Journal of Memory and Language 129 (2023) 104402

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Retrieval practice and verbal-visuospatial transfer: From memorization to inductive learning

Gregory I. Hughes [*,1], Ayanna K. Thomas

Tufts University, Medford, MA, United States

ARTICLE INFO

Keywords:
Categorization
Retrieval practice
Transfer
Testing effect
Induction

ABSTRACT

Retrieval practice, the act of recalling information on a practice test, leads to better long-term memory than non-testing study activities (*the testing effect*). This effect occurs even when the contexts of the practice and final test differ, suggesting that retrieval practice fosters transferable learning. For example, practice tests involving the recall of targets (A-?) not only enhance performance on final tests of the targets (A-?), but this effect can also extend to tests of the non-recalled cues (?-B). Simple memory tests can also facilitate the inference of underlying rules or principles that can be used to answer completely new questions or problems. However, these transfer effects have been overwhelmingly demonstrated with verbal materials. Further, research suggests that transfer effects diminish as the type of information tested during the practice and final tests diverge. In the present study, we explored the influence of retrieval practice on paired associates consisting of the names and visuospatial diagrams of molecules. In two experiments using a standard paired-associate learning paradigm, practice tests of name targets (?-diagram) or diagram targets (name-?) did not enhance performance on final tests of their respective cues. In a final experiment using a category induction paradigm, we found a benefit of retrieval practice on the memorization of cues and the induction of underlying rules simultaneously.

Introduction

Retrieval practice, the act of recalling information from memory, is not simply a way of assessing the contents of the mind but is also a potent learning event. However, the nature of this learning event remains elusive. One way to shed light on the subject is to explore the kinds of learning that occur during a cued-recall test, the most commonly investigated form of retrieval practice (see Pan & Rickard, 2018). The typical finding is that, compared to an equivalent amount of time studying information, taking an initial cued-recall test (cue - ?) yields higher performance on later final tests of the targets (cue - ?; for a review, see Adesope et al., 2017). This is called the *testing effect*. Research increasingly suggests that the benefits of retrieval practice are not confined to the specific items that are recalled, but transfers to other, related information. In the present study, we investigate how properties of the stimuli influence these phenomena.

Studies on retrieval practice use a common paradigm. During an initial encoding phase, participants learn material through a non-testing study activity and then review the material again in an identical fashion (*restudy* or *study practice*) or by taking memory tests (retrieval practice). Participants then take a final memory test, which is used to compare performance as a function of the prior encoding activity. To ensure a fair comparison, experimenters equate the total time that participants spend during each encoding activity.

Two criticisms are commonly made about the literature on retrieval practice. One criticism is that the vast majority of studies have participants encode and retrieve verbal materials, like word pairs (Bulevich et al., 2016; Carpenter et al., 2006; Hughes et al., 2018; Tullis et al., 2013), simple facts (Pan, Gopal, et al., 2016), and text passages (McDaniel et al., 2009; McDaniel et al., 2015; Roediger & Karpicke, 2006). Only a handful of studies had participants produce non-verbal information during tests, like the visuospatial arrangement of landmarks on maps (Carpenter & Pashler, 2007; Rohrer et al., 2010), an array of objects (Carpenter & Kelly, 2012), and the strokes that compose

* Corresponding author.
E-mail address: gregory.hughes@tufts.edu (G.I. Hughes).
[1] Gregory I. Hughes is now at the U.S. Army Combat Capabilities Development Command Soldier Center (DEVCOM SC), Natick, MA, USA, and the Center for Applied Brain and Cognitive Sciences (ORCID: 0000-0002-0010-0613); Ayanna K. Thomas, Department of Psychology, Tufts University (ORCID: 0000-0001-8173-5911). This study was not preregistered. Correspondence concerning this article should be addressed to Gregory Hughes, the U.S Army Combat Capabilities Development Command (DEVCOM SC) Soldier Center, 15 General Greene Ave, Natick, MA 01760.

https://doi.org/10.1016/j.jml.2022.104402
Received 29 April 2022; Received in revised form 14 December 2022; Accepted 19 December 2022
Available online 29 December 2022
0749-596X/Published by Elsevier Inc.

Chinese characters (with non-Chinese speakers; Kang, 2010). Another criticism of the literature on retrieval practice concerns the transfer of learning from one context to another. Overwhelmingly, in studies on retrieval practice, the initial and final tests are completely identical (for a discussion, see Pan & Rickard, 2018). In other words, apart from the passage of time, the contexts of the acquisition and assessment of knowledge are the same, and no substantive transfer can be demonstrated. In the present study, we explored how retrieval practice affects learning and transfer with visuospatial materials.

We focused on two kinds of transfer effects. First, we investigated the types of verbatim memorization that an initial cued-recall test can foster. A handful of studies have demonstrated what we term *the backward transfer effect* (see Fig. 1), the finding that initial tests of targets (cue - ?) enhances performance on not only on later tests of those targets (cue - ?) compared to restudy, but also tests of the *cues* (? - target) (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020). However, to our best knowledge, the backward transfer effect has only been demonstrated with simple, native language word pairs (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020) and with native-foreign language word pairs (Barenberg et al., 2021). Empirical (Bulevich et al., 2016; Rickard & Pan, 2020) and theoretical evidence (Blaxton, 1989; Morris et al., 1977) suggests that the effect will not always occur, such as when the members of pairs come from different classes of stimuli (e.g., verbal - visuospatial pairs)(Litt & Nation, 2014; Molander & Garvill, 1979). Second, we examined the possibility that the advantage of an initial cued-recall test is not constrained to the memorization of studied items (e.g., cues and targets), but also fosters the induction of underlying rules from examples (e.g., category learning). Several studies show that initial tests of the category name of studied items (exemplar - ?) enhances inductive learning, as measured by a transfer test that requires the categorization of new exemplars. However, it is unclear how effectively retrieval practice can influence verbatim item memorization and inductive learning simultaneously.

Memorization and backward transfer

We are aware of four studies that have investigated the backward transfer effect with paired associates (Barenberg et al., 2021; Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020). Three of these studies used simple, native language word pairs. In two experiments, Carpenter et al. (2006) had participants learn native-language word pairs through repeated study or cued recall of the targets. On final cued-recall tests of the cues, retrieval practice outperformed restudy by 14% (Exp 1) and 9% (Exp 2). Rickard and Pan (2020) found that retrieval practice led to an equivalent benefit on cues and targets with a 24-hour interval. They also replicated the backward transfer effect with a 1-week retention interval, although the benefit of retrieval practice was stronger for targets than cues. Similarly, Cheng (2014) observed a backward transfer effect, but used an initial multiple-choice test and a final cued-recall test. Using a population of native German speakers, Barenberg et al. (2021) found that with German-English pairs, initial tests of English targets (German - ?) led to superior performance on backward tests of the English cue (? - English) compared to a restudy condition ($M = .62$ and .54, respectively). It is important to note that although many other studies have investigated the influence of retrieval practice of targets on final tests of targets and cues, these did not include a study-practice control

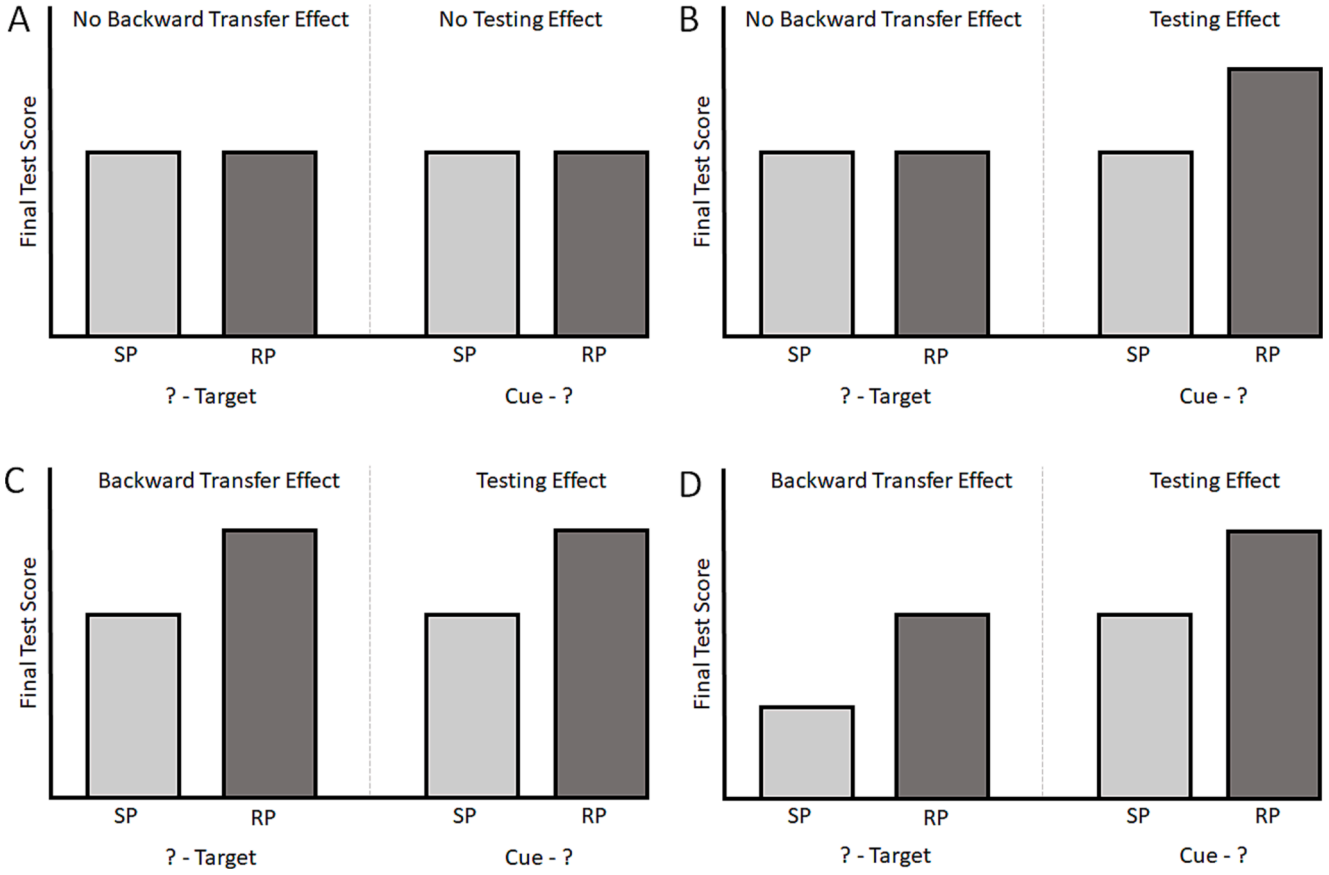

Fig. 1. *The Testing Effect and Backward Transfer Note.* Illustrative results comparing the effects of learning through study practice (SP) or retrieval practice (RP) of targets (cue - ?) on final tests of targets (cue - ?) and cues (? - target). (A) No benefit of retrieval practice on final tests of targets (no testing effect) or cues (no backward transfer effect). (B) Retrieval practice enhanced the learning of targets (testing effect), but not cues (no backward transfer effect). (C) Retrieval practice enhanced learning of targets and cues equally (both effects). (D) Retrieval practice enhanced the learning of targets more than cues, but outperformed study practice on both measures (both effects).

condition (for a review, see Kahana, 2002). Consequently, they did not investigate the standard testing effect or the backward transfer effect, which can only be demonstrated by comparing performance on criterial memory measures to a restudy condition (see Fig. 1).

Research suggests that the backward transfer effect is sensitive to the nature of the materials. For example, although the effect has been demonstrated several times with simple word pairs, it is attenuated or abolished with word triplets (Bulevich et al., 2016; Rickard & Pan, 2020; Pan, Wong, et al., 2016), or longer and more conceptually-meaningful materials, like multi-term facts (Hinze & Wiley, 2011; Pan, Gopal, et al., 2016; but see McDaniel, Anderson, Derbish, & Morrisette, 2007) and term-definition facts (McDaniel et al., 2015; Pan & Rickard, 2017). These findings have led some researchers to speculate that there is something unique about paired associates that enables backward transfer relative to longer lists of associates (Kahana, 2002; Rickard & Pan, 2018).

Even with paired associates, there is reason to expect that the backward transfer effect may not occur when properties of cues and targets substantially differ. According to the principle of transfer-appropriate processing, performance on a final test depends on the degree of overlap between the processes that take place during encoding and retrieval (Blaxton, 1989; Morris et al., 1977). With word pairs, there is likely to be significant overlap because participants are asked to retrieve verbal information during initial (word cue - ?) and final tests (? - word target). Furthermore, both words are individually well-known to participants before they arrive at the laboratory. An extensive network of pre-existing semantic knowledge can make forming flexible associations easier, either by enabling the deliberate use of strategies like mental imagery and sentence generation (Mondani & Battig, 1973; Paivio, 1971), or by automatically strengthening indirect cue-target links through spreading activation (Carpenter, 2011).

