Journal of Experimental Psychology: Learning, Memory, and Cognition

Levels of Retrieval and the Testing Effect

Ningxin Su, Zachary L. Buchin, and Neil W. Mulligan

Online First Publication, October 15, 2020. http://dx.doi.org/10.1037/xlm0000962

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

ISSN: 0278-7393

http://dx.doi.org/10.1037/xlm0000962

Levels of Retrieval and the Testing Effect

Ningxin Su
Beijing Normal University

Zachary L. Buchin and Neil W. Mulligan
University of North Carolina at Chapel Hill

Retrieval enhances subsequent memory more than restudy (i.e., the testing effect), demonstrating the encoding (or reencoding) effects of retrieval. It is important to delineate the nature of the encoding effects of retrieval especially in comparison to traditional encoding processes. The current study examined if the level of retrieval, analogous to the level of processing during encoding, has an effect on subsequent memory. In 4 experiments, participants studied short lists of words, each followed by a retrieval or restudy trial. A final free recall test was given at the end of the experiment. The level of retrieval was manipulated by asking participants to retrieve words with a semantic or phonemic cue in the retrieval trial. In order to isolate the effects of retrieval per se, the semantic or phonemic cue was also presented in the restudy trial. Experiment 1 manipulated levels of retrieval (and restudy) between subjects while Experiment 2 manipulated levels within subjects. Experiment 3 sought to enhance the levels effect by adding an overt levels judgment, and Experiment 4 sought to rule out an alternative account of the equality of the testing effects across levels by increasing the list length. In all 4 experiments, a robust testing effect was obtained but it was not moderated by level of retrieval, a result supported by a small-scale meta-analysis, which demonstrated an overall effect of levels and testing condition, but no interaction.

Keywords: levels of retrieval, testing effect, levels of processing

Testing is not simply a tool that reveals the contents of memory but has also proven to be a powerful way to enhance memory (i.e., the testing effect, for recent reviews see Adesope, Trevisan, & Sundararajan, 2017; Karpicke, 2017; Rowland, 2014). In a typical testing effect experiment, participants first study some items and then either restudy or practice retrieving those items. On a final memory test, performance is typically better for the previously retrieved items than the restudied items (e.g., Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006). In fact, the testing effect has emerged as one of the most robust memory phenomena in cognitive psychology (e.g., Karpicke, 2017). The learning benefits of retrieval practice have been replicated in both the laboratory and the classroom (e.g., Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006), and with various types of materials, including single words and word pairs (e.g., Carpenter & DeLosh, 2006; Carpenter, Pashler, & Vul, 2006), text passages, (e.g., Butler, 2010; Chan, McDermott, & Roediger, 2006), and academic facts (e.g., Carpenter, Pashler, & Cepeda, 2009). Further, the testing effect generalizes across final test type (e.g., recognition, cued recall, and free recall; Carpenter & DeLosh, 2006) as well as retention interval (e.g., minutes, days, weeks, and months; Kornell, Bjork, & Garcia, 2011; Rowland & DeLosh, 2015).

The enhanced memory retention from retrieval practice suggests that retrieval modifies memory representations (e.g., Bjork, 1975). In line with this idea, recent research has examined the similarity between what might be referred to as the encoding (or reencoding) effects of retrieval and those processes more typically labeled as encoding (Buchin & Mulligan, 2017, 2019; Mulligan & Picklesimer, 2016). For example, it is well-known that typical encoding processes (e.g., studying and reading) can be easily disrupted by dividing attention (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996; Mulligan, 2008). Mulligan and colleagues conducted multiple studies to determine if the encoding benefits of retrieval were similarly impaired under divided attention. Participants were asked to retrieve or restudy previously studied items under full or divided attention. Across a variety of conditions, final test performance indicated that the encoding effects of retrieval were more resilient to disruption from divided attention than the effects of study-based encoding (Buchin & Mulligan, 2017, 2019; Mulligan & Picklesimer, 2016). In the present study, we examined another classical factor that moderates the effectiveness of typical encoding processes—the depth or level of processing (Craik & Lockhart, 1972; Craik & Tulving, 1975).

Ningxin Su, Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University; Zachary L. Buchin and Neil W. Mulligan, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill.

This study was supported by the Natural Science Foundation of China [Grant 31671130].

Correspondence concerning this article should be addressed to Ningxin Su, Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, No.19, Xinjiekouwai Street, Haidian District, Beijing 100875, China, or to Neil W. Mulligan, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 East Cameron Avenue, Chapel Hill, NC 27599-3270. E-mail: suningxin@mail.bnu.edu.cn or nmulligan@unc.edu

Levels of processing is one of the most widely applied concepts in memory research (e.g., Cermak & Craik, 1979; Craik, 2002; Craik & Lockhart, 1972; Craik & Tulving, 1975; Rose & Craik, 2012; Rose, Myerson, Roediger, & Hale, 2010). The main idea is that deep processing (semantic analysis including meaning, inference, implication) produces better memory than shallow processing (superficial analysis including processing of structural, perceptual, or syntactic features). The memorial advantage of deep processing generalizes across the type of final test (recognition or free recall), the nature of instruction (incidental learning or intentional learning), the duration of study time (200 ms or 6 s), and the reward (with or without; Craik & Tulving, 1975). Further, the levels of processing framework emphasizes that the memorial enhancement from deeper processing is not simply due to increased effort, difficulty, or processing time (Craik & Tulving, 1975). Considering the robust levels of processing effect on encoding, the first aim of the present study was to determine if there is an analogous effect on the encoding effects of retrieval—what might be referred to as an effect of levels of retrieval on subsequent memory.[1]

Although the idea that deep or semantic processing produces better memory than shallow or nonsemantic processing has been extensively assessed (e.g., Craik, 2002; Craik & Tulving, 1975), the potential influence of processing level on the mnemonic benefits of retrieval is not as well understood. Specifically, does deep, semantic retrieval enhance memory retention more than shallow, nonsemantic retrieval?

Theoretical accounts of the testing effect provide some suggestions about the likely answer to this question. The elaborative retrieval account (Carpenter, 2009, 2011; Carpenter & Yeung, 2017) proposes that the effectiveness of retrieval lies in its ability to activate semantic associates of the cue and target words, which can then be used as additional retrieval routes during subsequent retrieval. For example, if the cue-target pair *mother-child* is to be learned, then practicing retrieval with the cue *mother-* leads to the activation of semantic associates like *father* whereas restudying *mother-child* is less likely to lead to such semantic elaborations and in turn less likely to produce effective semantic mediators for later retrieval (Carpenter & Yeung, 2017). Given that deep processing involves semantic analysis and shallow processing involves nonsemantic analysis, it can be inferred that deep processing during retrieval provides more opportunities for semantic elaboration which would benefit retention compared to shallow processing during retrieval. Thus, this account suggests that deep retrieval should enhance later memory more than shallow retrieval.

A second hypothesis argues that retrieval entails more effortful processing of the target stimulus than does restudy, and that this difference in effortful processing produces the testing effect (e.g., Bjork, 1975; Endres & Renkl, 2015; Pyc & Rawson, 2009). This account is consistent with the desirable difficulties framework (e.g., Bjork, 1994, 1999) and the finding that more difficult retrieval conditions can enhance the size of the testing effect (e.g., Halamish & Bjork, 2011; Pyc & Rawson, 2009). However, if retrieval effort or difficulty is equated as in the present study, the hypothesis does not predict an effect of retrieval level.

