subjects who engaged in problem solving during training improved their performance as they progressed through the experiment. For subjects in the problem-solving conditions, this pattern was statistically reliable in Experiments 2 and 4 (both $ps < .023$), but did not reach statistical significance in the other experiments (all $ps > .051$). The results are less noisy when the data from Experiments 1–4 are combined: Subjects in the problem-solving conditions performed better on the posttest than on the first and second problems in the learning phase (both $ps < .004$, both $ds > 0.225$) and subjects in the mixed study conditions performed better on the posttest than on the learning phase problem ($p = .035$, $d = 0.161$)

# Experiment 2

In Experiment 1, the posttest occurred 1 week after the learning phase. One possibility is that differences in learning exist between the training conditions, but that information was forgotten over the delay, which obscured differences in posttest performance. Thus, we conducted a second experiment, identical to Experiment 1, except that the posttest was immediate.

## Methods

### Subjects and procedure

One hundred fifteen undergraduate students from ISU participated in this experiment for course credit in an introductory psychology course. Aside from replacing the delayed posttest with an *immediate* posttest, the design, materials, and procedures were identical to Experiment 1. Subjects took approximately 11 min to complete the learning phase ($M = 10.62$ min; see second row of Table 1). After completing the learning phase, all subjects were notified that they would be shown one more scenario and would be required to solve it. Subjects were then given an immediate posttest, which consisted of a novel problem scenario.

**Table 4** Adjusted (for time on learning phase) and unadjusted mean performance on the posttest (and repeated problems), partitioned by condition, in each experiment

|  | Example study | Problem-solving | Mixed | Control |
|---|---|---|---|---|
| **Adjusted mean** |  |  |  |  |
| Experiment 1 | .466 (.048) | .455 (.048) | ---------- | ---------- |
| Experiment 2 | .608 (.046) | .468 (.045) | ---------- | ---------- |
| Experiment 3 | .472 (.048) | .361 (.048) | .482 (.046) |  |
| **Experiment 4** |  |  |  |  |
| Posttest | .430 (.033) | .421 (.034) | .387 (.033) | ---------- |
| Repeated | .528 (.029) | .611 (.029) | .625 (.028) | ---------- |
| Experiments 1-4 | .472 (.021) | .436 (.022) | .423 (.027) | ---------- |
| **Unadjusted mean** |  |  |  |  |
| Experiment 1 | .483 (.049) | .438 (.048) | ---------- | ---------- |
| Experiment 2 | .602 (.045) | .491 (.045) | ---------- | ---------- |
| Experiment 3 | .419 (.043) | .404 (.044) | .494 (.044) | .231 (.044) |
| **Experiment 4** |  |  |  |  |
| Posttest | .426 (.033) | .424 (.034) | .387 (.033) | ---------- |
| Repeated | .523 (.029) | .616 (.029) | .626 (.029) | ---------- |
| Experiments 1–4 | .471 (.021) | .438 (.022) | .423 (.027) | ---------- |

Standard errors of the mean are shown in parentheses

## Results and discussion

An ANCOVA (with learning phase time as a covariate) revealed no reliable posttest differences between conditions, $F(1,112) = 3.570$, $p = .062$, $MSE = .117$, $\eta_p^2 = .031$. A follow-up Bayesian version of this ANCOVA found support for the null hypothesis ($BF = .278$), as the results were 3.60 times more likely to occur under the null model. These findings replicate the results from Experiment 1 and demonstrate that when the training conditions are carefully controlled, both produce similar levels of transfer on an immediate posttest.

## Experiment 3

One possibility is that the results from Experiments 1 and 2 were due to floor effects, wherein subjects failed to learn enough during the learning phase to solve novel, analogous problems. Another possibility is that for problem-solving practice to be effective, learners must have enough knowledge about the problem type before engaging in problem solving (see Corral et al., 2020). Subjects in the problem-solving practice condition were only presented one example to study, which may have been insufficient for them to learn enough about the problem type to benefit from problem-solving practice.