Further, some studies suggest that perfect transfer is less likely to occur when cues and targets involve processing or producing different types of information during initial and final tests, such as with visual-aural pairs (Litt & Nation, 2014) and visual-tactile pairs (Molander & Garvill, 1979). That is, an initial test of one type of target preferentially enhanced the learning of that target rather than its cue. However, it is important to emphasize that these cross-modal studies did not include a restudy control condition. Even if an initial test of targets preferentially enhanced the learning of its target compared to its cue, it could still have enhanced cue learning relative to a restudy condition (see Fig. 1D), which was not assessed in these studies. Indeed, the results of Barenberg et al. (2021) suggest that the backward transfer effect can occur when cues and targets differ qualitatively, since German and English words are unlikely to be processed identically by native German speakers who are learning a second language.

We expanded on the prior literature by exploring the backward-transfer effect with pairs consisting of a verbal and visuospatial member. Participants learned the names of molecules (verbal) and diagrams of their atomic structure (visuospatial; name-diagram pairs). Of interest was whether initial tests of diagrams (name - ?) enhanced performance on final tests of names (? - diagram), and vice versa, compared to restudy. We also limited participants' prior expertise in chemistry by excluding those who had taken any chemistry course beyond the standard high school level. Compared to studies on simple word pairs, flexible memorization could be doubly challenged both by the comparative lack of pre-existing knowledge and qualitative differences in the cues and targets.

Category induction and transfer

As we have discussed, there are at least two ways that the backward transfer effect could be reduced. One way is when the processes of retrieving cues and targets are sufficiently different. Another way is when participants lack sufficient pre-existing knowledge structures that can be used to link cues and targets flexibly. It is therefore conceivable that with novel, heterogenous pairs (e.g., verbal - visuospatial), interventions that help scaffold new networks of conceptual or associative-knowledge structures would make the backward transfer effect more likely to occur. In other words, with novel or less familiar materials, building new networks of knowledge could increase the size of a backward-transfer effect.

One way to foster new associative-knowledge structures is through a category induction task. In experiments on category induction, participants study cue-target pairs, which consist of a category name and an exemplar. Unlike typical paired associate paradigms, the same category name is associated with multiple exemplars, which are different from one another in some ways, but not others. By observing the features that repeat across a given category's exemplars, participants can infer the rules that govern category membership. For example, participants could notice that whereas all exemplars of arachnids have eight legs, all exemplars of insects have six, and correctly infer this as a categorization rule. Ultimately, the job of a participant in these experiments is to infer these rules, such that they can categorize novel exemplars on a later test (e.g., an un-studied arachnid). In other words, their learning must transcend simple memorization of the studied exemplars, which is inadequate for mastering performance on a later transfer test with new exemplars. Essentially, the category induction process yields new knowledge that links cues and targets together bidirectionally (cue ← rules → target).

Several studies demonstrate that retrieval practice can foster visual category induction relative to restudy (Cho & Powers, 2019; Jacoby et al., 2010; Yang & Shanks, 2018; Yang et al., 2019). For example, Jacoby et al. (2010) had participants study categories of birds through restudy or multiple-choice retrieval practice (exemplar – category A, B, or C?) with correct-answer feedback. In three experiments, retrieval practice led to greater performance on a multiple-choice categorization test. The effect occurred with previously studied items and, more importantly, un-studied items, which demonstrates the acquisition of categorical knowledge. This effect has been replicated in studies that used an initial cued-recall test, both with exemplars of painters (Lee & Ahn, 2018; Yang & Shanks, 2018) and categories of Chinese characters (Cho & Powers, 2019).

Note that the studies on retrieval practice and category induction were not expressly concerned with the verbatim memorization of exemplars. In these studies, retrieval practice involved presenting an image of a studied exemplar and asking participants to recall or recognize its associated label. This approach makes sense, because the explicit goal of inductive learning is not to memorize exemplars, but to isolate the subset of exemplar features that are always associated with one category label, but not the others. However, this approach contrasts with the majority of studies on retrieval practice, which are expressly concerned with verbatim memorization and often require the production of entire items on initial and final tests. Consequently, these studies do not address how effectively retrieval practice can foster verbatim exemplar memorization and inductive learning simultaneously.

It is unclear how retrieving entire exemplars during initial tests would influence category induction. On the one hand, verbatim memorization of exemplars has been assumed to be detrimental to inductive learning. This is because research suggests that category induction is maximized when people selectively focus on the features that are relevant to the categorization rules and ignore those that are irrelevant (Carvalho & Goldstone, 2014; Goldstone, 1996; Lancaster et al., 2013; Zulkiply & Burt, 2013; for a review, see Hughes & Thomas, 2021). In contrast to that aim, verbatim memorization requires attention to the irrelevant features as well. On the other hand, exemplar models of categorization regard item memorization as the primary, or even sole, determinant of categorization learning and judgments (see, Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). According to these models, categorization involves comparing novel exemplars to those that are stored in memory. During the learning process, the presentation of one exemplar could prompt the recall of a previously studied

exemplar (i.e., *study-phase retrieval*; Greene, 1989), which could help participants isolate the category-relevant features through comparison processes. From this perspective, it stands to reason that enhancing the encoding of exemplars would likewise profit performance on later categorization tests with novel exemplars.

We addressed these questions in Experiment 3 by having participants learn paired associates belonging to categories of organic chemistry molecules. In contrast to the prior literature, we explicitly asked participants to recall items from memory verbatim (either the name of a molecule or an entire diagram). We then compared performance on tests of verbatim memorization and category induction as a function of initial encoding group.

Overview of the present study

In three experiments, we had participants with limited chemistry experience learn the names of molecules (verbal) and diagrams of their atomic structure (visuospatial; name-diagram pairs). In the first two experiments, we used a standard paired-associate paradigm, in which participants studied the pairs once by rote copying, and then either restudied those pairs in the same fashion three times (study practice) or by taking three cued-recall tests with feedback (retrieval practice). We fully crossed the type of initial and final tests, which allowed us to examine the backward transfer effect (e.g., name - ? → ? - diagram). In a final experiment, we explored backward transfer in the context of a category induction paradigm. Participants studied pairs in the same way as the prior experiments, but this time, the pairs belonged to multiple categories. The underlying category rules could be inferred by observing patterns across the set of studied items. On the final test, we included studied items to test memorization and new items, which could only be answered correctly if participants had inferred the rules.

Experiment 1

The purpose of Experiment 1 was to explore how retrieval practice influences the learning of molecular name-diagram pairs. In the study-practice group, participants copied these pairs by hand into a paper booklet four times. In the retrieval-practice group, participants copied these pairs by hand once, and then took three cued-recall tests. For half of the items, participants were tested on the diagram (name - ?) and for the other half, they were tested on the names (? - diagram). Responses were drawn or written into a test booklet, respectively. Feedback, in the form of the intact name-diagram pair, was provided after every trial because it maximizes the benefit of retrieval practice on long-term memory (e.g., Butler & Roediger, 2008).

In this experiment, we crossed the type of retrieval-practice question and the type of final-test question, which resulted in four conditions (see Fig. 2). In the two congruent trials, the types of retrieval-practice and final-test question were identical (e.g., name - ? and name - ?). In the two incongruent trials, the types of questions did not match (e.g., name - ? and ? - diagram). Performance on the two congruent trials allowed us to replicate the standard testing effect with the retrieval of verbal (diagram - ? / diagram - ?) (Bulevich et al., 2016; Carpenter et al., 2006; Hughes et al., 2018; Tullis et al., 2013) and visuospatial (name-? / name -?) (Carpenter & Pashler, 2007; Kang, 2010; Rohrer et al., 2010) associates. If we observed the same effects on the two incongruent trials (e.g., name -? / ? - diagram), then we would extend the prior work on backward transfer with word pairs materials (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020) to name-diagram pairs.

Data availability

The data, analysis materials, and stimuli are available through the Open Science Framework (https://osf.io/tkmv3/).

Method

Design

We used a 2 (encoding group: study practice, retrieval practice) × 2

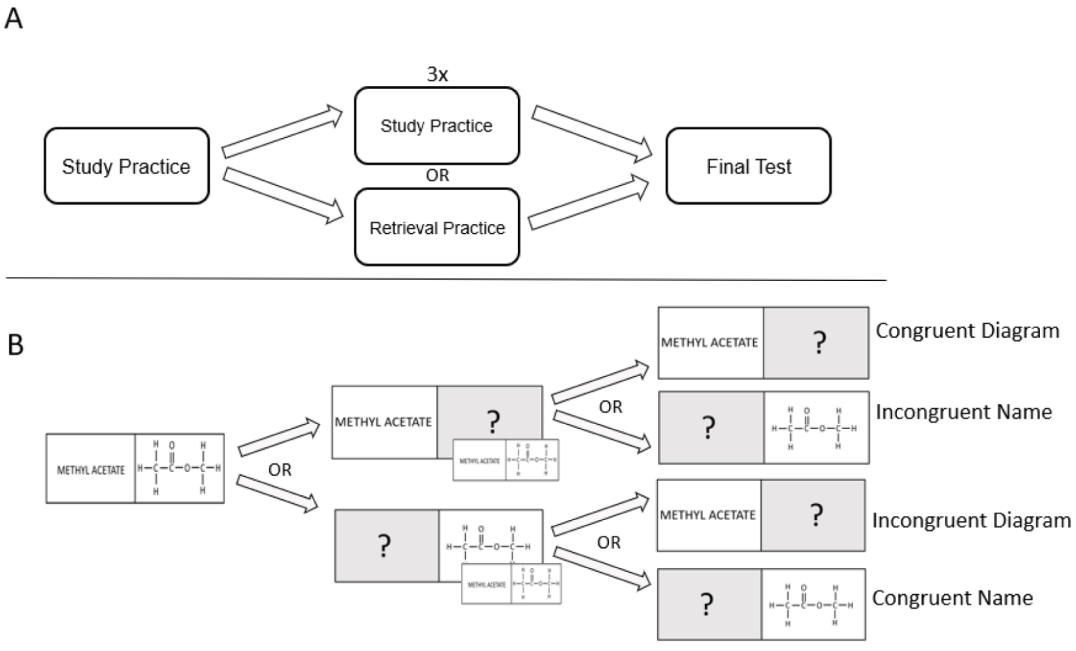

Fig. 2. *Design of Experiment 1 Note.* Schematic of initial learning and final test procedure. (A) Participants in both groups copied pairs by hand and then repeated this three times (study practice; SP) or took three cued-recall tests (retrieval practice; RP) with trial-by-trial feedback in the form of the entire item. The final test took place after 48 h. (B) Congruency manipulation. Across all three RP tests, a given pair was consistently tested on its name or diagram. During the final test, each pair was either tested on its name or diagram. For a given pair, the target (name or diagram) during RP was followed by a final test of that same target (congruent) or its cue (incongruent), yielding four conditions.

(target type on final test: name, diagram) × 2 (trial type: congruent, incongruent) mixed design. Learning technique was a between-subjects variable, and the other two variables were within-subjects.

Power analysis

To determine adequate sample size, we conducted an a priori power analysis. In a three-way mixed design, a power analysis can be conducted for any of the nested statistical tests. We conducted our power analysis to detect a between-subjects effect of encoding group (retrieval practice > study practice). We based this choice on the finding that, with paired associate learning, cued-recall retrieval practice (A-?) improves performance relative to study practice roughly to the same degree on forward (A-?) and backward (?-B) tests (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020). The average weighted effect size of these studies was medium-to-large ($\eta_p^2 = .117$). However, all of these studies required the production of words during tests. Therefore, we chose to use the smaller effect size reported by Kang (2010), who observed a testing effect when the recall of visuospatial information (Chinese characters) was required during retrieval practice and the final tests ($\eta_p^2 = .097$). We conservatively assumed a moderate-to-high correlation among repeated measures (.70), thereby increasing the number of participants needed to achieve the desired level of power (Winer et al., 1991). The analysis demonstrated that a minimum of 33 participants per encoding group would be needed to achieve adequate power to detect a between-subjects effect of retrieval practice at a power level of 80% ($\alpha = .05$). This sample size would also be sufficient to detect a small-to-medium interaction effect in which retrieval practice benefited performance congruent trials, but not on incongruent trials ($\eta_p^2 = .018$).

Participants

Eighty-three undergraduate students from Tufts University ($M_{age} = 20.46$, $SD_{age} = 1.65$) participated in the experiment. Participants were either compensated with course credit or $10 per hour and were invited to participate only if they had never taken chemistry beyond the standard high-school level. Participants were randomly assigned to the study (n = 41) and retrieval-practice encoding groups (n = 42).

Materials

Participants learned 16 molecular name-diagram pairs. The names averaged 9.89 letters in length ($SD = 2.63$). The diagrams averaged 6.44 atoms ($SD = 1.31$) and 6.19 bonds ($SD = 1.33$). Collapsed across atoms and bonds, each diagram contained 12.69 components ($SD = 2.60$). The name lures averaged 9.65 letters in length ($SD = 3.5$).

Procedure

Participants were run in groups of two to six. The experiment was programmed with the E-Prime software (Version 2.1; Schneider, Eschman, & Zuccolotto, 2002). The instructions and stimuli were projected onto a large screen. Research assistants ensured that all participants could read the instructions and view the stimuli clearly. We informed participants that the goal of the experiment was to learn the name-diagram pairs for a final memory test in 48 hr. Both experimental encoding groups began with a short practice session with two name-diagram pairs, which familiarized them with the encoding and testing procedures. During the practice of the final test, participants were asked to retrieve the name for one item and the diagram for the other item. Participants were told that for any given pair on the final test in 48 hr, they might be expected to retrieve the name, the diagram, or both. The presentation of items was randomized during all learning and testing activities.