Finally, the episodic context account proposes that successful retrieval updates the contextual representation of targets by including features from both the original study context and the present test context (e.g., Karpicke, Lehman, & Aue, 2014; Whiffen & Karpicke, 2017). The resulting composite trace in the retrieval condition provides varied contextual information that is more likely to match whatever contextual cues are used during the final recall test, restricting the search set of candidate information to a greater degree than in the restudy condition (Karpicke, 2017; see Lehman & Karpicke, 2016 for contrast with the elaborative retrieval account). Because semantic and nonsemantic retrieval both provide the possibility for contextual updating, the theory provides no reason to predict a difference between the two conditions.

Thus, the elaborative account predicts a levels-of-retrieval effect, the effort account proposes no effect (provided retrieval difficulty is equated), and the episodic-context account does not make a clear prediction but seems most consistent with an absence of such an effect. Finally, traditional theories of memory encoding, such as the levels-of-processing approach, predict a levels-of-retrieval effect. That is, it is reasonable to assume that semantic retrieval generally entails semantic processing to a greater degree than nonsemantic retrieval, which in turn should enhance still later memory.

Several studies predating the current interest in the testing effect manipulated levels during initial retrieval, and either found a mnemonic advantage of semantic retrieval or no difference between the semantic and nonsemantic conditions on a final memory test (Bartlett, 1977; Bartlett & Tulving, 1974; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985; Whitten, 1978). Bartlett (1977) asked participants to study lists of six words and after each immediately recall three of the words when given: temporal cues about the serial positions of the targets; orthographic cues consisting of the final one to three letters of the targets; or semantic cues that were meaningfully related to the targets. Memory performance on a final free recall test was influenced by the initial retrieval mode; retrieval with a semantic cue enhanced performance the most, independent of serial position (Bartlett, 1977). Whitten (1978) found a similar result using rhyme cues in the shallow retrieval condition. In contrast, neither McDaniel and Masson (1985) nor McDaniel, Kowitz, and Dunay (1989) found a consistent memorial benefit from semantic processing during retrieval practice compared to phonological (nonsemantic) processing. Similarly, although Bartlett and Tulving (1974) observed a benefit of initial semantic cuing over temporal cuing when retrieving from short-term memory (STM), there was no difference when retrieving from long-term memory.

Although these studies provide preliminary information, certain methodological factors prevent unambiguous conclusions regarding retrieval-based learning. First, none of the studies included restudy control conditions to compare to the retrieval practice conditions (Bartlett, 1977; Bartlett & Tulving, 1974; McDaniel et

[1] Two points merit comment. First, it is well known that the effects of levels of processing can be affected by the nature of the later memory test, with perceptually-driven tests often producing a reversed levels-of-processing effect (captured by the idea of transfer appropriate processing, Morris et al., 1977). In the present explorations, the final test is free recall, a test that virtually always exhibits the usual levels-of-processing effect. Second, research on source-constrained retrieval (e.g., Jacoby, Shimizu, Daniels, & Rhodes, 2005) has used the term *retrieval depth* to describe the qualitative level of processing at retrieval produced by different levels of processing during initial encoding. However, we are examining a different question—the effect of processing level during retrieval practice on the size of the testing effect.

al., 1989; McDaniel & Masson, 1985; Whitten, 1978). This is a critical issue in isolating the effects of retrieval, per se, on later memory. Final recall in a retrieval condition can be influenced by prior retrieval itself and by the reexperience with the stimulus that results from retrieval (or from feedback after a retrieval attempt). To isolate the effect of retrieval itself, one must compare the retrieval condition with an appropriate restudy condition, which controls for reexperience (see Buchin & Mulligan, 2017; Roediger & Karpicke, 2006; Rowland, 2014, for discussion). Consequently, the foregoing studies do not isolate the effects of specific levels of retrieval, per se, on subsequent memory performance. This is especially important in the present case because retrieval with semantic cues entails not just retrieval but also reprocessing of the target item in the company of a semantic cue. Likewise, in the shallow retrieval condition, retrieval and additional reprocessing occurs in the company of a shallow cue. Reprocessing a stimulus with a semantic cue may enhance memory more than reprocessing a stimulus with a shallow cue. If so, any differences between the semantic and shallow retrieval conditions might reflect different mnemonic effects of retrieval or of reprocessing in the presence of different types of cues (or both). To eliminate this confound, the mnemonic benefits of each retrieval condition must be assessed relative to its appropriate restudy baseline.

Second, initial retrieval success was generally higher for items in the semantic condition than in the nonsemantic condition (Bartlett, 1977; Bartlett & Tulving, 1974; McDaniel et al., 1989; McDaniel & Masson, 1985; Whitten, 1978). This raises two countervailing concerns. First, conditions that produce greater initial retrieval are more likely to demonstrate higher recall on the final test (e.g., Kang, McDermott, & Roediger, 2007; Karpicke et al., 2014). This means that any differences between the semantic and nonsemantic retrieval conditions on the final test could reflect levels-of-retrieval effects, the beneficial effects of greater successful retrieval practice, or both. Second, the lower level of initial retrieval success in the nonsemantic condition indicates that it was generally more difficult than the semantic condition. This adds further ambiguity because the mnemonic benefits of successful retrieval practice (without feedback) increase as the difficulty (or effort) of retrieval increases (e.g., Halamish & Bjork, 2011; Pyc & Rawson, 2009). Taken together, the two concerns indicate that these prior studies are unable to provide a conclusive answer regarding the effect of retrieval level on later memory.[2]

To the best of our knowledge, only one study has equated initial retrieval success and used a restudy control condition to examine the influence of processing depth during retrieval practice on the size of the testing effect (Veltre, Cho, & Neely, 2015). In this experiment, Veltre, Cho, and Neely (2015) assessed the transfer appropriate processing account (Morris, Bransford, & Franks, 1977) as a possible explanation of the testing effect, according to which memory improves if the type of processing used during the final test matches the type of processing used during learning. Specifically, participants studied words (ABOVE) and then either restudied them (ABOVE), retrieved them given a semantic cue (BEYOND-____), or retrieved them given an orthographic cue (A_OV_). Two days later, participants took a cued-recall test using cues identical to the initial retrieval cue, new cues of the same level, or new cues of a different level. To assess the size of the testing effect in the semantic final test cue conditions, Veltre et al. (2015) subtracted performance in the restudy condition (restudy ABOVE, final test cue BEYOND-____) from the identical (BEYOND-____, BEYOND-____), same level (BELOW-____, BEYOND-____), and different level (A_OV_, BEYOND-____) retrieval practice groups. A similar analysis was conducted for the orthographic final test cue conditions; performance in the restudy condition (restudy ABOVE, final test cue A_OV_) was subtracted from performance in the identical (A_OV_, A_OV_), same level (AB_V_, A_OV_), and different level (BEYOND-____, A_OV_) retrieval practice groups. Veltre et al. (2015) found larger testing effects in the semantic final test cue conditions when the initial retrieval cue was semantic (identical or same level) rather than orthographic, but there were no clear differences between the orthographic final test cue conditions.