Thus, in addition to the problem-solving practice and example study conditions, Experiment 3 included a control condition, wherein subjects attempted to solve each of the four problems, but were not shown any examples nor provided feedback. This condition provides a baseline measure of how well subjects can solve these problems without any training, and thus allows for a direct assessment of whether the training conditions produce sufficient learning to facilitate the transfer of knowledge to analogous, novel scenarios. Furthermore, to allow subjects greater opportunity to acquire the necessary knowledge about the to-be-learned problem type before engaging in problem-solving practice, a mixed study condition was included, in which subjects studied two examples before attempting to solve a problem.

If learning occurs during training, subjects who receive training should outperform control subjects on the posttest. Furthermore, if subjects require more than one example to benefit from problem-solving practice, then perhaps studying a second example before engaging in problem solving is beneficial. If so, subjects in the mixed study condition should perform best on the posttest.

## Methods

### Subjects

Two hundred twenty-eight undergraduate students from ISU participated in this experiment for course credit in an introductory psychology course.

### Design, materials, and procedure

Subjects were randomly assigned to one of four conditions: (a) mixed study ($n = 58$), (b) problem-solving practice ($n = 56$), (c) example study ($n = 58$), and (d) control ($n = 56$). The example study and problem-solving practice conditions were identical to those in Experiments 1 and 2.

In the mixed study condition, after subjects were presented with the first example for study, subjects were shown a second example (as in the example study condition). Next, these subjects were presented with a third problem and were asked to solve it. After entering a response, these subjects were shown the correct solution and were asked to study it (as in the problem-solving practice condition). Subjects took approximately 7 min to complete the learning phase ($M = 7.45$ min; see third row of Table 1).

In the control condition, subjects were asked to solve four problems (one at a time); no feedback was presented on any of these problems. All other procedures and materials were identical to Experiment 2.

## Results and discussion

### Control versus training conditions

For problems the control subjects completed, a repeated-measures ANOVA revealed no differences in performance between the first ($M = .208$, $SE = .039$), second ($M = .220$, $SE = .044$, third ($M = .226$, $SE = .040$), and fourth problems ($M = .232$, $SE = .046$), $F(3,165) = 0.057$, $p = .982$, $MSE = 0.102$, $\eta_p^2 = .001$. For this reason, control subjects' mean performance on these problems was compared to training subjects' posttest performance.

An ANOVA revealed that subjects in the training conditions performed better on the posttest than control subjects, $F(3,224) = 6.380$, $p < .001$, $MSE = 0.110$, $\eta_p^2 = .079$; this outcome occurred for each of the training conditions (all $p$s $< .007$ and all $d$s $> 0.611$; see unadjusted means in Table 4). Thus, subjects in each training condition were able to learn and comprehend the problem solutions well enough to transfer this knowledge to novel, analogous problems.

### Training conditions

An ANCOVA (with time on the learning phase as a covariate) revealed no posttest differences among the training conditions, $F(2,168) = 1.952$, $p = .145$, $MSE = 0.123$, $\eta_p^2 = .023$ (see third row of Table 4). Moreover, a Bayesian version of this ANCOVA found support for the null hypothesis ($BF = .297$), as the results were 3.37 times more likely to occur under the null model. Given that subjects in the training conditions demonstrated better problem solving than control subjects, these findings indicate

that each of the training conditions facilitate learning and transfer, but do so to a comparable degree.

## Experiment 4

Experiments 1–3 show that subjects who engage in problem-solving practice do not transfer solutions to new problems better than example study subjects. Better memory of the content that learners practice retrieving is arguably the primary mechanism through which retrieval practice benefits learning (Butler et al., 2017; Pan & Rickard, 2018). Thus, if subjects who engage in problem-solving practice do not have better memory of the problem's solution strategy than example study subjects, it would highlight a critical limitation in using retrieval practice to improve problem solving. Experiments 1–3 do not reveal whether problem-solving practice fails to benefit memory for the problem solution, application of that solution to a new problem, or both.

We therefore conducted a fourth experiment, which was similar to Experiment 3, but did not include a control condition. After the posttest, subjects were asked to solve the problems they were presented with during the learning phase (i.e., repeated problems). If problem-solving practice benefits memory of solution strategies, subjects in the problem-solving practice and mixed study conditions should outperform example study subjects on repeated problems.