Initial Learning. Participants studied all 24 name-diagram pairs. Each pair was presented, one at a time, for 30 s. Participants wrote/drew each name-diagram pair on a single page in a paper booklet. We instructed participants to use all 30 s to study a given pair and not to revisit any pairs from previous trials. After 30 s elapsed, a red screen flashed for 500 ms to signal the end of the trial and for participants to flip the page (preventing review of the previous item).

Study-Practice Group. Participants in the study-practice group learned the 24 name-diagram pairs, in an identical fashion, across three additional blocks. At the end of each of the four study-practice blocks, we collected their completed booklet and provided a new one.

Retrieval-Practice Group. Participants in the retrieval-practice group were told that they would be taking a test on the pairs they had just studied. They were told that they would be presented with either the name or diagram of a pair and would be asked to write/draw the corresponding member in their test booklet. Participants were encouraged to provide as much of an answer as they could remember, but to respond, "I don't know," if they could not remember the answer at all. On a given trial, participants were presented with the name or diagram for 20 s. After the 20 s elapsed, they were presented with the intact pair for 10 s as feedback and instructed not to correct their answers. A red screen flashed for 500 ms to signal the start of the next trial and turn the page in their booklet. Again, participants were asked to spend the entire time on the current trial and not to revisit any previously written answers. Participants repeated these practice tests in three additional blocks. At the end of each of the three retrieval-practice blocks, we collected their completed booklets and provided a new one.

For each participant, half of the pairs were assigned for diagram testing (name - ?) and the other half for name testing (? - diagram). A given pair was tested in the same way across blocks. That is, if a participant was tested on the diagram for one pair (name - ?), they were tested on the diagram again for that pair in the next two retrieval-practice blocks. Likewise, if they were tested on the name for that pair (? - diagram), they were again tested on the name for that pair two more times. We counterbalanced the assignment of the 24 items to each type of testing condition.

Results

We complemented our frequentist analyses with the Bayesian equivalent to those analyses. To foreshadow our results, the main motivation to include these analyses was to make direct inferences about null results in the frequentist tests of our hypotheses. Of interest is whether these results are true null effects and were not merely due to low power (Dienes, 2014). We report a Bayes Factors (BF_{10}) at the end of each frequentist test which is a ratio of the probability of obtaining the data given the alternative hypothesis relative to the null hypothesis. Values above 1 marshal evidence for the alternative hypothesis, and values below 1 provide evidence for the null hypothesis (Jeffreys, 1939, 1961; Lee & Wagenmakers, 2014).

We conducted our analyses with the JASP software (JASP Team, 2022) and followed analysis procedures recommended by van Doorn et al. (2021) and van den Bergh et al. (2022). We used a uniform prior for all of our Bayesian analyses. Deviating from a uniform prior requires rigorous justification that is anchored in the results of similar research designs (van Doorn et al., 2021), which we lacked. We also could not use the results of each experiment to modify the priors of each subsequent experiment due to significant changes in the design and/or materials. All Bayesian post-hoc tests corrected for multiple comparisons with the method reported by Westfall et al. (1997).

Scoring

Names and diagrams were only scored as correct if they were produced without any error. To account for the possibility that participants ignored the instructions and corrected some of their answers after viewing feedback during retrieval practice, we only scored responses

from the initial encoding phase as correct if no portion of it had been crossed out and/or replaced.

Initial learning

During the retrieval practice tests, performance increased across the three tests of the names (M = .36, .47, and .59, respectively) and diagrams (M = .27, .38, and .43, respectively). A 2 (target type: name, diagram) × 3 (test number: 1, 2, 3) ANOVA showed a main effect of target type, $F(1, 37) = 10.07$, $p = .003$, $\eta_p^2 = .214$, $BF_{10} = 12.94$. Retrieving names (M = .47) was easier than diagrams (M = .36). There was also a main effect of test number, $F(2, 74) = 29.26$, $p < .001$, $\eta_p^2 = .442$, $BF_{10} > 10{,}000$, confirming the significance of increased performance across the trials, but there was no interaction between target type and test number, $F(2, 74) = 1.25$, $p = .293$, $\eta_p^2 = .033$, $BF_{10} = 0.27$. Although retrieving the names was easier than the diagrams, the rate of learning for each type of target across tests did not statistically differ.

Final test

The results are depicted in Fig. 3. We conducted a 2 (encoding group: study practice, retrieval practice) × 2 (target type of final test: name, diagram) × 2 (trial type: congruent, incongruent) mixed ANOVA on proportion correct on the final test. There was a main effect of encoding group, $F(1, 81) = 6.48$, $p = .013$, $\eta_p^2 = .074$, $BF_{10} = 3.12$. Retrieval practice (M = .31) led to higher performance than study practice (M = .23). There was also a main effect of target type, $F(1, 81) = 36.48$, $p < .001$, $\eta_p^2 = .311$, $BF_{10} > 10{,}000$. Performance was higher when participants retrieved names (M = .33) than diagrams (M = .21).

There was also a main effect of trial type, $F(1, 81) = 15.78$, $p < .001$, $\eta_p^2 = .163$, $BF_{10} = 52.42$, which was qualified by an interaction with encoding group, $F(1, 81) = 14.67$, $p < .001$, $\eta_p^2 = .153$, $BF_{10} = 88.22$. Simple effects analysis showed that on congruent trials, retrieval practice (M = .40) led to higher performance than study practice (M = .23), $F(1, 81) = 14.62$, $p < .001$, $\eta_p^2 = .153$, $BF_{10} = 97.13$. However, on incongruent trials, retrieval practice (M = .23) did not lead to higher performance than study practice (M = .23), $F(1, 81) = 0.10$, $p = .920$, $\eta_p^2 < .001$, $BF_{10} = 0.23$. In other words, there was no backward transfer effect, and the Bayesian analysis marshaled evidence for the null hypothesis, as the BF_{10} value indicated that the data were 4.35 times likelier to occur given the null compared to the alternative hypothesis (1/0.23).

Neither the interaction between encoding group and target type was significant, $F(1, 81) = 3.43$, $p = .068$, $\eta_p^2 = .041$, $BF_{10} = 1.34$, nor target type by trial type, $F(1, 81) = 3.15$, $p = .080$, $\eta_p^2 = .037$, $BF_{10} = 0.58$. The three-way interaction was not significant, $F(1, 81) = 0.06$, $p = .801$, $\eta_p^2 < .001$, $BF_{10} = 0.24$.

Discussion

We replicated the standard testing effect that has been observed with verbal (Carpenter et al., 2006) and visuospatial (Kang, 2010; Carpenter & Pashler, 2007) materials, but not the backward transfer effect that has been documented with purely verbal materials (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020). That is, retrieving a target (A - ?) only benefited the later retrieval of that target (A - ?) but not its cue (? - B). This occurred both when the target was a diagram (name - ?) or a name (diagram - ?). The Bayesian analysis suggested that the lack of a backward transfer effect was a true null effect and was not due to low statistical power.

These results suggest that the backward transfer effect may not generalize across materials. Rather, the effect may be obtainable only to the extent that the processes involved during the retrieval of cues and targets sufficiently overlap. In this experiment, the demands of retrieving a name and diagram were different, thereby reducing this overlap. Although each type of stimulus required the production of verbal information during tests (letters), only the diagrams required the placement of those letters in a complex visuospatial configuration. Indeed, research suggests the spatial processing of diagrams is different than that of words (see, Tversky, 2011; Winn, 1991; Winn & Sutherland, 1989). Moreover, the names could easily be read or spoken aloud, while this is not true of the associated diagrams. For example, it is likely that participants read "Oxirane" like any other word, but it is quite unlikely that they could have done the same for its associated diagram.

Our use of a cued-recall final test makes it difficult to interpret our null results. Note that successful performance on a cued-recall test reflects (a) item accessibility and (b) item memory. Although we found that recalling a target did not enhance performance on tests of its cue, it is difficult to know where the breakdown occurred. On the one hand, it is possible that an initial test of targets neither enhanced item accessibility, nor cue memory. On the other hand, it possible that retrieving a target

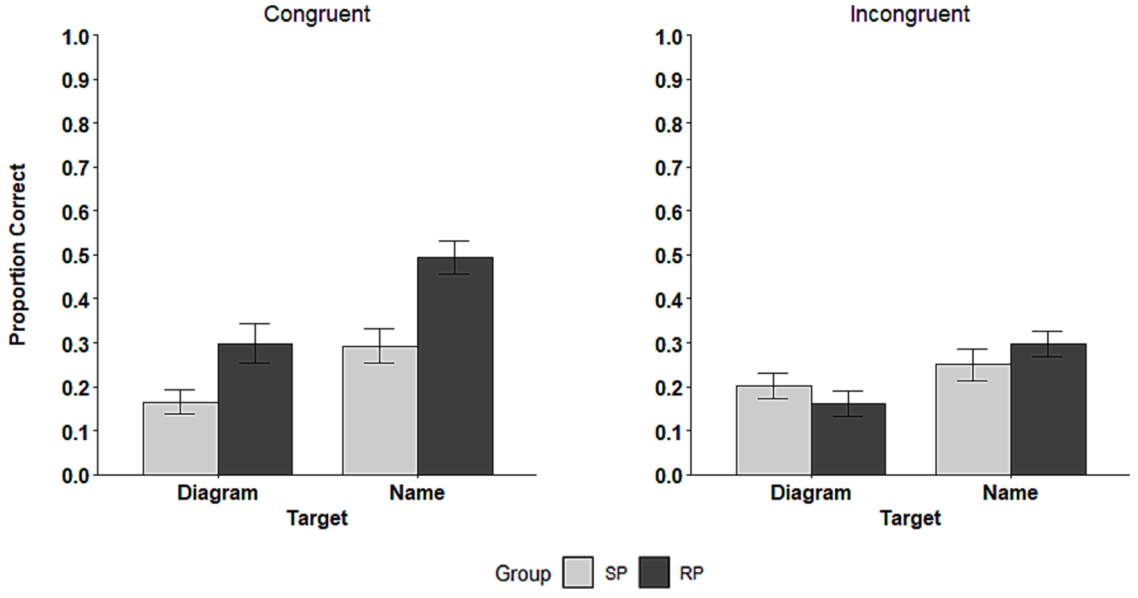

Fig. 3. *Experiment 1: Final Test Results Note.* SP = Study Practice. RP = Retrieval Practice. Error bars reflect standard error of the mean.

did enhance cue memory, but not the accessibility of that cue. To investigate that possibility, it would be necessary to use a final test that minimized the role of associative memory, but still probed item memory. This is what we explored in Experiment 2.

Experiment 2

In Experiment 1, retrieving targets did not benefit performance on tests of cues, both when the target was a name and when the target was a diagram. Given that memory for an item can be accurate, but not accessible on a recall test, the influence of retrieval practice on memory for the cue remains unclear. One way to explore this question is by using a recognition test to minimize the contribution of accessibility on performance (see, Koriat, 1993; Vaughn & Rawson, 2014).

In the present experiment, we used an identical encoding procedure as Experiment 1, but substituted the final cued-recall test with an item-by-item recognition test. During the recognition test, we presented participants with an intact name-diagram pair. When the names were tested, participants were told that its associated diagram was drawn correctly and were simply asked to assess if the name was correctly spelled (yes or no). When the diagrams were tested, participants were told that the name was correctly spelled, and were asked to assess if the diagram was constructed without error. Critically, we informed participants that this was not an associative recognition test (i.e., we would never pair the diagram of one pair with the name of another and vice versa). To make the lures for the names and diagrams, we made minor changes to the spelling, or structural composition, respectively. For example, the lure for "Norflurane" was "Norfluryne." An example of a diagram lure is shown in Fig. 4. This test therefore reduced the role of accessibility on successful performance, but still measured the accuracy of item memory by assessing participants' sensitivity to small deviations of correct items.

We also included confidence judgments to the final test. On each trial, participants provided confidence ratings on the correctness of their response on a 0 (complete guess) to 100 (completely certain) scale in increments of 25. We included these confidence ratings in order to assess the impact of correct guessing on the final recognition test, which adds noise to the measurement of memory. Even if a participant has no representation of an item in memory, they can still respond correctly by pure chance on a recognition test. The addition of confidence ratings also makes it possible to assess participants' metacognitive sensitivity to the impacts of study and retrieval practice on memory. Some studies suggest that learners are not always sensitive to the relative efficacy of study and retrieval practice (see, Kornell & Son, 2009; Roediger & Karpicke, 2006). For example, Roediger and Karpicke (2006) found that although participants encoded more material with retrieval compared to study practice, their confidence ratings in their learning outcomes were higher after engaging in study practice. However, other studies have found the opposite pattern (Hughes et al., 2018; Tullis et al., 2013).

Method

Design

We used a 2 (encoding group: study practice, retrieval practice) × 2 (target type on final test: name, diagram) × 2 (trial type: congruent, incongruent) mixed design. Learning technique was a between-subjects variable, and the other two variables were within-subjects.