The results seem to suggest that semantic retrieval leads to larger testing effects than nonsemantic retrieval, as long as the final test cue also induces semantic retrieval (Veltre et al., 2015). This in turn suggests that the encoding effects of retrieval exhibit the same sort of levels-of-processing effect routinely found when the manipulation is implemented during initial study (again, provided the final test is conceptual in nature, Morris et al., 1977). However, there is a limitation with the study. The restudy condition used to calculate the testing effects presented the study item in isolation, unaccompanied by the cue presented in the retrieval conditions. This confound may be critical, as the restudy condition did not explicitly prompt the same level of processing as did the retrieval conditions. The mere presence of the (semantic or orthographic) cue might have an influence on processing in the retrieval conditions beyond any specific effect of retrieval. The role of the restudy condition is to eliminate just those differences to allow an assessment of the mnemonic effects of retrieval. For example, using an appropriate restudy comparison condition for orthographic retrieval that guides encoding of orthographic information might have revealed a comparable testing effect to that from semantic retrieval. The cues given to the retrieval practice conditions should be presented in the restudy conditions along with the target words to enhance comparability between the test and restudy conditions, and isolate the effects of a specific level of processing on the testing effect.

To address the foregoing issues, we conducted four experiments using an adaptation of the design used by Whitten (1978). In this design, participants are presented with a series of short lists of (e.g., three or four) words, each followed by a brief (e.g., 30 s) distractor task and a retrieval or restudy trial. A final test is given at the end of the experiment. This design is useful in the present case for two reasons. First, it produces high retrieval success during initial retrieval, allowing for a robust testing effect even with brief delays and maximizing the influence of retrieval, per se, on the final recall test (e.g., Kuo & Hirshman, 1996). Second, in designs using long study lists, retrieval with shallow cues is often

[2] It should be noted that a majority of these studies conducted additional analyses conditionalizing final performance on successful initial retrieval in an attempt to mitigate item-selection effects, including analyzing *a* scores (McDaniel et al., 1989; see Lockhart, 1975) and *s* scores (Bartlett, 1977; McDaniel & Masson, 1985; see Modigliani, 1976). However, the conditionalized analyses do not correct for the possibility that more difficult initial retrieval can produce a more potent effect for those items successfully retrieved. Even more important, these analyses do not remedy the more critical issue of the lack of appropriate restudy comparison conditions.

quite low compared with retrieval with semantic cues (Morris et al., 1977; Mulligan & Picklesimer, 2012). In the present case, it is desirable to use a paradigm in which initial retrieval is high and comparable across the two retrieval conditions (see Veltre et al., 2015, for an alternative strategy). In the present experiments, participants were asked to learn a series of words for a later memory test. During the learning phase, participants briefly studied a set of three words, solved math problems for 30 s, and then either restudied or retrieved two of the three words in the presence of semantic or rhyme cue words. After repeating this for all the to-be-learned words, participants completed a 5-min filler task and then took a final free recall test. Because this differs from a traditional levels-of-processing manipulation which occurs during initial study, we refer to our manipulation simply as "levels" and do not use the term "levels-of-processing."

Experiment 1

The experiments addressed two issues in examining the effect of level of retrieval on the testing effect. First, as stressed above, differences in reexperience with study items can influence the size of the testing effect; the testing effect increases with greater reexperience induced either by feedback (vs. no feedback) or high initial retrieval performance (vs. low performance; Rowland, 2014). To control the influence of reexperience (and eliminate any reexperience confounds), feedback was given after initial retrieval to make sure that all conditions reexperienced all items. Second, retrieval effort contributes to the testing effect with easier retrieval leading to less memory enhancement (Pyc & Rawson, 2009). In some of the earlier studies, initial retrieval performance was better with semantic than nonsemantic cues (Bartlett, 1977; Bartlett & Tulving, 1974; but see Veltre et al., 2015), implying that semantic retrieval was easier than nonsemantic retrieval for these tasks, which in turn might have inflated the mnemonic benefit of the latter form of retrieval. In light of this issue, we conducted pilot research to develop methods under which the semantic and nonsemantic cues produced approximately equal initial retrieval accuracy and retrieval time. The methods of the current experiments were based on the pilot study and the results indicate that difficulty was generally matched.[3]

The purpose of Experiment 1 was to ascertain whether semantic retrieval produces a larger testing effect than phonemic retrieval when compared to the appropriate baseline restudy condition. Level at retrieval or restudy was manipulated between-subjects. Participants in the semantic group either retrieved or restudied words with a semantic cue or restudied words in isolation; participants in the phonemic group engaged in the same tasks except with rhyme cues (or no cue in the isolated restudy condition). It should be noted that we included a restudy condition without cues to determine if restudying with a semantic or phonemic cue affected the baseline comparison for computing the testing effect (cf. Veltre et al., 2015).

Method

Participants. A power analysis was conducted based on an effect size of the testing effect from a previous experiment that used similar initial and final tests (i.e., $d = 0.75$ from Mulligan & Peterson, 2015, Experiment 3), it was found that 21 participants per group were needed to detect an effect of that size with 90% power ($\alpha = .05$, two-tailed). Therefore, 43 participants from UNC at Chapel Hill were recruited in exchange for course credit. One participant was excluded because of a computer error. The remaining 42 participants (30 females; age $M = 18.68$, $SD = 0.99$, one participant did not provide their age) were randomly assigned to the semantic group ($n = 21$) or the phonemic group ($n = 21$). The study received research ethics committee (Instructional Review Board) approval.

Design and materials. The experiment used a 2 (level: semantic vs. phonemic) × 3 (review condition: retrieval practice vs. cued restudy vs. restudy) design with level as a between-subjects variable and review condition as a within-subjects variable.

All materials were drawn from Nelson, McEvoy, and Schreiber (2004). Forty-two critical target words were selected. The average length, frequency, and concreteness for target words was $M = 4.33$ ($SD = 0.76$, *Range* = 3 to 6), $M = 210.78$ ($SD = 360.74$, *Range* = 3 to 1,599), and $M = 4.98$ ($SD = 1.18$, *Range* = 2.93 to 6.92), respectively. For each target word (e.g., cold), we selected a semantic cue (e.g., chill) and a phonemic cue (e.g., hold). The average length, frequency and concreteness for semantic cues was $M = 5.17$ ($SD = 0.77$, *Range* = 3 to 6), $M = 48.97$ ($SD = 81.05$, *Range* = 1 to 373), and $M = 4.97$ ($SD = 1.16$, *Range* = 2.48 to 6.94), respectively. The average length, frequency and concreteness for phonemic cues was $M = 4.39$ ($SD = 1.65$, *Range* = 3 to 10), $M = 72.81$ ($SD = 144.37$, *Range* = 2 to 794), and $M = 4.67$ ($SD = 1.35$, *Range* = 1.49 to 6.96), respectively. Semantic and phonemic cues did not differ in frequency or concreteness ($ps > .1$), but the semantic cues were almost one letter longer on average ($p = .01$). The average forward strength and backward strength of semantic cue–target pairs was $M = 0.67$ ($SD = 0.11$, *Range* = 0.5 to 0.91) and $M = 0.21$ ($SD = 0.24$, *Range* = 0 to 0.87), respectively.