## Methods

### Subjects

Three hundred forty-one undergraduate students from Syracuse University participated in this experiment for course credit in an introductory psychology course. This experiment was approved by the Institutional Review Board at Syracuse University. The experiment site was changed from the previous experiments because the first author changed institutions, which afforded the opportunity to increase our external validity and extend the findings from the previous experiments to a different and more diverse population. Approximately 61% of students who attend Syracuse University are 21 years of age or younger, 54% identify as female, and approximately 56% are White.

To decrease the chance of committing a type two error, the present experiment adopted a more conservative a priori power analysis than Experiments 1–3. We based the sample size for Experiment 4 on an a priori power analysis with at least 90% power to detect a small-medium effect size ($f = .20$; $\alpha = .05$).

### Design, materials, and procedure

Subjects were randomly assigned to one of three conditions: (a) example study ($n = 116$), (b) problem-solving practice ($n = 109$), and (c) mixed study ($n = 116$). Apart from not including a control condition, through the posttest, the design and procedure in Experiment 4 were identical to those of Experiment 3. Subjects took approximately 10 min to complete the learning phase ($M = 10.34$ min; see fourth row of Table 1).

After the posttest, all subjects were asked to solve the same problem scenarios that they were presented with during the learning phase. The order in which these problems were presented was randomized for each subject. No feedback was presented on repeated problems.

Lastly, to get a better sense of subjects' knowledge about the problems, we asked them two supplementary questions: (a) whether they recognized any connection or commonality among the problem scenarios, and if so, what it was (structural similarity question), and (b) whether they used a common rule or solution for solving the problems, and if so, what it was (solution similarity question). For each of these questions, subjects were asked to type their response into a textbox, which was located directly beneath the question.

**Scoring of similarity questions** Both structural and solution similarity questions were scored on a scale of 0–4. As noted earlier, the problem scenarios consist of the same *relational structure*, such that there are (a) sufficient resources/forces to reach the goal state, but (b) doing so requires concentrating all available resources on a given point, and (c) there is an obstacle that prevents this from happening. Accordingly, these problems can be solved using the same solution strategy, as the (a) resources/forces in the scenario must be partitioned and (b) sent along different routes and then (c) converge simultaneously on the target. Thus, there are three shared structural components in the problem scenarios and three corresponding shared components in the solution strategy.
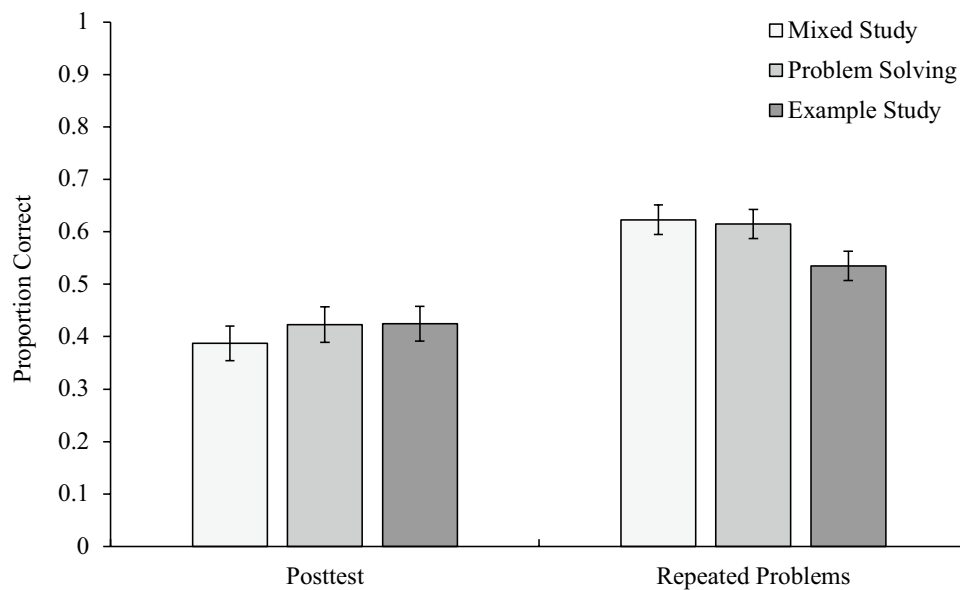
*Structural similarity questions.* For the structural similarity questions, subjects received a 0 if they reported not recognizing any commonalities among the problem scenarios; subjects received a 1 if they reported noticing a commonality, but did not note any of the problems' structural or solution components, a 2 if they noted one of the problems' structural or solution components, a 3 if they noted two of the problems' structural or solution components, and a 4 if they noted three of the problems' structural or solution components.