Power analysis

To determine adequate sample size, we conducted an a priori power analysis. In Experiment 1, there was a large effect of retrieval practice on congruent trials ($\eta_p^2 = .153$), but essentially no effect on incongruent trials ($\eta_p^2 < .001$). Assuming the same average correlation among repeated measures as in the prior experiment ($\bar{r} = .35$), we would need 20 participants per encoding group to replicate the between-subjects effect of group on congruent trials. However, it is possible that retrieval practice would enhance performance on incongruent trials, but not to an equivalent degree. To detect an effect of retrieval practice on incongruent trials that would be half as large as incongruent trials ($\eta_p^2 = .077$), we would need 41 participants per encoding group. Due to changes in our design, we conservatively assumed a higher correlation among repeated measures (.70), thereby increasing the required sample size to 51 participants per encoding group.

Participants

Participants included 110 Tufts University undergraduate students aged 18 to 25 ($M_{age} = 20.65$, $SD = 1.37$), who were equally divided into the study-practice and retrieval-practice groups. Participants were either compensated with course credit or $10 per hour and could only participate if they had never taken a chemistry course beyond the standard high-school level.

Materials

The name-diagram pairs used during the encoding phase were the same as those in Experiment 1. Each name averaged 9.89 letters in length ($SD = 2.63$). Each diagram had an average of 6.44 atoms ($SD = 1.31$) and 6.19 bonds ($SD = 1.33$). Collapsed across atoms and bonds,

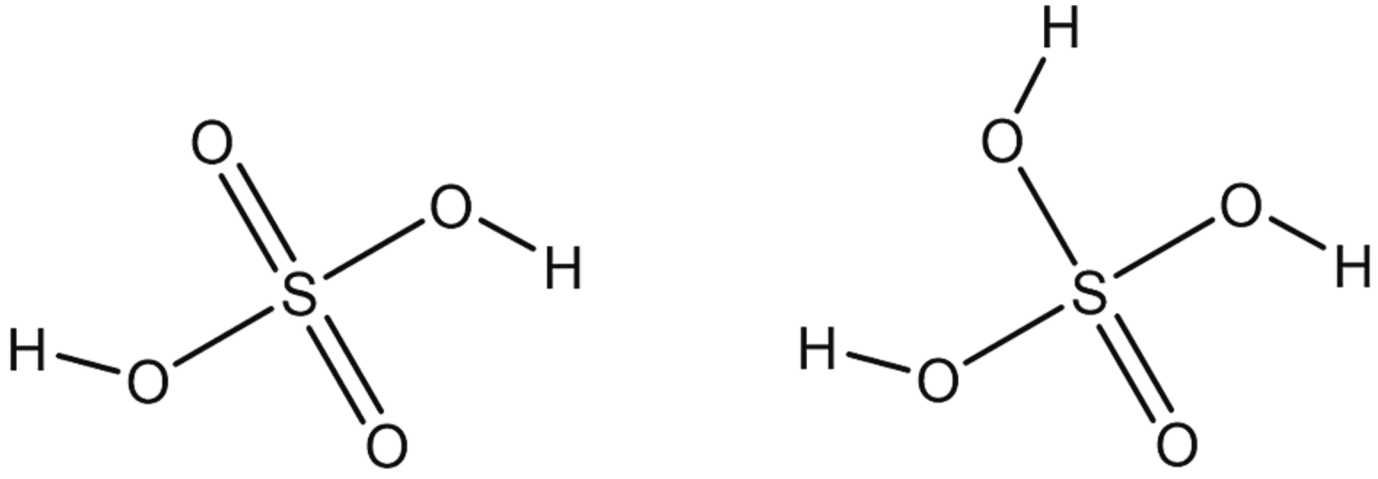

Fig. 4. *Experiment 2: Example of a Diagram Lure on the Final Recognition Test Note.* Left: correct diagram. Right: lure diagram.

the diagrams had an average of 12.69 components (SD = 2.60). The name lures averaged 9.65 letters in length (SD = 3.5). The diagrams had an average of 6.69 atoms (SD = 1.81), 6.63 bonds (SD = 1.50), and 13.13 (SD = 3.20) total components collapsed across atoms and bonds. The diagram lures were constructed so as not to violate the rules that govern the placement of atoms and the bonds between them. This was intended to mitigate the possibility that participants would reject any particular deviation from a studied diagram on the basis of (explicit reasoning about the correctness of a molecule. For example, a single hydrogen atom can never be connected to more than one atom and consequently, hydrogen items are always on the perimeter of a molecular structure. Participants saw many hydrogen atoms across the 16 molecules, and therefore if we used a lure that had a single hydrogen that connected to more than one atom, participants could reject that lure without having a specific memory representation of the corresponding correct structure of that lure.

Procedure

Initial Learning. This phase was the same as Experiment 1 in every way. Of note, participants were told to expect a cued-recall final test (as in Experiment 1). In this way, we attempted to minimize the possibility of test-expectancy effects, in which anticipating a recognition as opposed to recall test can reduce effort during encoding (e.g., d'Ydewalle, 1981; Thiede, 1996).

Final Test. The final test took place 48 hr after initial learning. As with Experiment 1, participants were tested on either the molecule name or the diagram for half of the items. When the name was tested, participants were presented with a name-diagram pair and informed that the diagram was completely correct, but that the name may or may not contain a small misspelling. We told participants that we would never pair a name, correct or misspelled, with the diagram of another pair. Thus, their job was simply to determine whether the name was spelled completely correctly. Participants recorded their answer on a test sheet and rated confidence on a five-point scale from 0 (complete guess) to 100 (completely certain). The procedure for testing the diagrams was identical, except that we informed participants that the name was spelled perfectly but the diagram might contain an error. Each item on the name test was presented twice: correct on one trial, and incorrect on a different trial. The same was true for the diagram test. Thus, there were 32 total questions on the final test.

Results

We examined performance on the final test with three measures: the proportion of hits (saying "correct" for a correct name or diagram), the proportion of false alarms (saying "correct" for an incorrect name or diagram), and discrimination (d'), which integrates hits and false alarms into a single metric. Because measures of d' cannot be calculated in cases with extreme values (e.g., perfect hits and false alarms), we applied the log-linear correction to the frequencies of the hits and false alarms before calculating d' (Snodgrass & Corwin, 1988). This approach involves adjusting the raw hits by adding .5 to the frequencies of hits and false alarms, and then dividing by the number of trials in the denominator plus one. Notably, the log-linear correction applies to all hits and false alarms, not just those that are extreme, which has been shown to bias the data (Verde et al., 2006).

Initial learning

Performance increased across tests of the names (M = .29, .46, and .57, respectively) and diagrams (M = .22, .37, and .47, respectively). A 2 (target type: name, diagram) × 3 (test number: 1, 2, 3) ANOVA showed a main effect of target type, $F(1, 54) = 12.94$, $p < .001$, $\eta_p^2 = .193$, $BF_{10} = 35.72$. Retrieving names (M = .44) was easier than diagrams (M = .35). There was also a main effect of test number, $F(1, 54) = 66.99$, $p < .001$, $\eta_p^2 = .554$, $BF_{10} > 10{,}000$, confirming increases in performance across trials, but no interaction between target type and test number, $F(2, 108) = 0.405$, $p = .668$, $\eta_p^2 = .007$, $BF_{10} = 0.09$. Although retrieving the names was easier than the diagrams, the rate of learning for each type of target across tests did not statistically differ.

Final test

Hits. We conducted a 2 (encoding group: study practice, retrieval practice) × 2 (target type: name, diagram) × 2 (trial type: congruent, incongruent) mixed ANOVA on proportion of hits during the final test. As shown in Table 1, there was no main effect of encoding group, $F(1, 108) = 0.54$, $p = .464$, $\eta_p^2 = .005$, $BF_{10} = 0.18$. There was a main effect of target type, $F(1, 108) = 11.78$, $p < .001$, $\eta_p^2 = .098$, $BF_{10} = 10.48$. Hits were higher for names (M = .82) than for diagrams (M = .76). There was also no main effect of trial type, $F(1, 108) = 0.003$, $p = .954$, $\eta_p^2 < .001$, $BF_{10} = 0.13$, but there was an interaction with encoding group, $F(1, 108) = 5.58$, $p = .020$, $\eta_p^2 = .049$, $BF_{10} = 1.77$. Retrieval practice led to a higher hit rate than study practice on congruent trials (M = .82 and M = .77, respectively), but lower hits on incongruent trials (M = .76 and M = .80, respectively). Neither the interaction between encoding group and target type, $F(1, 108) = 1.41$, $p = .238$, $\eta_p^2 = .013$, $BF_{10} = 0.27$, nor target by trial type, $F(1, 108) = 0.007$, $p = .697$, $\eta_p^2 = .001$, $BF_{10} = 0.18$, were significant. The three-way interaction was also not significant, $F(1, 108) = 0.898$, $p = .345$, $\eta_p^2 = .008$, $BF_{10} = 0.31$.

False Alarms. We conducted a 2 (encoding group: study practice, retrieval practice) × 2 (target type: name, diagram) × 2 (trial type: congruent, incongruent) mixed ANOVA on proportion of false alarms during the final test. There was a main effect of encoding group, $F(1, 108) = 9.40$, $p = .003$, $\eta_p^2 = .080$, $BF_{10} = 7.84$. Retrieval practice (M = .20) led to lower false alarms than study practice (M = .27). There was also a main effect of target type, $F(1, 108) = 73.51$, $p < .001$, $\eta_p^2 = .405$, $BF_{10} > 10{,}000$. There were lower false alarms for names (M = .15) than diagrams (M = .31). There was no main effect of trial type, $F(1, 108) = 1.87$, $p = .175$, $\eta_p^2 = .017$, $BF_{10} = 0.31$. None of the two-way interactions were significant: encoding group by target type, $F(1, 108) = 1.37$, $p = .244$, $\eta_p^2 = .013$, $BF_{10} = 0.29$; target type by trial type, $F(1, 108) = 0.29$, $p = .591$, $\eta_p^2 = .003$, $BF_{10} = 0.19$; or encoding group by trial type, $F(1, 108) = 2.18$, $p = .143$, $\eta_p^2 = .020$, $BF_{10} = 0.46$. The three-way interaction was also not significant, $F(1, 108) = 0.81$, $p = .371$, $\eta_p^2 = .007$, $BF_{10} = 0.24$.

Discrimination. The results are depicted in Fig. 5. We conducted a 2 (encoding group: study practice, retrieval practice) × 2 (target type: name, diagram) × 2 (trial type: congruent, incongruent) mixed ANOVA on discrimination (d'). There was a main effect of encoding group, $F(1, 108) = 4.58$, $p = .035$, $\eta_p^2 = .041$, $BF_{10} = 1.18$. Retrieval practice (M = 1.46) led to better discrimination than study practice (M = 1.25). There was also a main effect of target type, $F(1, 108) = 58.62$, $p < .001$, $\eta_p^2 = .350$, $BF_{10} > 10{,}000$. Discrimination was better for names (M = 1.63) than diagrams (M = 1.07). There was no main effect of trial type, $F(1, 108) = 1.28$, $p = .261$, $\eta_p^2 = .012$, $BF_{10} = 0.23$, but there was an interaction with encoding group, $F(1, 108) = 4.27$, $p = .015$, $\eta_p^2 = .053$, $BF_{10} = 2.06$. Retrieval practice led to better discrimination than study practice on congruent trials (M = 1.60 and 1.20, respectively), $F(1, 108) = 10.21$, $p = .002$, $\eta_p^2 = .086$, $BF_{10} = 17.37$, demonstrating the standard testing effect. However, it did not lead to a benefit on incongruent trials (M = 1.31 and 1.30, respectively), $F(1, 108) = 0.006$, $p = .937$, $\eta_p^2 < .001$, $BF_{10} = 0.20$, demonstrating no backward transfer effect. Notably, the BF_{10} value provided direct evidence for this null result, as the BF_{10} value indicated that the data were 5 times likelier to occur given the null

Table 1
Experiment 2: Mean hits and false alarms on the final recognition test, split by target type (name, diagram) and trial type (congruent, incongruent).

	Hits				False Alarms			
	Name		Diagram		Name		Diagram	
Group	Cong.	Incong.	Cong.	Incong.	Cong.	Incong.	Cong.	Incong.
SP	.77 (.21)	.83 (.23)	.75 (.24)	.77 (.21)	.17 (.16)	.18 (.20)	.36 (.26)	.35 (.26)
RP	.87 (.24)	.81 (.24)	.77 (.21)	.74 (.23)	.11 (.16)	.15 (.16)	.22 (.24)	.31 (.20)

Note. SP = study practice, RP = Retrieval Practice. Standard deviations given in parentheses. Cong. = congruent. Incong. = incongruent.