Twenty-one nontarget words (with similar features as the target items) were added to the set of 42 target words to create 21 short lists of three words each (e.g., cold, fish, leave), with each study list containing two target words and one nontarget word. Although all three words in a list were initially studied, only the two targets were subsequently reviewed via retrieval, cued restudy, or restudy.[4] To be clear, the specific words that acted as targets or nontargets in each list were the same for all participants (and the

[3] As will be seen, there are two comparisons assessing difficulty in each experiment, one based on retrieval accuracy and the other on retrieval time, yielding a total of eight comparisons across the four experiments. Two of the comparisons favored semantic cues (higher retrieval accuracy than for nonsemantic in Experiments 1 and 4), one favored the nonsemantic cues (quicker retrieval times compared to the semantic cues in Experiment 3), and five of the comparisons found no significant difference. In sum, the semantic and nonsemantic retrieval tasks do not appear to differ substantially in difficulty.

[4] We included three words in each study list to require selective retrieval guided by the supplied cue. We required retrieval practice for only two of the three items to reduce the time for forgetting after the study list presentation in order to facilitate high initial retrieval success (on the assumption that retrieval success would decrease over more retrieval trials). The use of two retrieval trials in turn dictated the comparable structure for restudy trials. Thus, this design represents a compromise between the needs for a slightly longer study list (to induce selective retrieval) and a smaller number of retrievals (to facilitate higher initial retrieval success). See Experiment 4 for more on this issue.

nontarget was equally often in each serial position). Of the 21 total lists, three were used for practice and six were used in each of the three review conditions. The lists were counterbalanced across subjects such that each was used equally often in each condition.

During review, targets were either retrieved with a semantic (e.g., chill–?) or phonemic (rhyme) cue (e.g., hold–?), restudied with a semantic (e.g., chill–cold) or phonemic (e.g., hold–cold) cue, or restudied without a cue (e.g., cold). Participants were asked to retrieve or read each target word aloud and a microphone was used to record response times.

Procedure. The main experiment consisted of a learning phase (Phase 1; see Figure 1) and a testing phase (Phase 2). Before Phase 1, participants completed a preliminary voice key calibration to test the quality of their oral responses. Five words were presented on a computer screen (2 s each) and participants were instructed to say each word aloud into a microphone as quickly and accurately as possible. If the microphone successfully recorded their response, the computer displayed "Correct!" as feedback. Otherwise, participants were informed to read the next word louder.

After successful voice calibration, participants were told that they would be presented with a series of words that should be studied for an upcoming memory test. They were given a basic overview of Phase 1 and ran through the procedure in each review condition using the three practice lists. The entire procedure depicted in Figure 1 was repeated for each list, one list at a time, in a random order.

First, a fixation cross was presented for 1 s followed by a blank screen lasting 0.5 s. Then the three words in a given list were presented one at a time (2 s each) in a random order for initial study (e.g., cold, fish, leave). After studying each of the three words, another fixation cross was presented for 1 s before participants were asked to solve math problems for 30 s, typing their answers into the computer. The maximum time to solve each problem was 8 s.

After the math problems, the review portion of the learning phase began. First, participants were shown instructions for 3 s that indicated which type of review was going to take place. For the semantic retrieval condition, the instruction "retrieve with a semantic cue" was shown on the computer, followed by a fixation cross (1 s). Participants then saw a semantic cue for 4 s (e.g., chill–?) and were instructed to retrieve the associated studied word, saying it aloud into the microphone as quickly and accurately as possible. Regardless of their response, the correct target word was then displayed as feedback for 2 s (e.g., chill–cold) and participants were asked to read the word aloud if they had not successfully recalled it. This was repeated for a second word from the studied list (e.g., vacate–? for 4 s, vacate–leave for 2 s), before the learning phase was repeated with a new list of three words.

For the semantic cued restudy condition, the instruction "restudy with a semantic cue" was shown on the computer screen (for 3 s followed by the 1-s fixation cross) before a semantic cue-target pair was presented for 6 s (e.g., chill–cold). Participants were asked to read the cue word silently and the target word aloud. After 6 s, the second target word from the studied list was presented for restudy alongside its corresponding semantic cue (e.g., vacate–leave). The phonemic retrieval and cued restudy conditions were similar to their semantic counterparts but used phonemic cues (e.g., hold–cold, weave–leave) instead of semantic cues. In the (uncued) restudy condition, the instruction "restudy" was displayed for 3 s. before one of the target words appeared without a cue for 6 s (e.g., cold). Participants were asked to read the target word aloud. Afterward, the second target word was displayed and restudied for 6 s (e.g., leave).

After completing Phase 1 for all lists, participants were asked to solve math problems for 5 min before starting Phase 2, which consisted of the final free recall test. Participants were given a blank sheet of paper and asked to write as many of the studied words as they could recall in 5 min. After the free recall test, participants answered a brief postexperiment questionnaire. Because the questions were secondary to the main point of the current study, the questions and results are presented in Appendix A.

Results

Initial cued recall. Mean cued recall accuracy and median retrieval times in Experiments 1–4 can be found in Table 1. The proportion of target words correctly recalled in the semantic group was significantly higher than in the phonemic group, $F(1, 40) = 9.197$, $MS_e = 0.026$, $p = .004$, $\eta_p^2 = .187$. There was no difference in median retrieval time between the semantic and phonemic groups, $F(1, 40) = 0.638$, $p = .429$. Occasionally, the vocal response did not trip the voice key (e.g., because a partici-

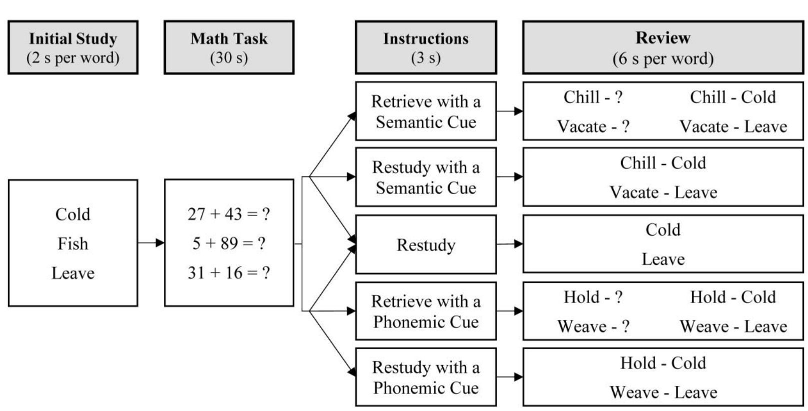

Figure 1. Learning phase (Experiment 1). After reviewing the second word, the entire process repeated with a new list of words until all lists were presented. In this example, *cold* and *leave* are targets and *fish* is a nontarget.