*Solution similarity questions.* For the solution similarity questions, subjects received a 0 if they reported not using any similar or common solution to solve the problems;

**Table 5** Mean scores on structural and solution similarity questions partitioned by training condition in Experiment 4

|  | Example study | Problem-Solving | Mixed |
|---|---|---|---|
| Structural Similarity | 1.73 (.121) | 1.61 (.124) | 1.63 (.121) |
| Solution Similarity | 1.91 (.130) | 1.95 (.134) | 1.91 (.130) |

Standard errors of the mean are shown in parentheses. Scores on both questions were positively related with performance on the posttest and repeated problems (all βs > .218 and all ps < .001), as subjects with higher scores performed better on the posttest and repeated problems. However, no differences were observed among the training conditions on either question (both ps > .732). For both questions, follow-up Bayesian ANCOVAs (with learning phase time as a covariate) revealed strong support for the null hypothesis (both BFs < .044)

repeated problems). A reliable interaction was observed between the training conditions and the test phase, $F(2,337) = 5.29$, $p = .005$, $MSE = 0.056$, $\eta_p^2 = .030$, such that differences in performance among conditions depended on test phase.

Specifically, no performance differences were observed on the posttest, $F(2,337) = 0.476$, $p = .622$, $MSE = 0.128$, $\eta_p^2 = .003$. Additionally, a follow-up Bayesian ANCOVA (as in Experiment 3) revealed very strong evidence for the null hypothesis ($BF = .014$), as the posttest results were 73.39 times more likely to occur under the null model.

However, performance differences *did* emerge on repeated problems, $F(2,337) = 3.36$, $p = .036$, $MSE = 0.093$, $\eta_p^2 = .020$, as subjects in the mixed study and problem-



**Fig. 2** Mean performance and standard errors of the mean on the posttest and repeated problems for each condition in Experiment 4

subjects received a 1 if they reported using a common or similar solution, but did not note any of the problems' solution components, a 2 if they noted one of the problem's solution components, a 3 if they noted two of the problem's solution components, and a 4 if they noted three of the problems' solution components. Table 5 includes mean scores on both similarity questions partitioned by training condition.

## Results and discussion

Figure 2 shows mean performance on the posttest and repeated problems for each condition. To examine performance differences among the training conditions, we conducted a mixed ANCOVA, with training condition as a between-subjects factor (mixed study vs. problem-solving practice vs. example study), learning phase time as a covariate, and test phase as a within-subjects factor (posttest vs.

solving practice conditions performed better on repeated problems than subjects in the example study condition (both ps < .050).

Thus, subjects who engaged in problem solving during learning (i.e., mixed study and problem-solving practice conditions) were better able to retrieve and apply the solutions to those same problems later than subjects in the example study condition. Nevertheless, this benefit did not lead to better transfer on the posttest.

## General discussion

Across four experiments (also see the combined analysis in the Appendix), subjects in the training conditions performed comparably on the posttest. Experiment 3 suggests that this

finding is due to the training conditions producing similar levels of transfer, as subjects in all of the training conditions outperformed control subjects on the posttest. This finding indicates that the training conditions facilitated the transfer of learning (otherwise posttest differences between the training and control conditions should not have emerged). Further support for this conclusion comes from subjects in the problem-solving practice and mixed study conditions generally performing better on novel problems as the study progressed (see Table 3). Thus, the null findings reported among the training conditions were the result of these subjects being able to transfer what they learned to a comparable degree.