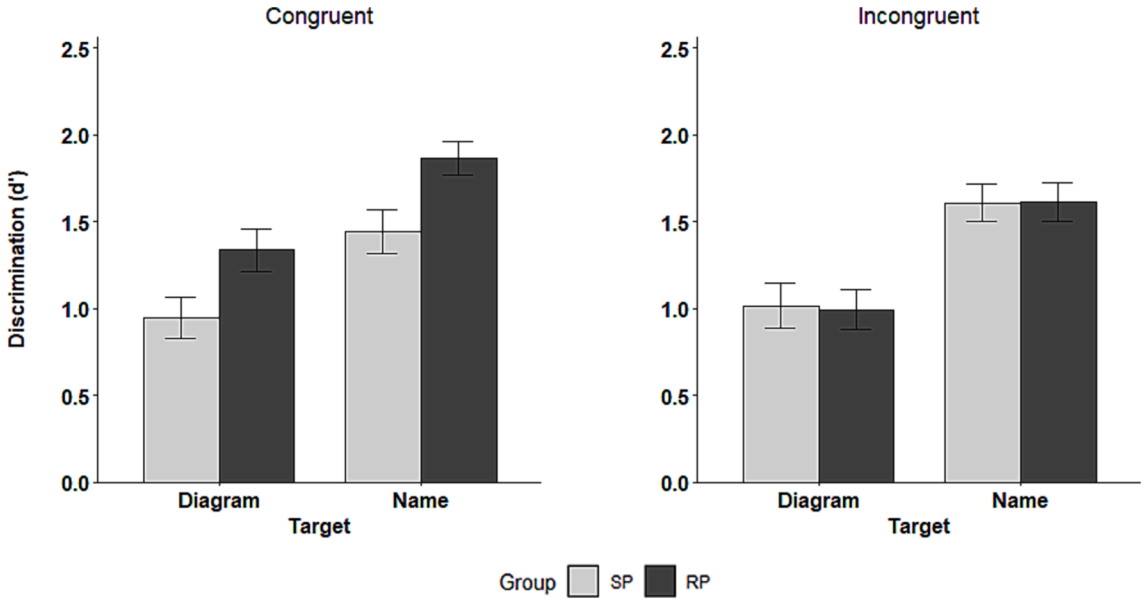

Fig. 5. *Experiment 2: Final Test Results Note.* SP = Study Practice. RP = Retrieval Practice. Error bars reflect standard error of the mean.

hypothesis compared to the alternative hypothesis (1/0.20).

Neither of the remaining two-way interactions were significant: encoding group by target type, $F(1, 81) < 0.01$, $p = .985$, $\eta_p^2 < .001$, $BF_{10} = 0.16$; target type by trial type, $F(1, 81) = 0.38$, $p = .536$, $\eta_p^2 = .004$, $BF_{10} = 0.17$. The three-way interaction was also not significant, $F(1\,0\,8) = 0.36$, $p = .849$, $\eta_p^2 < .001$, $BF_{10} = 0.21$.

Confidence. The motivation to include confidence was to examine the impact that correctly guessing an answer might have influenced the results. The rate of correct guessing was low ($M = .04$). When participants rated that their answer was a complete guess, they performed at chance levels ($M = .53$). For completeness, we present the results of a mixed ANOVA on average confidence, which showed a main effect of target type, wherein participants were more confident for names ($M = 74.82$) than diagrams ($M = 59.61$), $F(1, 108) = 99.73$, $p < .001$, $\eta_p^2 = .480$, $BF_{10} > 10{,}000$. There was no main effect of encoding group, $F(1, 108) = 0.83$, $p = .364$, $\eta_p^2 = .007$, $BF_{10} = 0.47$, or trial type, $F(1, 108) = 0.64$, $p = .425$, $\eta_p^2 = .006$, $BF_{10} = 0.16$. The two-way interactions were not significant, which included encoding group by target type, $F(1, 108) = 1.14$, $p = .289$, $\eta_p^2 = .010$, $BF_{10} = 1.68$, and target by trial type, $F(1, 108) = 2.84$, $p = .095$, $\eta_p^2 = .026$, $BF_{10} = 0.50$. The three-way interaction was not significant, $F(1, 108) = 3.64$, $p = .059$, $\eta_p^2 = .033$, $BF_{10} = 1.10$.

Discussion

The results of Experiment 2 do not support the hypothesis that retrieval practice of targets (A - ?) improves memory for cues in the backward direction (? - B). Rather, we replicated the same pattern of findings as was observed in Experiment 1. On congruent trials, retrieval practice ($M = 1.60$) led to statistically better memory accuracy than study practice ($M = 1.22$), but there was no benefit on incongruent trials. The Bayesian analysis suggested that the lack of an effect on incongruent trials was not an issue of lower power, but was a true null effect. These results suggest that the failure to observe a benefit of retrieval practice on backward transfer in Experiment 2 did not simply owe to a failure to promote accessibility of the solicited memories.

As discussed in the introduction, flexible memorization with paired associates can be difficult to achieve due to at least two barriers. One barrier is when the cues and targets are sufficiently different in nature or encourage different types of processing. This may have been the case here. Although the names and diagrams contained verbal information (letters), only the names could be easily read or spoken aloud, and only the diagrams required the memorization of visuospatial information. This reduces the likelihood that the processes that occur during encoding and the final test will match. Another barrier is when participants lack a sufficient network of pre-existing semantic and association knowledge, which makes associative memory easier. By choosing chemistry novices, we minimized the impact of such knowledge on performance. Although names and targets may inherently be processed in different ways, it is conceivable that scaffolding new networks of knowledge would nevertheless make flexible association easier. We investigate this possibility in Experiment 3.

Regarding the confidence ratings, our results suggest that participants were not sensitive to the relative benefit of retrieval practice on long-term memory. Discrimination performance was higher in the retrieval practice group. However, average confidence ratings during the final test were only numerically, but not statistically, higher in the retrieval ($M = 69.12$) compared to the study practice group ($M = 65.80$). This study echoes prior work finding that people's metacognitive judgments do not always reflect the memorial advantage of retrieval over

study practice (Kornell & Son, 2009; Roediger & Karpicke, 2006). One factor that may have contributed to this null finding was the use of a between-subjects design. Research suggests that metacognitive judgments are more sensitive to differences in encoding or processing conditions in within-subject designs because they allow people to compare their experiences across conditions directly (see, Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Carroll & Nelson, 1993; Shaw & Craik, 1989; Tullis, Finley, & Benjamin, 2013). For example, Tullis et al (2013) found that retrieval practice led to higher final-test performance and higher average metacognitive ratings than study practice in a within-subjects design. However, Hughes et al. (2018) did observe higher average metacognitive judgments for participants who used retrieval practice in a between-subjects design. It is possible that people's metacognitive judgments are generally sensitive to larger, but not smaller, differences in the relative efficacy of encoding practices, especially in a between-subjects design.

Experiment 3

If the absence of an extensive network of pre-existing knowledge makes flexible association more difficult, then methods to create new networks of knowledge should make it easier. A category induction paradigm is one method of accomplishing this aim, as it results in new types of knowledge that link members of individual items together (cue ← rules → target) and contextualizes them with knowledge about other items. We had participants learn categories of organic chemistry molecules. As shown in Fig. 6, all organic molecules are composed of chains of carbon and hydrogen atoms (C's and H's) that vary in length (number of C's). The names of organic molecules are derived from a rule-based system, in which subcomponents of the name specify the precise type, quantity, and spatial arrangement of its atoms. Subcomponents of the name (affixes) communicate different types of information: (a) the presence of a category-defining cluster of items or (b) the length of the carbon-hydrogen chain. Knowledge of these rules allows people to translate names into diagrams, and vice versa, and can thereby enhance flexible associative memory.

There were two goals of Experiment 3. The first goal was to determine if the backward transfer effect would occur in a category induction task. The second goal was to determine if retrieval practice could foster memorization of exemplars and inductive learning simultaneously. In the previous studies on category induction, retrieval practice involved testing the category name of a previously studied item (exemplar - ?) during initial and final tests. These studies did not ask participants to memorize individual exemplars verbatim, and the final tests likewise did not probe this type of knowledge.

It is possible that the enhancement of retrieval practice on inductive learning comes at the expense of verbatim item memorization. Indeed, research has consistently found that asking people to label exemplars with a category name preferentially enhances the encoding of rule-relevant features (see, Chin-Parker & Ross, 2004; Jones & Ross, 2011; Medin et al., 1987; Nosofsky et al., 1994; Yamauchi & Markman, 1998). This is because these tasks explicitly incentivize participants to isolate the subset of features that are consistently associated with one category label, and not another. Forcing participants to retrieve entire exemplars requires processing the rule-irrelevant features, perhaps obstructing inductive processes. Alternatively, memorizing individual exemplars could profit inductive learning, since research suggests that category rules can be inferred by comparing presented items to those stored in memory (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986).

Note that the typical category induction paradigm is not well-suited to investigating the verbatim memorization of individual items. This is because a single category name is associated with multiple, distinct exemplars. A category name is therefore not an effective cue for the recall of a single, specific exemplar target. For cued-recall tests to elicit memory of a specific exemplar, the name of that exemplar must include information that points to it specifically, and not any other exemplar of that category. Therefore, in this experiment, the label of each exemplar not only included its category membership, but also a component that uniquely specified an individual exemplar within that category. The name of each molecule included a category label (e.g., "Hydroxyl") and a number (1–4) that uniquely specified one of a category's exemplars. When tested on the names, participants were asked to produce both the category label and the number. The stimuli are shown in Fig. 7.

Overview of Experiment 3

The initial learning phase was the same as the previous experiments. Participants studied name-diagram pairs through study practice or

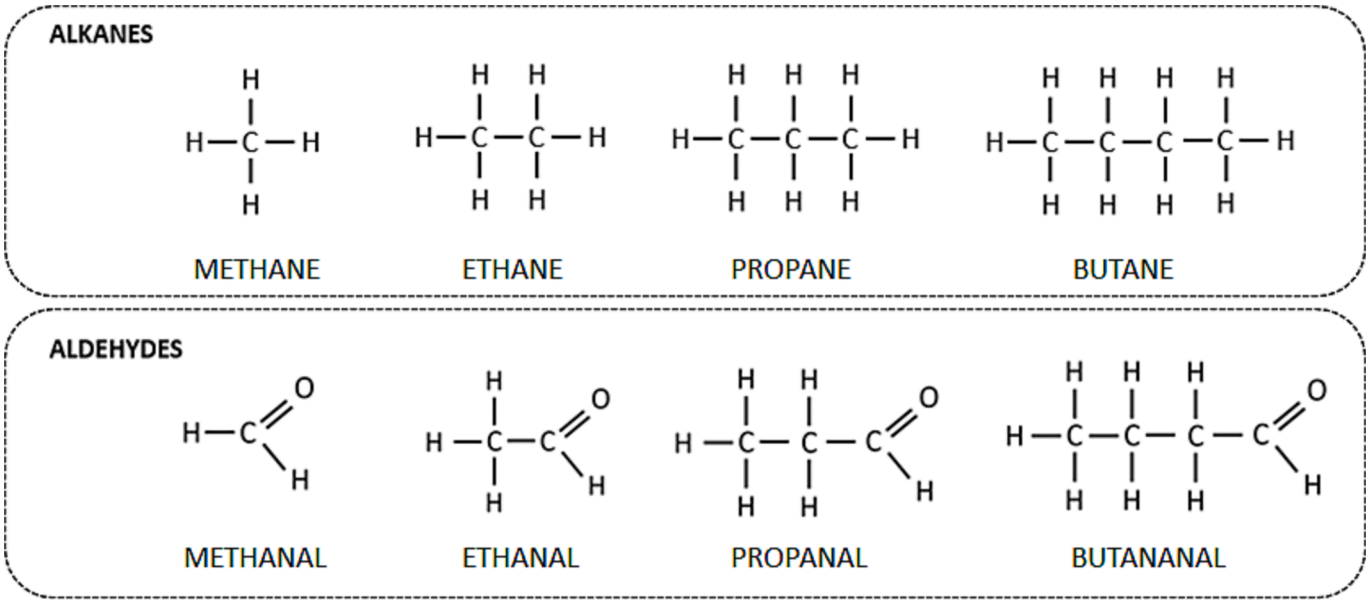

Fig. 6. *Examples of Organic Chemistry Molecule Categories Note.* Examples of organic chemistry molecules: alkanes and aldehydes. Each name consists of a prefix (meth-, eth-, but- etc.) that designates the length of the carbon/hydrogen chain (number of carbon atom [C's]), and has a suffix that designates the present of a category defining feature (-ane means only C's and H's), and -anal means C's and H's plus a single, double-bonded O attached to one C.