Table 1
Initial Cued Recall Proportion Correct and Median Retrieval RTs (ms): M (SD)

Experiment	Proportion correct		Median retrieval RT (ms)	
	Semantic	Phonemic	Semantic	Phonemic
Experiment 1	0.86 (0.15)	0.71 (0.17)	1434 (273)	1367 (275)
Experiment 2	0.78 (0.16)	0.81 (0.18)	1457 (337)	1413 (311)
Experiment 3	0.75 (0.21)	0.78 (0.18)	1578 (411)	1410 (285)
Experiment 4	0.72 (0.18)	0.63 (0.18)	1618 (360)	1515 (400)

pant spoke too softly) and thus did not record retrieval time for that particular trial (although accuracy could still be recorded for these trials). The percentage of missing retrieval times did not significantly differ between the semantic ($M = 2.69\%$, $SD = 4.37\%$) and phonemic ($M = 4.79\%$, $SD = 9.85\%$) groups, $F(1, 40) = 0.799$, $p = .377$.

Final free recall. There are two types of items, those referred to as targets, which were either retrieved or restudied, and those referred to as nontargets, which were the third item from each study list that was not restudied or retrieved. The final recall of nontargets is not important for our primary research questions and is reported for all experiments in Appendix B.

Final free recall performance for targets in Experiments 1–4 can be found in Table 2. A 2 (level: semantic vs. phonemic) × 3 (review condition: retrieval vs. cued restudy vs. restudy) mixed factorial analysis of variance (ANOVA) was conducted with level as a within-subjects factor and review condition as a between-subjects factor (see Table 2). Only the main effect of review condition was significant, $F(2, 80) = 10.209$, $MS_e = 0.019$, $p < .001$, $\eta_p^2 = .20$. Post hoc tests revealed that retrieval practice enhanced final free recall compared with both cued restudy, $t(41) = 2.517$, $p_{bonf} = .047$, $d = 0.388$, and restudy, $t(41) = 4.615$, $p_{bonf} < .001$, $d = 0.712$, demonstrating the testing effect. There was no significant difference between the cued restudy and restudy conditions, $p_{bonf} = .175$. Although recall in all three semantic conditions was numerically greater than the corresponding three phonemic conditions, the main effect of level was nonsignificant, $F(1, 40) = 2.685$, $p = .109$. Most critically, the interaction between level and review condition was also nonsignificant, $F(2, 80) = 0.142$, $p = .868$, indicating a comparably sized testing effect for semantic and phonemic retrieval.

Discussion

In Experiment 1, we found that retrieval practice produced better final free recall performance than both restudy conditions, demonstrating a testing effect. First, this is consistent with earlier research that found robust testing effects after a short retention interval when initial retrieval success was high and/or feedback was provided (both attributes of the present experiment; Rowland & DeLosh, 2015). Second, this result is critical because the ultimate goal of the current study is to determine if the testing effect differs across the retrieval conditions, requiring that we first observe a robust testing effect. Having satisfied that requirement, it is interesting that we found a similar sized testing effect for both the semantic and phonemic groups, implying similar beneficial effects of retrieval for semantic and phonemic cues.

However, this should be considered preliminary at this point, and there are several other aspects of the results to consider. First, as described earlier, these materials were pilot tested in an attempt to equate initial performance between the semantic and phonemic retrieval groups. In Experiment 1, initial retrieval time was similar between the retrieval groups but the semantic group had higher initial recall success. This raises the concern that retrieval with phonemic cues may be more difficult than retrieval with semantic cues. Despite this particular outcome (and as noted in Footnote 3), examining data across all four of the current experiments indicates approximately equal difficulty for both retrieval groups. Thus, generally speaking, differences in retrieval difficulty are not likely to be decisive to the final-recall results (which are quite consistent across the four experiments, as will be seen).

Second, the current results appear to conflict with the results of Veltre et al. (2015), who reported a significant effect of retrieval level on the testing effect. A concern about that study was the use of a single, common (uncued) restudy condition for the comparison with semantic and nonsemantic retrieval conditions, which prompted the present use of two different cued restudy conditions. However, the present results do not demonstrate any clear differences between the semantic and phonemic restudy conditions or between the cued and uncued restudy conditions. This raises questions about the basis of the difference in results between Veltre et al. (2015) and the present experiment. This issue is deferred until additional experiments further evaluate these issues and discussed in detail in the General Discussion.

Third, the lack of a main effect of level requires comment. It might be expected that the semantic group would significantly outperform the phonemic group on the final recall test, exhibiting a type of levels-of-processing effect. Although final recall was numerically greater in the semantic group, the difference was not significant. This may raise concerns about the level manipulation. For example, perhaps in the cued restudy condition, participants ignored the cues while restudying thus diminishing the differences in processing between the two groups. We examined this possibility in two ways. First, one of the postexperiment questions asked participants if they followed our instructions to silently read the cue word first in the cued restudy trials. Generally, participants reported following this instruction, with only three participants in

Table 2
Final Free Recall Proportion Correct: M (SD)

Experiment and review condition	Semantic	Phonemic
Experiment 1		
Restudy	0.24 (0.15)	0.19 (0.11)
Cued restudy	0.28 (0.13)	0.25 (0.16)
Retrieval practice	0.37 (0.13)	0.32 (0.15)
Experiment 2		
Cued restudy	0.20 (0.15)	0.18 (0.15)
Retrieval practice	0.28 (0.13)	0.25 (0.11)
Experiment 3		
Cued restudy	0.19 (0.12)	0.14 (0.11)
Retrieval practice	0.32 (0.14)	0.26 (0.14)
Experiment 4		
Cued restudy	0.11 (0.08)	0.09 (0.10)
Retrieval practice	0.26 (0.12)	0.20 (0.12)

each group claiming that they didn't read the cue before attending to the target. We then removed these participants and reran the 2 (level) × 3 (review condition) ANOVA on final free recall performance. No changes in results were found, indicating that the presence of the testing effect, and its equality between the two groups persisted.

We can also assess this issue by looking at the time it took participants to identify the target item during the cued restudy and restudy trials. In the former, participants were instructed to silently read the cue word before identifying the target aloud. If participants followed this directive, naming times for those trials should be longer than in the restudy group, in which no cue was presented. To assess, we conducted a 2 (level: semantic vs. phonemic) × 2 (review condition: cued restudy vs. restudy) ANOVA on median naming times during the restudy trials. The main effect of review condition was significant, $F(1, 40) = 74.945$, $MS_e = 26{,}503.463$, $p < .001$, $\eta_p^2 = .652$, with slower naming times in the cued restudy condition ($M = 1216.21$, $SD = 303.72$) than the restudy condition ($M = 908.67$, $SD = 148.19$). Therefore, it appears that participants did process the cues in the cued restudy condition.

Another possibility is that levels was manipulated between-subjects, which is somewhat unusual (it is usually manipulated within-subjects, e.g., Craik & Tulving, 1975; Fisher & Craik, 1977; Moscovitch & Craik, 1976). In Experiment 2, levels was manipulated within-subjects and included just the retrieval and cued restudy conditions. This experiment serves a second important purpose. In a between-subjects design, it is possible that the groups encoded the words differently upon their initial presentation in anticipation of either semantic or phonemic tasks (cf. Cho & Neely, 2017). In Experiment 2, each participant experienced both the semantic and phonemic tasks and did not know upon initial presentation which type of task would follow, precluding any concerns about differential initial encoding across conditions.

Experiment 2

Method

Participants. Thirty-two participants (21 females; age $M = 19.13$, $SD = 0.75$) from UNC at Chapel Hill were recruited in in exchange for course credit.