Furthermore, Experiment 4 showed that subjects who engaged in problem-solving practice performed better than subjects who engaged in example study on repeated problems. This outcome can be thought of as a type of testing effect, as subjects who had the opportunity to practice retrieving solution strategies during training were better able to recall those solutions than subjects who studied examples. This result is similar to recent data showing that problem-solving practice benefits subsequent performance on identical problems (Yeo & Fazio, 2019). Critically, however, this superior memory was not enough to produce differential posttest transfer among the training conditions.

These results are in line with recent work suggesting that memory of to-be-learned material is necessary, but insufficient for the transfer of learning (Butler et al., 2017). Memory alone, however, does not facilitate transfer unless learners recognize the relevance of the learned information in the current situation.

Indeed, in Experiment 4, although problem-solving practice and mixed study subjects had the best memory for the solution strategies, the similarity questions revealed that there were no condition differences in subjects' recognizing the relevance of these solution strategies across different problem scenarios (see Table 5). These findings might explain why transfer more often occurs under conditions where learners are provided with hints about the relevance of learned information to a current situation (Barnett & Ceci, 2002; Butler, 2010; Gick & Holyoak, 1983), as learners often do not recognize that their prior knowledge is applicable.

Thus, even when learners have superior knowledge of a problem's solution strategy, this knowledge is not enough to facilitate the transfer of learning. This takeaway highlights that other aspects of problem solving might be particularly important for facilitating transfer (e.g., recognizing a problem type's structure; see Corral & Kurtz, 2023). Accordingly, work in mathematics suggests that students do not struggle to learn a problem's solution strategy or in how to use it, but rather in recognizing *when* to apply it (Mayer, 1998). Therefore, although retrieval practice can aid learners' knowledge of solution strategies, our findings suggest that additional training is needed to facilitate learners' application of that knowledge to new situations.

These findings offer important new insights into the mechanistic relationship between retrieval practice and the transfer of learning, and how the former impacts the latter. Retrieval practice seems to provide a greater benefit than studying to the learning and memory component that is involved in the transfer of learning. However, one limitation is that this benefit might be restricted to the information that learners attempt to retrieve (e.g., solution strategies). Indeed, when compared to example study, retrieval practice does not appear to better improve the recognition component of knowledge transfer.

Critically, the recognition component of transfer seems to give rise to the inert knowledge problem (Snoddy & Kurtz, 2021), and has been posited to be the most central component in the successful transfer of learning (Corral & Kurtz, 2023). It is thus particularly noteworthy that retrieval (via problem-solving practice) does not appear to better improve the recognition of when to apply a corresponding solution strategy any more so than studying examples. One possibility is that for problem-solving practice to produce a greater benefit to transfer than example study, it must better aid learners' recognition of when to apply the corresponding solution strategy.

The present results differ from the multitude of studies showing advantages of retrieval practice over restudy (for reviews, see Agarwal et al., 2021; Pan & Rickard, 2018). It is worth noting, however, that most studies on retrieval practice are based on tasks involving recall of information from memory (which we also find evidence for in Experiment 4); tasks involving transfer of learning – and in particular transfer of a solution strategy to novel scenarios – have not been thoroughly explored with retrieval practice (see Carpenter et al., 2020).

The current results thus contribute critical new data showing that retrieval practice benefits memory, but not necessarily transfer. These findings point to potential boundary conditions and limitations to the benefits that retrieval practice provides. These takeaways highlight an often-overlooked distinction by theoreticians between memory and the transfer of learning. Indeed, theories on retrieval practice primarily focus on how retrieval engages mechanisms that strengthen memory of the information that is retrieved (Carpenter et al., 2022; Pan & Rickard, 2018). However, these theories do not directly explain how this enhanced memory might facilitate the transfer of learning, nor (and perhaps more importantly) how it might aid learners in recognizing when to apply their corresponding knowledge. The present findings thus call for theoreticians to consider more careful and nuanced hypotheses on the relationship between retrieval practice and the transfer of learning.

# Appendix

## Combined analyses (Experiments 1–4)

To further increase confidence in our conclusions, we conducted a set of supplementary analyses, in which we pooled the data from Experiments 1–4 into a single dataset. In total, this combination led to 734 subjects: (a) mixed study ($n =$ 174), (b) problem-solving practice ($n = 277$), and (c) example study ($n = 283$).