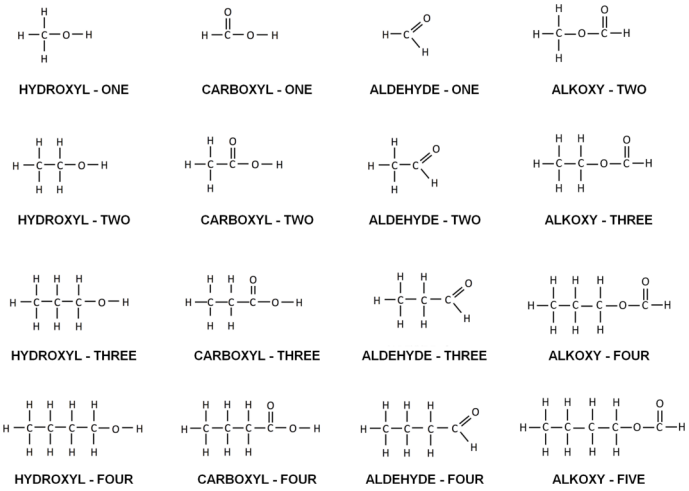

Fig. 7. *Experiment 3: Category Stimuli Note.* The 16 molecules participants studied during initial learning. The stimuli were divided evenly into four categories. The identity of the category was defined by a unique cluster of atoms on the right-hand side of the diagram.

retrieval practice of the names or diagrams. The pairs were equally divided into four categories. The names consisted of two parts, a category label and a number. The category label was uniquely associated with the portion of the diagram that determined the category membership of the molecule. In contrast, the number referred to the length of the carbon-hydrogen chain to which the diagnostic feature is attached. For example, "hydroxyl four" and "aldehyde four" both share a carbon-hydrogen chain consisting of four carbon atoms, but each differed in the cluster of atoms that designated their respective categories. As before, we fully crossed initial tests of the diagrams and names, which permitted an investigation of the backward transfer effect.

In addition to testing participants on the 16 previously studied (old) exemplars, we also tested them on 16 new exemplars of those categories. Whereas test performance on the old items can be accomplished through verbatim memorization, performance on the novel items reflects inductive learning. For the novel items, we increased the length of the carbon-hydrogen chain of the molecules studied during initial learning. For example, the four exemplars of "aldehyde" during initial learning ranged from having one to four carbon atoms in its carbon-hydrogen chain. On the final test, we included aldehydes with five to eight carbon atoms (novel items). Producing the name or diagram of these novel items could not be done with memorization, as they had not been seen previously.

In contrast to the previous experiments, we used two between-subjects retrieval practice groups. That is, one retrieval-practice group learned all name-diagram pairs through cued-recall tests of the diagram (name - ?; $RP_{DIAGRAM}$) and the other through cued-recall tests of the name (? - diagram; RP_{NAME}). We made this change because the categorical nature of the stimulus set means that within-subjects encoding manipulations with one set of items would be likely to influence the learning of other items. In other words, retrieving the diagrams of one category of molecule could influence learning the names of another. Using two different retrieval practice groups allowed a better comparison of how the two types of retrieval (name or diagram) influence learning compared to the within-subjects approach used in the previous experiments.

Method

Design

We used a 3 (encoding group: study practice, RP_{NAME}, $RP_{DIAGRAM}$) × 2 (target type on final test: name, diagram) × 2 (item type on final test: old, new) mixed design. Encoding group was manipulated between subjects. Final-test type and item-type were manipulated within encoding groups.

Power analysis

To ensure an adequate sample size, we conducted an a priori analysis. In Experiments 1 and 2, we observed a large benefit of retrieval practice when the type of target during initial learning matched that of the final test ($\eta_p^2 = .153$ and $\eta_p^2 = .086$, respectively), but essentially no effect when they did not match (both $\eta_p^{2'}s < .001$). Due to the changes of the present experiment, we did expect a testing effect when the type of target retrieved during initial learning did not match the final test. To detect a medium effect size of encoding group ($\eta_p^2 = 0.06$), the analysis showed that we would need 120 participants to achieve 80% power ($\alpha = .05, \overline{\rho} = .70$). The power level of all the statistical tests was assessed with the effect-size sensitivity analyses reported in the results section.

Participants

We randomly and evenly assigned 120 participants from Tufts University (*M*age = 21.23, *SD* = 1.33) to the three experimental encoding groups: study practice, RP_{NAME}, and $RP_{DIAGRAM}$. Participants were compensated with course credit and invited to participate only if they had never taken chemistry beyond the standard high-school level.

Materials

***Old Items*.** Participants learned 16 molecular name-diagram pairs. The diagrams consisted of three types of atoms: carbon (C), hydrogen (H), and oxygen (O). The diagrams contained two pieces. First, all diagrams consisted of a chain of carbon atoms that were surrounded by hydrogen atoms (1 to 4 carbon atoms). Second, the diagrams contained a unique configuration of atoms that separated it into one of four categories. All diagrams therefore contained information that specified its category membership and information that did not.

The names of the molecules likewise contained two pieces of information: a category label and a number. The label was uniquely associated with the category-defining configuration of atoms in the diagram. The number reflected the number of carbon atoms in the chain. The names and diagrams both contained information that was diagnostic or non-diagnostic of category membership. The labels consisted of a word that ranging from $6 - 8$ letters ($M = 7.50$) letters in length and a number (one to five). Because two of the category labels are commonly known amongst non-experts (e.g., alcohol, ester), we substituted them with the

names of their defining atom clusters, called functional groups (e.g., hydroxyl, alkoxy).

The diagrams were line drawings of corresponding molecular structures that had an average of 10.25 atoms (SD = 3.79) and 10.00 bonds (SD = 3.79). Collapsed across atoms and bonds, each diagram contained an average of 20.25 total components (SD = 7.57). Each molecule was connected to another via single bonds (one line) or double bonds (two lines).

New Items. For each category, the final test included four name-diagram pairs that were never encountered during the initial learning phase (16 new pairs). These stimuli were created by increasing the length of the carbon chain from the old items (2–5) to the new items (5–8). For example, participants learned "hydroxyl - one" through "hydroxyl four" during the initial learning phase, but were later tested on "hydroxyl - five." Each name averaged 8 letters in length (SD = 1.26). Each diagram contained an average of 20.25 atoms (SD = 4.55), 20.00 bonds (SD = 4.62), and 40.25 total components collapsed across atoms and bonds (SD = 9.16).

Procedure

Initial Learning. The procedure and instructions were identical to that of Experiments 1 and 2, except the retrieval-practice group was split into two between-subjects groups: retrieval practice of the names (RP_{NAME}) and of the diagrams ($RP_{DIAGRAM}$). As in the previous experiments, participants were told to expect tests of the names and/or diagrams, but were never informed that the stimuli were categorical in nature. The items were presented in a random order, such that the category exemplars were intermixed.

Final Test. The final test took place 48 hr after initial learning. Participants took a cued recall test on the 16 items studied during initial learning phase, which consisted of testing the names or diagrams of a given pair in an intermixed fashion. For half of these items (two categories), participants were tested on the name, and for the other half, the diagram (the other two categories). We used four counterbalances to assign items/categories to the name and diagram conditions. For example, in one counterbalance, the names of hydroxyl and aldehyde molecules were tested, and the diagrams of carboxyl and alkoxy molecules were tested. In another counterbalance, the names of hydroxyl and carboxyl molecules were tested, and the diagrams of aldehyde and alkoxy molecules were tested. As in the previous experiments, participants had 30 s to provide an answer, after which the screen advanced to the next item. The test also included the 16 new items, half of which were tested on the names, and the other half which were tested on the diagrams. The new and old items were intermixed. The assignment of the 16 new items to the two final test conditions matched the old items. That is, if participants were tested on the name of old hydroxyls, they were likewise tested on new hydroxyls. At the start of the testing phase, participants were informed that the test may include items that had not been seen during the study phase, but to attempt to provide an answer. We confirmed that participants did not have experience with these organic molecules by asking them at the end of the testing phase.

Results

Initial learning

Performance in the RP_{NAME} group increased across the three tests (M = .54, .76, and .88). Performance in the $RP_{DIAGRAM}$ group was slightly lower (M = .43, .59, and .77). A 2 (encoding group: RP_{NAME}, $RP_{DIAGRAM}$) × 3 (test number: 1, 2, 3) ANOVA showed a main effect of encoding group, $F(1, 78) = 6.15$, $p = .015$, $\eta_p^2 = .073$, $BF_{10} = 3.46$. On average, participants retrieved names (M = .73) at a higher rate than diagrams (M = .60). The main effect of test number was also significant, $F(2, 156) = 104.24$, $p < .001$, $\eta_p^2 = .572$, $BF_{10} > 10{,}000$, and there was no interaction with encoding group, $F(2, 156) = 0.927$, $p = .398$, $\eta_p^2 = .012$, $BF_{10} = 0.17$. Therefore, although performance was higher in the RP_{NAME} group, increases in learning across the tests occurred at a similar rate for the $RP_{DIAGRAM}$ group.

Final test

As shown in Fig. 8, the two retrieval practice groups led to higher performance than study practice. The results also suggest a backward transfer effect. Compared to study practice, the retrieval of names led to higher performance on studies of the diagram, and retrieval of the diagrams led to better performance on retrieval of the names. This pattern occurred both for old and new items. Overall, performance on the old items and new items were roughly equivalent, demonstrating that participants had learned rules.

We conducted a 3 (encoding group: study practice, RP_{NAME}, $RP_{DIAGRAM}$) × 2 (type of final test: name, diagram) × 2 (item type: old, new) mixed ANOVA. As shown in Fig. 6, there was a main effect of encoding group, $F(2, 117) = 14.22$, $p < .001$, $\eta_p^2 = .196$, $BF_{10} = 3{,}258.21$. showing that RP_{NAME} (M = .56) and $RP_{DIAGRAM}$ (M = .57) led to higher performance than study practice (M = .25). There was also a main effect of target type $F(1, 117) = 31.24$, $p < .001$, $\eta_p^2 = .211$, $BF_{10} > 10{,}000$, again replicating the finding that names (M = .55) were easier to retrieve than diagrams (M = .37). The main effect of item type was not significant, $F(1, 117) = 1.23$, $p = .269$, $\eta_p^2 = .010$, $BF_{10} = 0.21$. Performance was equivalent on old items (M = .46) and new items (M = .45). There was no interaction between encoding group and item type, $F(2, 117) = 0.884$, $p = .416$, $\eta_p^2 = .015$, $BF_{10} = 0.16$, and the three-way interaction was not significant, $F(2, 117) = 2.34$, $p = .101$, $\eta_p^2 = .038$, $BF_{10} = 0.52$.

There was an interaction between encoding group and target type, $F(2, 117) = 11.52$, $p < .001$, $\eta_p^2 = .165$, $BF_{10} = 533.43$. Simple effects analysis showed that the effect of encoding group was significant both for names, $F(1, 108) = 14.50$, $p < .001$, $\eta_p^2 = .199$, $BF_{10} = 7{,}281.21$, and diagrams, $F(1, 108) = 12.67$, $p < .001$, $\eta_p^2 = .178$, $BF_{10} = 1{,}867.27$. Bonferroni post-hoc tests showed that on the final test of the names, RP_{NAME} (M = .74) did not result in statistically superior performance than $RP_{DIAGRAM}$ (M = .56), $t(78) = 2.34$, $p = .062$, $d = 0.53$, $BF_{10} = 2.35$, but did compared to study practice (M = .34), $t(78) = 5.37$, $p < .001$, $d = 1.22$, $BF_{10} > 10{,}000$. Most importantly, $RP_{DIAGRAM}$ led to higher performance than study practice on the name test, $t(78) = 3.03$, $p = .009$, $d = 0.69$, $BF_{10} = 7.28$, demonstrating a backward transfer effect.

Similarly, on the final test of the diagrams, $RP_{DIAGRAM}$ (M = .57) led to higher performance than RP_{NAME}, (M = .38), $t(78) = 2.44$, $p = .048$, $d = 0.55$, $BF_{10} = 2.02$, and study practice (M = .17), $t(78) = 5.03$, $p < .001$, $d = 1.14$, $BF_{10} = 9{,}966.14$. RP_{NAME} also resulted in superior performance than study practice on the diagram test, $t(78) = 2.58$, $p = .033$, $d = 0.59$, $BF_{10} = 5.66$, again demonstrating a backward transfer effect.

There was also an interaction between target type and item type on the final test, $F(2, 117) = 7.06$, $p = .009$, $\eta_p^2 = .057$, $BF_{10} = 4.29$. For diagrams, performance was higher on the old items (M = .39) than new items (M = .36), $t(117) = 2.83$, $p = .005$, $d = 0.52$, $BF_{10} = 5.10$. For names, performance on the old items (M = .54) and new items (M = .55) did not significantly differ, $t(117) = 0.95$, $p = .345$, $d = 0.17$, $BF_{10} = 0.16$.

Discussion

There were two main findings of Experiment 3. First, we found that retrieval practice enhanced inductive learning and verbatim item memorization simultaneously. Second, we observed a backward transfer effect. Retrieval practice of name and diagram targets not only led to a testing effect on final tests of those targets, but also their cues, which were never the subject of recall attempts during encoding. This finding

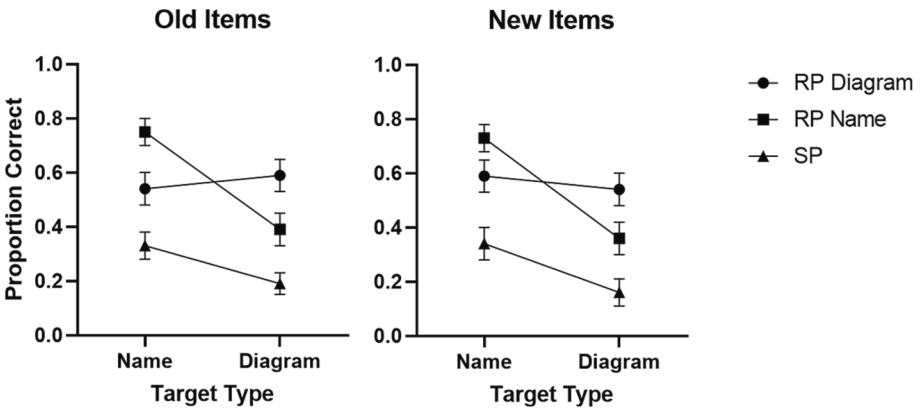

Fig. 8. *Experiment 3: Final Test ResultsNote.* Error bars represent *SEM*. SP = Study Practice. RP Diagram = Retrieval Practice group that retrieved diagrams during encoding. RP Name = Retrieval Practice group that retrieved names during encoding.

replicates previous research conducted with word pairs (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020). The results of the present experiment contrast starkly with the preceding experiments, which demonstrated no enhancement of target retrieval on cue learning, both when the final tests involved recall and recognition.