Design and materials. The experiment used a 2 (level: semantic vs. phonemic) × 2 (review condition: retrieval vs. cued restudy) within-subjects design. Because there was no significant difference between the cued restudy and restudy conditions in Experiment 1, the restudy condition was eliminated.

Experiment 2 required 56 target words along with 28 nontargets to constitute the study lists (most of the words were used in Experiment 1 with additional words drawn from Nelson et al., 2004). The total of 84 words were divided into 28 short lists of three words each (two targets and one nontarget). Four lists were used for practice and six lists were used for each review condition (semantic retrieval, semantic restudy, phonemic retrieval, and phonemic restudy). As in Experiment 1, semantic and phonemic cues were selected for each target word.

The average length, frequency, and concreteness for target words was $M = 4.21$ ($SD = 0.82$, *Range* = 3 to 6), $M = 191.33$ ($SD = 358.09$, *Range* = 3 to 1,772), and $M = 5.08$ ($SD = 1.16$, *Range* = 2.56 to 6.92), respectively. The average length, frequency and concreteness for semantic cues was $M = 4.90$ ($SD = 1.17$, *Range* = 3 to 10), $M = 80.58$ ($SD = 319.99$, *Range* = 1 to 2216), and $M = 4.98$ ($SD = 1.12$, *Range* = 2.23 to 6.94), respectively. The average length, frequency and concreteness for phonemic cues was $M = 4.21$ ($SD = 0.80$, *Range* = 3 to 6), $M = 67.83$ ($SD = 126.62$, *Range* = 2 to 794), and $M = 4.59$ ($SD = 1.33$, *Range* = 1.49 to 6.96), respectively. Semantic and phonemic cues did not differ in frequency or concreteness ($ps > .1$), but the semantic cues were almost one letter longer on average ($p < .001$). The average forward strength and backward strength of semantic cue–target pairs was $M = 0.66$ ($SD = 0.11$, *Range* = 0.5 to 0.91) and $M = 0.23$ ($SD = 0.25$, *Range* = 0 to 0.87), respectively. The study materials were fully counterbalanced across experimental conditions.

Procedure. The procedure for Experiment 2 was based on Experiment 1 with the following changes. Phase 1 consisted of four types of trials, namely retrieval or restudy with semantic or phonemic cues. The trial types were randomly ordered so that when a short list was initially presented, the participant did not know if it would be followed by retrieval or restudy, or with semantic or phonemic cues. Phase 2 was identical to Experiment 1. At the end of the experiment, participants were asked whether they read the cue first when restudying and ranked final recall performance of the four conditions (results reported in Appendix A).

Results

Initial cued recall. The proportion of target words correctly recalled did not significantly differ between the two retrieval levels, $t(31) = 1.123$, $p = .270$, nor did the median retrieval times, $t(31) = 0.745$, $p = .462$. The percentage of missing retrieval times also did not significantly differ between the semantic ($M = 2.13\%$, $SD = 8.07\%$) and phonemic ($M = 1.15\%$, $SD = 3.75\%$) conditions, $t(31) = 0.608$, $p = .547$.

Final free recall. Final recall of the targets was submitted to a 2 (level: semantic vs. phonemic) × 2 (review condition: retrieval vs. cued restudy) repeated measures ANOVA. The same pattern of results was obtained as in Experiment 1. Only the main effect of review condition was significant, $F(1, 31) = 8.534$, $MS_e = 0.020$, $p = .006$, $\eta_p^2 = .216$, with retrieval practice leading to greater final recall performance than cued restudy. Neither the main effect of level, $F(1, 31) = 0.777$, $p = .385$, nor the interaction between level and review condition were significant, $F(1, 31) = 0.207$, $p = .653$.

Discussion

In Experiment 2, initial recall and retrieval time was approximately equal between the semantic and phonemic conditions. In addition, we again observed a robust testing effect on final free recall, and this effect was not moderated by the level of retrieval, thus replicating the results of Experiment 1. Further, the equality at initial retrieval supports the idea that the final recall results are not due to any substantial differences in retrieval difficulty between the semantic and phonemic conditions.

A similar testing effect for semantic and phonemic retrieval was found when manipulating level between-subjects (Experiment 1) and within-subjects (Experiment 2). One concern with Experiment 1 is that knowing what type of retrieval is required may induce different encoding during initial presentation. This in turn might have modified the results. In the present experiment, this is not possible as all participants experience both semantic and phonemic cues, and do not know which will be relevant during the initial presentation on any given trial. This indicates that the equality of the testing effect across levels does not depend on the between- or within-subject manipulation of this variable.

Finally, it should be noted that the main effect of levels did not significantly affect final recall in the present experiment, when levels was manipulated within subjects, nor in Experiment 1 when manipulated between subjects. The semantic condition produced numerically higher performance on final recall as it did in Experiment 1, producing four of four comparisons with a numerical advantage for the semantic compared to phonemic condition (the cued recall and cued restudy conditions of both Experiments 1 and 2), but the effect was not significant in either experiment. This issue is further explored in Experiment 3.

Experiment 3

Although the results of Experiments 1 and 2 were consistent in finding a similar effect of semantic and phonemic retrieval on the testing effect, it is reasonable to wonder if this equality only holds when the levels manipulation fails to produce a clear effect on final recall. The current levels manipulation differs from traditional levels-of-processing in two ways. First, the traditional manipulation guides processing of the stimulus during its initial presentation whereas our manipulation is introduced during the second experience with the stimulus. Second, the traditional levels-of-processing manipulation requires an overt processing task, typically a type of judgment task, which was not used in our manipulation. Given that we are interested in the effect of level during retrieval or restudy, it is necessary that our manipulation occurs when reexperiencing the stimulus. However, the second characteristic—the lack of an overt judgment task—is not a necessary characteristic, and its introduction may well induce a robust effect of levels in the context of our manipulation. Experiment 3 again examined the effect of level of retrieval on the testing effect but introduced an overt judgment into the manipulation.

Method

Participants. Thirty-two participants (19 females; age $M = 19.84$, $SD = 2.57$) from UNC at Chapel Hill were recruited in in exchange for course credit.

Design, materials, and procedure. Experiment 3 was identical to Experiment 2 except for the following modifications. First, the retrieval trials began as in Experiment 2, with the cue presented for 4 s during which time the participant tried to retrieve the appropriate target. After this, the target word joined the cue on the screen (for feedback), accompanied by a question presented below the words for 4 s. Participants were asked to read the word to verify that they recalled the correct target, saying it aloud if they had not, before answering the question. For the semantic retrieval trials, participant judged which word (the cue or target) was more pleasant; for the phonemic retrieval trials, they judged which word had more consonants. Second, cued restudy trials began with the cue—target pair for 4 s, then the question was presented below the pair for an additional 4 s. For semantic restudy, they made the pleasantness judgment and for phonemic restudy, they made the consonant judgment. For both judgments, participants pressed “j” to choose the left word (cue); “k” if the answer was “equal”; and “l” to choose the right word (target).

Results

Initial cued recall. The proportion of target words correctly recalled did not significantly differ between the two retrieval levels, $t(31) = 0.753$, $p = .457$. However, median retrieval time for the semantic condition was significantly slower than the phonemic condition, $t(31) = 2.888$, $p = .007$, $d = 0.511$. The percentage of missing retrieval times did not significantly differ between the semantic ($M = 6.03\%$, $SD = 10.83\%$) and phonemic ($M = 5.02\%$, $SD = 10.07\%$) conditions, $t(31) = 0.644$, $p = .524$.