We conducted a $3 \times 4$ ANCOVA with training condition (example study vs. problem-solving practice vs. mixed study) and experiment (Experiment 1 vs. Experiment 2 vs. Experiment 3 vs. Experiment 4) as between-subject factors, time on the learning phase as a covariate, and posttest performance as the dependent measure. This analysis revealed no differences in performance among the training conditions on the posttest (see Table 4 for condition means), $F(2, 723) = 1.16$, $p = .315$, $MSE = 0.127$, $\eta_p^2 = .003$.[4] Moreover, a Bayesian version of this ANCOVA found strong support for the null hypothesis ($BF = .041$), as these findings were 24.10 times more likely to occur under the null model. These findings therefore confirm the patterns from the four experiments and suggest that the three training conditions seem to produce similar amounts of learning and transfer during problem solving.

# References

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review, 33*, 1409–1453.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*, 35–44. https://doi.org/10.1037/0003-066X.48.1.35

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

Benassi, V., Overson, C., & Hakala, C. (Eds.). (2014). *Applying science of learning in education: Infusing psychological science into the curriculum*. American Psychological Association.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1. From learning theory to connectionist theory; Vol. 2. From learning processes to cognitive processes* (pp. 35–67). Lawrence Erlbaum Associates Inc.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied, 23*, 433–446.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*, 632–642. https://doi.org/10.3758/bf03202713

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1563–1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributed to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1547–1552.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283.

Carpenter, S. K., Endres, T., & Hui, L. (2020). Students' use of retrieval in self-regulated learning: Implications for monitoring and regulating effortful learning experiences. *Educational Psychology Review, 32*, 1029–1054. https://doi.org/10.1007/s10648-020-09562-w

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology, 1*, 496–511.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1147–1156.

Corral, D., & Carpenter, S. K. (2020). Facilitating transfer through incorrect examples and explanatory feedback. *Quarterly Journal of Experimental Psychology, 73*, 1340–1359.

Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition, 132*, 280–300.

Corral, D., & Kurtz, K. (2023). *Improving the transfer of learning in education through category learning*. Manuscript submitted for publication.

Corral, D., Quilici, J. L., & Rutchick, A. M. (2020). The effects of early schema acquisition on mathematical problem solving. *Psychological Research, 84*, 1495–1506.

Corral, D., Carpenter, S. K., & Clingan-Siverly, S. (2021). The effects of delayed versus immediate explanatory feedback on complex concept learning. *Quarterly Journal of Experimental Psychology, 74*, 786–799.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347–362.

Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs, 58*, i–113. https://doi.org/10.1037/h0093599

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

---

[4] A main effect of experiment was observed, $F(3, 723) = 4.07$, $p = .007$, $MSE = .127$, $\eta_p^2 = .017$, but no interactions occurred between the training conditions and the experiments, $F(4, 723) = 1.19$, $p = .314$, $MSE = .127$, $\eta_p^2 = .007$.

educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.

Hostetter, A. B., Penix, E. A., Norman, M. Z., Batsell, W. R., & Carr, T. H. (2018). The role of retrieval practice in memory and analogical problem-solving. *Quarterly Journal of Experimental Psychology, 72*, 858–871. https://doi.org/10.1177/1747021818771928

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–46). Academic Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15*, 332–340. https://doi.org/10.3758/BF03197035

Mayer, R. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science, 26*, 49–63. https://doi.org/10.1023/a:1003088013286

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*, 710–756.

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory, 26*, 1460–1466.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629–639. https://doi.org/10.1037/0278-7393.13.4.629

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 456–468. https://doi.org/10.1037/0278-7393.15.3.456

Snoddy, S., & Kurtz, K. J. (2021). Preventing inert knowledge: Category status promotes spontaneous structure-based retrieval of prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*, 571–607. https://doi.org/10.1037/xlm0000974

Sweller, J., & Cooper, G. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89.

Sweller, J., Mawer, R., & Ward, M. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 112*, 639–661.

Whitehead, A. N. (1929). *The aims of education and other essays*. Macmillan.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science, 36*, 1–10.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218.

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*, 73–90. https://doi.org/10.1037/edu0000268