Our results suggest that verbatim memorization of items does not necessarily conflict with inductive learning. However, one reason that we did not observe this conflict could be that most of the features of the exemplars were task relevant. That is, the rules that could be learned about the exemplars reflected all of their features, in one way or another, thereby aligning memorial and inductive processes. Although the category-defining cluster of atoms was only a small portion of each exemplar, the exemplar name (the number) pertained to the rest of the features. This leaves open the possibility that our results would not replicate when the rule-relevant features take up a smaller proportion of the item's features.

Although we obtained a backward transfer effect, there were some notable differences in how retrieval of names affected the learning of diagrams, and vice versa. The testing effect was twice as large when the type of information tested during encoding and the final test matched (+.40) compared to when they did not match (+.20). That is, retrieving names preferentially enhanced encoding of those names compared to their corresponding diagrams, and vice versa.

Although comparisons to the study practice group suggest that retrieval of the name and diagram targets led to learning that was equally transferable to tests of their cues (+.20 and +.19, respectively), comparisons within the two retrieval practice groups tell a different story. Practicing retrieval of the diagrams led to equivalent performance on final tests of the diagrams ($M = .57$) and names ($M = .56$). In contrast, practicing the retrieval of names led to much higher performance on final tests of the names ($M = .74$) than diagrams ($M = .38$). Another way to examine this apparent asymmetry is to compare how the performance during the retrieval-practice tests compared to the final test. On the last of the three retrieval-practice tests during initial learning, performance in recalling diagrams was .77. On the final tests of the diagrams and names, this level of performance dropped by .20 and .19, respectively. This means that the learning that occurred during the retrieval of the diagrams during encoding transferred equally to final tests of the diagrams and cues. However, the same pattern was not found for the retrieval of names during encoding. Performance in retrieving the names on the last retrieval-practice test was .88, which dropped by .14 on final tests of the name and .50 on final tests of the diagram. In other words, retrieving the names affected the learning of those names to a much greater degree than their corresponding diagrams.

It is possible that comparative difficulty of retrieving the diagrams, compared to name, accounts for this pattern of findings. On average, more difficult initial tests lead to larger testing effects (see, Carpenter et al., 2006). Presumably, this is because the longer that a target eludes retrieval, the more that participants will process the cue and/or search their memory of prior learning episodes for evidence that can facilitate access to the sought-after information (Carpenter, 2009, 2011; Lehman, Smith, & Karpicke, 2014; Pyc & Rawson, 2009; Karpicke, Lehman, & Aue, 2014). If the target is retrieved, then the processing of the cue and the debris from the retrieval attempt can become bound together to form an exceptionally rich and detailed memory trace. In contrast, retrieving targets quickly could reduce or obviate these processes, leading to a smaller testing effect. Therefore, it is conceivable that the difficulty of retrieving the diagram targets led to a proportionally higher time spent reviewing name cues. If so, the act of producing the diagram may not have transferred to learning of the cue, but rather the increased duration of time that was spent attempting to produce it and/or processing the cue. That said, during initial learning, the diagrams were only modestly more difficult to retrieve than the names (an average difference of .10 across the retrieval practice tests).

Of note, since we used identical instructions for the learning phase as Experiments 1 and 2, we did not inform participants of the categorical nature of the stimuli at any point during the experiment. The explicit task goal was verbatim memorization. Category learning in this experiment must have occurred spontaneously, which is a common phenomenon in real-world and laboratory settings (see, Love et al., 2004). Category learning could occur by making inter-item comparisons within and between categories (e.g., noticing that items that look alike share a name). The explicit goal of memorization could itself have fostered spontaneous category learning. Research suggests that in more typical memorization studies, people spontaneously use a wide variety of strategies to make the encoding process easier, like sentence generation and imagery (e.g., Wollen & Lowry, 1971). It is conceivable that without any instruction to do so, participants in this experiment searched for simple heuristics, rules, or regularities that would make memory recall easier. This in turn could have led to the discovery of the underlying category rules.

General discussion

In the present study, we explored the influence of retrieval practice on the learning and transfer of paired-associate stimuli, which consisted of the names (verbal) and diagrams (visuospatial) of molecules. In all three experiments, participants learned name-diagram pairs either through retrieval practice with feedback or study practice. In two experiments, retrieval practice of the targets enhanced memory for those targets on final cued-recall (Exp 1) and recognition (Exp 2) tests, but did not influence memory for the cues. In these experiments, we replicated the benefit of retrieval practice on learning of verbal (Bulevich et al., 2016; Carpenter & DeLosh, 2006; Hughes et al., 2018; Tullis et al., 2013;

Thomas et al., 2020) and visuospatial targets (Carpenter & Pashler, 2007; Kang, 2010; Rohrer et al., 2010), but did not find a benefit on backward transfer, as has been demonstrated with word pairs (Carpenter et al., 2006; Cheng, 2014; Rickard & Pan, 2020).

In a final experiment, we explored the hypothesis that retrieval practice would benefit backward transfer in a category learning task, which encouraged the induction of rules that governed the relationships between names and diagrams. Under these conditions, retrieval practice of targets led to better performance than study practice on final tests of the targets and cues. Further, this not only occurred when the final test probed memory for the studied items, but also when it required the production of completely new names and diagrams, which could only be accomplished by having inferred rules from the studied items. Therefore, we also replicated the finding that retrieval practice benefits visual category learning (Jacoby et al., 2010; Cho & Powers, 2019).

A common theme emerged across the experiments: although retrieval practice of targets reliably enhanced the long-term retention of those targets, extension of that enhancement to novel contexts was less reliable. In the present experiments, retrieval of name and diagram targets consistently enhanced performance on final tests of those targets relative to study practice. However, only in Experiment 3 did this enhancement extend to final tests of the cues, and even then, the transfer was only partial. That is, the testing effect was twice as large on tests of the targets (+.40) than the cues (+.20). Further, obtaining this effect involved a considerable departure from the standard paired-associate paradigm used in the studies that documented it with word pairs (Experiments 1 and 2). Rather, in the present study, a benefit of retrieval practice on backward transfer only occurred in a category learning task, which permitted the inference of rules through inter-item comparisons. This study therefore coheres with the growing consensus that, although recalling an item on an initial test reliably enhances its later retrieval, the transfer of that enhancement to other contexts is not as dependable.

It is worth noting that in Experiment 3, the learning obtained through initial cued-recall tests of targets transferred to a recognition test. This replicates the common finding that the testing effect does not evaporate entirely when the formats of the initial and final tests do not match (McDermott et al., 2014; Putnam & Roediger, 2013; Thomas et al., 2020). This accords with the general finding that formats of retrieval practice that are more demanding, like those that involve recall, transfer better to final tests that are less demanding, like recognition, than the other way around (see, Carpenter & Delosh, 2006).

Backward transfer

Our results suggest that the backward transfer effect depends on properties of the materials. Whereas prior studies readily obtained the effect with simple word pairs, we did not replicate it with name-diagram pairs in a standard paired associate paradigm. We highlight two potential reasons that the backward transfer effect was not as easy to obtain with our materials. First, the processes involved in the retrieval of names and diagrams were likely too different, as only the retrieval of diagrams required the production of complex visuospatial information. Second, an absence of sufficient pre-existing knowledge could have made it too difficult to create flexible associative memories. The fact that we observed the backward transfer effect only in the context of a category induction task, which fosters new networks of knowledge, lends credence to this idea. However, this evidence is hardly decisive, as we did not directly demonstrate that the induction of rules mediated the forward and backward recall of names and diagrams. As it stands, we have merely demonstrated that the induction of rules and the backward transfer effect can co-occur, but more evidence is needed to establish a causal relationship. Future research should address these topics.

One alternative explanation for our results concerns differences in item difficulty across the experiments. The materials were easier to master in Experiment 3 than the prior two experiments. This is evidenced by higher performance during initial encoding in Experiment 3 (66%) compared to Experiment 1 (40%) and Experiment 2 (42%). Moreover, cued-recall final test performance was higher in Experiment 3 (46%) than Experiment 1 (27%). Perhaps it is easier to obtain a backward transfer effect with easier materials. However, arguing against this item difficulty account is the complete lack of a backward transfer effect in Experiments 1 and 2. If item difficulty accounted for our pattern of results, then one would expect at least a hint of a backward transfer effect in the first two experiments. The Bayesian analyses marshaled evidence that these were true null effects.

Retrieval practice theories

The prevailing theories of the testing effect are expressly concerned with the forward, not backward, effects of initial cued recall on memory. That is, they focus on why retrieving targets profits the later recall of those targets, and not the cues. Of these theories, the elaborative accounts of the testing effect specify a clear mechanism by which the backward transfer effect can occur. According to these accounts, the act of testing enhances memory by encouraging the automatic and/or deliberate creation of associative pathways that can mediate recall (Carpenter, 2009, 2011; Carpenter & Yeung, 2017; Chan et al., 2006; Pyc & Rawson, 2010; Vaughn & Rawson, 2011). If during an initial test, the presentation of the cue fails to elicit the target immediately, associated information may come to mind automatically (Carpenter, 2011), or the learner may actively search their memory for any information that can be used to access or reconstruct the target (Chan et al., 2006; Pyc & Rawson, 2009). If the learner subsequently retrieves the target or reviews it during feedback, then this associated information (*mediators*) can serve as memory bridges between the cue and the target. On a later test, if the cue does not prompt access to the target directly, it may nevertheless activate the mediators, which in turn could facilitate access to the target (cue → mediator → target). It is easy to see how this process could occur in reverse (cue ← mediator ← target).

For this elaborative mechanism to result in backward transfer, presentation of the target must prompt activation of the mediating information. However, research suggests that elaborative activities can be much stronger in one direction than another (Carpenter & Yeung, 2017; Vaughn & Rawson, 2014). For example, as Vaughn and Rawson (2014) note, a common elaborative strategy with Swahili-English word pairs is to explicitly generate mediators that are phonetically related to the cue but semantically related to the target. Consequently, the authors argue that such mediators would be more useful in one direction. The authors use the pair “wingu - cloud” to illustrate the point. A participant may generate the mediator “wing” during elaboration because it is phonetically similar to “wingu” and has a conceptual connection to the target (birds have wings, and birds can be found in clouds). However, the authors suggest that the presentation of the word “cloud” is less likely to elicit activation of “wing” than the much more phonetically-similar “wingu.” In Experiments 1 and 2, if retrieval practice did foster elaborative activities, it is conceivable that these activities fostered recall only in the forward direction. Anecdotally, one participant reported using semantic elaboration during the initial tests, such as generating keywords that described the global shape of the molecule (e.g., to them, the structure of propyne resembled a sword). It is possible that the diagram, which resembles a sword, would be more likely to activate that keyword than the name propyne, which does not. In Experiment 3, it is clear that elaboration did occur, as participants spontaneously extracted the category rules. It would make sense that these elaborative activities would foster recall bidirectionally, as the category rules thoroughly interrelated components of the cues and targets, making it likely that both items would prompt the associated rules. Again, however, we must emphasize that our study did not directly assess a mediation mechanism.

Category induction

Our results suggest that verbatim memorization of items does not

necessarily conflict with inductive learning. Both exemplar (Hintzman, 1986; Kruschke, 1992; Medin & Schaffer, 1978; Nosofky, 1986, 2011; Stewart & Brown, 2005) and prototype models (Minda & Smith, 2001; Reed, 1972; Vanpaemel & Storms, 2008) of category learning posit that during category learning, people selectively focus on the stimulus dimensions that are most relevant to the task. From this perspective, it could be argued that any factor that forces participants to reallocate their attention from the relevant to the irrelevant features would obstruct category learning. Forcing participants to encode items verbatim is one such factor, as the task explicitly requires attending to the task-irrelevant features. One reason that we may not have observed such a conflict is due to the way we constructed our materials. In Experiment 3, most of the features of the exemplars were task relevant. That is, the rules that could be learned about the exemplars reflected all of its features, in one way or another, thereby aligning memorial and inductive processes. Although the category-defining cluster of atoms was only a small portion of each exemplar, the exemplar name (the number) pertained to the rest of the features. The number of carbons was, itself, a category and therefore task-relevant. This leaves open the possibility that our results would not replicate when the rule-relevant features take up a smaller proportion of the item's features.