Final free recall. A 2 (level) $\times$ 2 (review condition) repeated measures ANOVA revealed a significant main effect of review condition, $F(1, 31) = 30.368$, $MS_e = 0.016$, $p < .001$, $\eta_p^2 = .495$ (i.e., a testing effect). Additionally, the main effect of level was also significant, $F(1, 31) = 6.857$, $MS_e = 0.015$, $p = .014$, $\eta_p^2 = .181$, with the semantic condition producing better performance than the phonemic condition. However, the interaction between level and review condition was nonsignificant, $F(1, 31) < 0.001$, $p > .999$.

Discussion

The earlier experiments showed a trend for an effect of levels but it was nonsignificant in both Experiments 1 and 2. Experiment 3 introduced an overt judgment as part of the manipulation, similar to the traditional levels-of-processing manipulation, and the main effect of levels was now significant with the semantic condition producing greater final recall than the nonsemantic condition. Critically, the rest of the results replicate the prior experiments. First, the retrieval condition produced greater final recall than the cued restudy condition, replicating the testing effect. Second, and more importantly, the size of the testing effect was once again comparable (indeed, nearly identical) in the semantic and nonsemantic conditions. This demonstrates that when the levels manipulation robustly affects final recall, the pattern of the earlier experiments persists. Specifically, the presence of a clear levels effect does not induce a larger testing effect in the semantic condition.

Experiment 4

The results of Experiments 1–3 imply that level of retrieval does not moderate the testing effect, at least under the present experimental circumstances. However, there is one final issue to consider, the extent to which participants actually used the semantic and phonemic cues to retrieve the studied words during retrieval practice. Clearly, participants used the provided cues to at least some degree because the appropriate target was reported in the presence of the cue. However, it is at least possible that despite the intent of the experimental procedure, participants engaged in free recall of the three studied words during retrieval practice and then chose the one which was related to the semantic or phonemic cue. That is, the retrieval practice could be free recall rather than cued recall (with different types of cues). If so, participants would not have engaged in retrieval at different levels (with different cues), at least for those trials in which this strategy was used.[5]

In Experiment 4, participants initially studied six words and then retrieved or restudied two of them on each trial. Increasing the list length was designed to increase the difficulty of free recall, making such a strategy less likely, and to promote the likelihood that the cues would be used to initiate recall of the studied target words. Otherwise, we reverted to the methods of Experiment 2.

Method

Participants. Thirty-two participants (21 females; age M = 18.34, SD = 1.04) from UNC at Chapel Hill were recruited in in exchange for course credit.

Design, materials, and procedure. Experiment 4 was identical to Experiment 2 with the exception that participants initially studied six words per list (two targets and four nontargets). A total of 84 new words were drawn from Nelson et al. (2004) and three were randomly assigned to act as additional nontargets in each list. This new set of words were similar to the words used in the prior experiments with average length, frequency, and concreteness of M = 5.06 (SD = 1.38; *Range* = 3 to 8), M = 107.43 (SD = 176.29; *Range* = 6 to 1,016), and M = 5.00 (SD = 1.38; *Range* = 2.20 to 7.00), respectively. All targets, semantic cues, and phonemic cues were the same as in Experiment 2.

Results

Initial cued recall. Although there was no significant difference in median retrieval time, $t(31) = 1.465$, $p = .153$, more target words were recalled in the semantic than phonemic condition, $t(31) = 2.314$, $p = .027$, $d = 0.409$. The percentage of missing retrieval times did not significantly differ between the semantic (M = 3.00%, SD = 6.81%) and phonemic (M = 7.13%, SD = 18.13%) conditions, $t(31) = 1.532$, $p = .136$.

Final free recall. A 2 (level) × 2 (review condition) repeated measures ANOVA revealed a significant main effect of review condition, $F(1, 31) = 49.000$, $MS_e = 0.011$, $p < .001$, $\eta_p^2 = .613$, demonstrating greater recall in the retrieval than restudy condition (i.e., a testing effect), and a significant main effect of level, $F(1, 31) = 4.794$, $MS_e = 0.010$, $p = .036$, $\eta_p^2 = .134$, demonstrating greater recall in the semantic than phonemic level. The interaction between level and review condition was nonsignificant, $F(1, 31) = 1.305$, $p = .262$.

Discussion

Experiments 1–3 demonstrated that level of retrieval does not influence the size of the testing effect. One factor that might have contributed to those results was whether participants engaged in free recall rather than cued recall on some initial retrieval trials. This would diminish the extent to which they engaged in different levels of retrieval and possibly lead to similar sized testing effects in both conditions. Experiment 4 was designed to increase the difficulty of free recall to ensure that participants engaged in cued recall during retrieval practice by increasing the list length. Importantly, the same pattern of results as in the prior experiments was obtained—a testing effect that was not moderated by the level of retrieval. Thus, it seems unlikely that a free recall strategy during retrieval practice was obscuring an effect of retrieval level on the size of the testing effect.

Bayesian Analyses and a Small-Scale Meta-Analysis

The present study was conducted to examine whether different levels of retrieval moderate the size of the testing effect. Based on the results of the four experiments, the answer seems to be no. However, the primary results are null interactions. In order to further explore the reliability of this result, we first conducted Bayesian analyses. We computed Bayes Factors using the statistical software program JASP (JASP Team, 2019; jasp-stats.org) for all four experiments and interpretations were based on Wagenmakers et al. (2018). The Bayes Factor (BF_{10}) is a measure of the fit of the data under one model (e.g., the alternative hypothesis/model) relative to the fit under a second model (e.g., the null model/hypothesis). Larger BF_{10} values reflect more support for the first (e.g., alternative) model versus the second (e.g., null) model. Its inverse, $BF_{01} = 1/BF_{10}$, has a similar interpretation, but now indicates the strength of the evidence for the second (e.g., null) model versus the first (e.g., alternative) model.

Below, we use this measure to assess the evidence in favor of the main effects only model compared with the main effects plus interaction model (i.e., support for the inclusion of the interaction term over and above the main effects, see Appendix C, Tables C1–C4). For all comparisons, we used the default prior settings in JASP such that the fixed effect scale factor (r_A) = 0.5 (e.g., Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Wagenmakers et al., 2018). Although we focus on the importance of the interaction term, the evidence in favor of all possible models was also assessed per the recommendations of Wagenmakers et al. (2018) and is presented in Appendix C (Tables C5–C8). To preview, the Bayesian analyses produced a very similar pattern of results to the ANOVA analyses (i.e., no evidence of an interaction between level and review condition).

For Experiment 1, the data are 2.99 times more likely under the main effects only model than under the main effects plus interaction model ($BF_{01} = 2.99$). For Experiment 2, the data are 3.80 times more likely under the main effects only model than under the main effects plus interaction model ($BF_{01} = 3.80$). For Experiment 3, the data are 3.92 times more likely under the main effects only model than under the main effects plus interaction model ($BF_{01} = 3.92$). For Experiment 4, the data are 2.18 times more likely under the main effects only model than under the main effects plus interaction model (BF_{01} = 2.18). Thus, these analyses indicate that the data provide support for the main effects only model over the main effects plus interaction model.[6]

[5] It is possible that postretrieval processes could still differ between the levels of retrieval because processing the retrieved words to see which is related to a semantic cue might differ from processing the retrieved words to see which is related to a phonemic cue. But any such differential (and critically, non-retrieval-based) processing would presumably be similar to the differential processing occurring across the two levels of restudy.