Recently, the topic of exemplar memorization has been the focus of studies that investigate the sequence that items are presented. Many of these studies examine the influence of *interleaving* the study of category exemplars in an alternating fashion (see, Kornell & Bjork, 2008; Guzman-Munoz, 2017; Kang & Pashler, 2012). Interleaving may enhance the inductive process through the juxtaposition of different categories, highlighting their most relevant differences (the discriminative-contrast hypothesis; Goldstone, 1996; Kornell & Bjork, 2008), or because it necessarily involves temporally spacing the exemplars of each category apart, which enhances memorization (Ebbinghaus, 1885; Cepeda et al., 2006). Similarly, other work has focused on promoting these inter-item comparisons not by presenting isolated items in a sequence, but in presenting multiple items simultaneously (Carvalho & Goldstone, 2014; Meagher et al., 2017; Nosofskyet al., 2022). Future work should investigate the relative contributions of memorization and comparisons processes. It is possible that in this study, retrieval practice benefited inductive learning by affecting the comparison processes, memorization, or both. At a minimum, studies on category learning should continue to investigate or account for processes involved in long-term memorization more thoroughly.

Retrieving entire exemplars from memory verbatim would not be feasible in all circumstances. For example, in the studies that use categories of paintings, it would be impractical or impossible to ask participants to paint an exemplar from memory perfectly (e.g., Picasso's *La Guernica*). The same could be said for any stimuli that are much more perceptually rich and complex compared to the molecular diagrams we used in this study. Indeed, only a small subset of people are gifted artists. However, it is conceivable that the entirety of an exemplar does not need to be retrieved for there to be a benefit of this kind of retrieval practice. The advantages of retrieval practice do not depend on people writing or drawing out their responses, but still occur if people recall or mentally imagine the solicited information in their mind (e.g., Kang, 2010). Therefore, even if people are constrained in their artistic ability, they could still benefit from the retrieval of exemplars from memory.

Applied considerations

Retrieval practice is commonly lambasted as a tool that is far more useful for rote memorization than fostering the transfer of learning. From an applied perspective, a positive benefit of retrieval practice on backward transfer would likely not counteract this view. In educational contexts, learning must be transferable across far more substantial divergences in the contexts of the acquisition of knowledge and assessments of performance. If transfer can be represented on a continuum spanning from learning extending from no change in context (memorization) to increasingly larger changes in context (more appreciative transfer), then backward transfer could be regarded as far closer to the memorization side of that continuum.

However, the results of this study suggest that backward transfer may not simply reflect verbatim memorization, but the abstraction of the kinds of underlying rules, principles, or concepts that can foster transfer across a wide range of contexts. In chemistry, for example, it is impossible to memorize the reactive properties of every single molecule. Simply memorizing a set of examples is unlikely to transfer to tasks that require consideration of the properties of molecules that were not memorized, let alone how they interact with one another in a reaction. Knowledge of the rules that govern these properties, though, are not only easier to obtain, but can be applied to understanding molecules that were never directly memorized.

The results of this study suggest that fostering flexible memorization and the induction of rules can be accomplished simultaneously (e.g., Agarwal, 2019). However, care must be taken to ensure that practice tests do not solely probe memory, but also incentivize the discovery of rules. This incentive does not need to be explicit. Rather, the inductive process can occur through making spontaneous inter-item comparisons. Such comparisons are likely to occur when questions on a test share superficial and/or deeper structural similarity (Gick & Holyoak, 1983; Holyoak & Koh, 1987).

Funding

This research was sponsored by the U.S. Army DEVCOM Soldier Center and was accomplished under Cooperative Agreement Number W911QY-19-2-0003. The opinions expressed herein are those of the authors and do not reflect those of the United States Army. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

CRediT authorship contribution statement

Gregory I. Hughes: Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization, Methodology, Formal analysis, Investigation. **Ayanna K. Thomas:** Writing – review & editing, Visualization, Project administration, Funding acquisition, Conceptualization, Methodology, Formal analysis, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data, analysis materials, and stimuli are available through the Open Science Framework (https://osf.io/tkmv3/).

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

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189–209. https://doi.org/10.1037/edu0000282

Barenberg, J., Berse, T., Reimann, L., & Dutke, S. (2021). Testing and transfer: Retrieval practice effects across test formats in english vocabulary learning in school. *Applied Cognitive Psychology, 35*(3), 700–710. https://doi.org/10.1002/acp.3796

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, 15(4), 657–668. https://doi.org/10.1037/0278-7393.15.4.657
Bulevich, J. B., Thomas, A. K., & Parsow, C. (2016). Filling in the gaps: Using testing and restudy to promote associative learning. *Memory, 24*(9), 1267–1277. https://doi.org/10.1080/09658211.2015.1098706
Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604
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*(6), 1563–1569. https://doi.org/10.1037/a0017021
Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547–1552. https://doi.org/10.1037/a0024140
Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268–276. https://doi.org/10.3758/BF03193405
Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*(3), 443–448. https://doi.org/10.3758/s13423-012-0221-2
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
Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826–830. https://doi.org/10.3758/BF03194004
Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008
Carroll, M., & Nelson, T. O. (1993). Effect of overlearning on the feeling of knowing is more detectable in within-subject than in between-subject designs. *The American Journal of Psychology, 106*(2), 227–235. https://doi.org/10.2307/1423169
Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0
Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354
Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553
Cheng, C.K. (2014). Effect of multiple-choice testing on memory retention—Cue-target symmetry (Order No. AAI3666589). Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/65649/1/Cheng_Cho_Kin_201406_PhD_thesis.pdf.
Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and Prototypicality in Category Learning: A Comparison of Inference Learning and Classification Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 216–226. https://doi.org/10.1037/0278-7393.30.1.216
Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition, 8*(2), 166–177. https://doi.org/10.1016/j.jarmac.2019.01.003
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. https://doi.org/10.3389/fpsyg.2014.00781
D'Ydewalle, G. (1981). Test expectancy effects in free recall and recognition. *Journal of General Psychology, 105*, 173–195. https://doi.org/10.1080/00221309.1981.9921071
Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Leipzig, Germany: Duncker & Humblot.
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
Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*(5), 608–628. https://doi.org/10.3758/BF03201087
Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(3), 371–377. https://doi.org/10.1037/0278-7393.15.3.371
Guzman-Munoz, F. (2017). The advantage of mixing examples in inductive learning: A comparison of three hypotheses. *Educational Psychology, 37*(4), 421–437. https://doi.org/10.1080/01443410.2015.1127331
Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121
Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411
Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15*(4), 332–340. https://doi.org/10.3758/BF03197035
Hughes, G. I., Taylor, H. A., & Thomas, A. K. (2018). Study techniques differentially influence the delayed judgment-of-learning accuracy of adolescent children and college-aged adults. *Metacognition and Learning, 13*(2), 109–126. https://doi.org/10.1007/s11409-018-9180-y
Hughes, G. I., & Thomas, A. K. (2021). Visual category learning: Navigating the intersection of rules and similarity. *Psychonomic Bulletin & Review, 28*(3), 711–731. https://doi.org/10.3758/s13423-020-01838-0
Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441–1451. https://doi.org/10.1037/a0020636
JASP Team (2022). JASP (Version 0.16. 3) [Computer software].
Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
Jones, E. L., & Ross, B. H. (2011). Classification versus inference learning contrasted with real-world categories. *Memory & Cognition, 39*(5), 764–777. https://doi.org/10.3758/s13421-010-0058-8
Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30*(6), 823–840. https://doi.org/10.3758/BF03195769
Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*(8), 1009–1017. https://doi.org/10.3758/MC.38.8.1009
Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97–103. https://doi.org/10.1002/acp.1801
Karpicke, J.D., Lehman, M., & Aue, W.R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), The psychology of learning and motivation (vol. 61); the psychology of learning and motivation (vol. 61) (pp. 237-284, Chapter x, 330 Pages) Elsevier Academic Press, San Diego, CA.
Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.609
Kornell, N., & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? *Psychological Science, 19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x
Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*(5), 493–501. https://doi.org/10.1080/09658210902832915
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22
Lancaster, M. E., Shelhamer, R., & Homa, D. (2013). Category inference as a function of correlational structure, category discriminability, and number of available cues. *Memory & Cognition, 41*(3), 339–353. https://doi.org/10.3758/s13421-013-0371-0
Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology, 110*(2), 203–217. https://doi.org/10.1037/edu0000211
Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian Cognitive Modeling: A Practical Course. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139087759
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*(6), 1787–1794. https://doi.org/10.1037/xlm0000012
Litt, R. A., & Nation, K. (2014). The nature and specificity of paired associate learning deficits in children with dyslexia. *Journal of Memory and Language, 71*(1), 71–88. https://doi.org/10.1016/j.jml.2013.10.005
Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309
Meagher, B. J., Carvalho, P. F., Goldstone, R. L., & Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review, 24*(6), 1987–1994. https://doi.org/10.3758/s13423-017-1251-6
McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x
McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154
McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied, 21*(4), 370–382. https://doi.org/10.1037/xap0000063
McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*(1), 3–21. https://doi.org/10.1037/xap0000004
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207
Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*(2), 242–279. https://doi.org/10.1016/0010-0285(87)90012-0
Minda, J. P., & Smith, J. D. (2001). Prototypes in Category Learning: The Effects of Category Size, Category Structure, and Stimulus Complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775
Molander, B., & Garvill, J. (1979). The invariance of asymmetric cross-modal transfer effects. *Scandinavian Journal of Psychology, 20*(3), 171–177. https://doi.org/10.1111/j.1467-9450.1979.tb00698.x
Mondani, M. S., & Battig, W. F. (1973). Imaginal and verbal mnemonics as related to paired-associate learning and directionality of associations. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 401–408. https://doi.org/10.1016/S0022-5371(73)80018-0
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nosofsky, M., Meagher, B. J., & Kumar, P. (2022). Contrasting exemplar and prototype models in a natural-science category domain. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*. https://doi.org/10.1037/xlm0001069

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53–79. https://doi.org/10.1037/0033-295X.101.1.5

Paivio. (1971). *Imagery and verbal processes.* Holt, Rinehart and Winston.

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108*(4), 563–575. https://doi.org/10.1037/edu0000074

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3), 278–292. https://doi.org/10.1037/xap0000124

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition, 44*(1), 24–36. https://doi.org/10.3758/s13421-015-0547-x

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36–48. https://doi.org/10.3758/s13421-012-0245-x

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). *Why testing improves memory: Mediator effectiveness hypothesis*. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review, 25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Rickard, T. C., & Pan, S. C. (2020). Test-enhanced learning for pairs and triplets: When and why does transfer occur? *Memory & Cognition, 48*(7), 1146–1160. https://doi.org/10.3758/s13421-020-01048-y

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233–239. https://doi.org/10.1037/a0017678

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide.* Pittsburgh: Psychology Software Tools Inc.

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4*(2), 131–135. https://doi.org/10.1037/0882-7974.4.2.131

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Stewart, N., & Brown, G. D. A. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology, 49*(5), 403–409. https://doi.org/10.1016/j.jmp.2005.06.001

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 49*(4), 901–918. https://doi.org/10.1080/027249896392351

Thomas, A. K., Smith, A. M., Kamal, K., & Gordon, L. T. (2020). Should you use frequent quizzing in your college course? giving up 20 minutes of lecture time may pay off. *Journal of Applied Research in Memory and Cognition, 9*(1), 83–95. https://doi.org/10.1016/j.jarmac.2019.12.005

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*(3), 429–442. https://doi.org/10.3758/s13421-012-0274-5

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive. Science, 3*(3), 499–535. https://doi.org/10.1111/j.1756-8765.2010.01113.x

van den Bergh, D., Wagenmakers, E., & Aust, F. (2022). Bayesian Repeated-Measures ANOVA: An Updated Methodology Implemented in JASP. *Preprint available on PsyArXiv*.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., … Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic bulletin & review, 28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying-abstraction model of categorization. *Psychonomic Bulletin & Review, 15*(4), 732–749. https://doi.org/10.3758/PBR.15.4.732

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*(9), 1127–1131. https://doi.org/10.1177/0956797611417724

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14–26. https://doi.org/10.1016/j.jml.2014.04.004

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics, 68*(4), 643–654. https://doi.org/10.3758/BF03208765

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian Perspective on the Bonferroni Adjustment. *Biometrika, 84*(2), 419–427. https://doi.org/10.1093/biomet/84.2.419

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* ((3rd ed.).). McGraw-Hill.

Winn, W. (1991). Learning from maps and diagrams. *Educational Psychology Review, 3*(3), 211–247. https://doi.org/10.1007/BF01320077

Winn, W. D., & Sutherland, S. W. (1989). Factors influencing the recall of elements in maps and diagrams and the strategies used to encode them. *Journal of Educational Psychology, 81*(1), 33–39. https://doi.org/10.1037/0022-0663.81.1.33

Wollen, K. A., & Lowry, D. H. (1971). Effects of imagery on paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 10*(3), 276–284. https://doi.org/10.1016/S0022-5371(71)80055-5

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*(1), 124–148. https://doi.org/10.1006/jmla.1998.2566

Yang, C., Chew, S., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology, 111*(5), 809–826. https://doi.org/10.1037/edu0000320

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(3), 485–492. https://doi.org/10.1037/xlm0000449

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition, 41*(1), 16–27. https://doi.org/10.3758/s13421-012-0238-9