[6] According to the descriptive classification scheme (e.g., Wagenmakers et al., 2018), the evidence against including the interaction term ranged from anecdotal (BF_{01} = 1 to 3) to moderate (BF_{01} = 3 to 10) across experiments (i.e., BF_{01} = 2.18 to 3.92). However, these discrete labels are only approximations of different standards of evidence and the specific Bayes factor value can fluctuate across categories due to error from the numerical integration routine (i.e., Markov chain Monte Carlo [MCMC]), which ranged from 2.69% to 4.04%.

Next, to increase the power of assessing the critical interaction, we conducted a small-scale meta-analysis using the single paper meta-analysis (SPM) proposed by McShane and Böckenholt (2017). The current meta-analysis was based on a 2 (level: semantic vs. phonemic) × 2 (review condition: retrieval vs. cued restudy) design (Tables 3 and 4). In Experiment 1, level was manipulated between subjects and review condition was manipulated within subjects. In Experiments 2–4, both variables were manipulated within subjects. The following contrasts were used to examine the main effect of level, the main effect of review condition, and the interaction effect (1 1 −1 −1), (1 −1 1 −1) and (1 −1 −1 1), respectively.

The results (see Figure 2) revealed a significant main effect of level (SPM estimate = 0.082, *SE* = 0.024, 95% CI [0.036, 0.128]), demonstrating an advantage of semantic over phonemic processing; a significant main effect of review condition (SPM estimate = 0.226, *SE* = 0.022, 95% CI [0.183, 0.269]), demonstrating an advantage of retrieval practice over cued restudy (i.e., a testing effect); and a nonsignificant interaction effect (SPM estimate = 0.012, *SE* = 0.018, 95% CI [−0.024, 0.048]), suggesting that level of initial retrieval did not moderate the testing effect.

For all analyses reported thus far, the final free recall data in the retrieval condition was not conditionalized on initial recall success. This is necessary to avoid the introduction of item-selection confounds into the assessment of the testing effect; items recalled during initial retrieval practice might simply be easier items in general and conditionalizing final recall on initial retrieval success might then produce a spurious advantage for the retrieval condition. However, to be sure that our critical results hold when we restrict consideration to those items successfully recalled during retrieval practice, we reanalyzed the final free recall data conditionalized on successful initial retrieval (note this only affects recall scores in the retrieval conditions and not in the restudy conditions). The critical results were all unchanged. For all experiments, the size of the testing effect did not significantly differ across the semantic and phonemic conditions. Likewise, Bayesian analyses of the conditionalized data produced nearly identical results with the unconditionalized data: the level-by-review-condition interaction favored the null hypothesis in all cases (BF_{01} of 2.64, 3.67, 3.54, and 3.78, for Experiments 1–4, respectively). Finally, when the meta-analysis was performed on the conditionalized results, the effects were identical: The main effects of testing and review condition were both significant and the interaction was not (SPM estimate = 0.007, *SE* = 0.021, 95% CI [−0.036, 0.049]). Thus, the finding that the testing effect is unmoderated by level of retrieval holds for the conditionalized as well as unconditionalized recall results.

Table 3
Summary Information Used in Single Paper Meta-Analysis (SPM)

Study	Factor 1	Factor 2	*M*	*SD*	*n*	wi
1	Semantic	Retrieval practice	0.3730	0.1334	21	1
1	Semantic	Cued restudy	0.2778	0.1326	21	1
1	Phonemic	Retrieval practice	0.3214	0.1474	21	2
1	Phonemic	Cued restudy	0.2540	0.1592	21	2
2	Semantic	Retrieval practice	0.2786	0.1263	32	3
2	Semantic	Cued restudy	0.1979	0.1537	32	3
2	Phonemic	Retrieval practice	0.2474	0.1130	32	3
2	Phonemic	Cued restudy	0.1823	0.1533	32	3
3	Semantic	Retrieval practice	0.3151	0.1400	32	4
3	Semantic	Cued restudy	0.1927	0.1242	32	4
3	Phonemic	Retrieval practice	0.2578	0.1441	32	4
3	Phonemic	Cued restudy	0.1354	0.1053	32	4
4	Semantic	Retrieval practice	0.2552	0.1196	32	5
4	Semantic	Cued restudy	0.1068	0.0770	32	5
4	Phonemic	Retrieval practice	0.1953	0.1171	32	5
4	Phonemic	Cued restudy	0.0885	0.0969	32	5

In sum, the Bayesian analyses and meta-analyses support the conclusion that the testing effect is comparable for the semantic and phonemic conditions. This is especially important for the present experiments for two reasons. First, the original power calculations ensured that we had robust power to detect a testing effect, which is a prerequisite for detecting any interaction but may raise question about the power to detect the interaction itself. Second, the effect of level of processing did not have a large effect. Performing Bayesian analyses helps further assess the diagnostic value of the null interactions, and the meta-analyses allows a more powerful assessment of this interaction.

General Discussion

Retrieval does not just reveal the contents of memory but can also modify memory representations, demonstrating what may be called the encoding (or reencoding) effects of retrieval. It is important to delineate the nature of the encoding effects of retrieval especially in comparison to traditional encoding processes (Buchin & Mulligan, 2017; Mulligan & Picklesimer, 2016). One important characteristic of encoding is captured by the well-documented levels-of-processing effect (Craik & Tulving, 1975) whereby deep or semantic processing produces better memory than shallow or nonsemantic processing. The current study examined whether a similar phenomenon applies to the subsequent mnemonic effects of retrieval by examining the effects of the level (or depth) of retrieval on the testing effect.

Experiment 1 manipulated levels of retrieval (and restudy) between subjects while Experiment 2 manipulated levels within subjects. Both experiments demonstrated the same pattern of results. First, the final test revealed a robust testing effect, and second, the testing effect was not moderated by level of retrieval. However, despite trends in the expected direction, the main effect of levels was not significant. Experiment 3 sought to enhance the levels effect by adding an overt levels judgment, and Experiment 4 sought to rule out an alternative account of the equality of the testing effects by increasing the list length. A testing effect and an effect of levels were found in both experiments, but levels did not moderate the testing effect. Finally, a small-scale meta-analysis provides an even more powerful analysis of the results, demonstrating an overall effect of levels and review condition, but no interaction—thus, there is no evidence of differential testing effects for semantic and phonemic retrieval, a result likewise supported by the Bayesian analyses.

Based on traditional research on encoding effects in memory and traditional theories of memory encoding (e.g., the levels-of-processing framework), one might expect a levels-of-retrieval effect, in which semantic retrieval enhances memory to a greater extent than nonsemantic retrieval (at least when the final test is free recall or other conceptually driven memory tests). Our results provide preliminary evidence of an asymmetry in the effects of levels on the mnemonic effects of encoding and retrieval. Specifically, in contrast to the

Table 4
Covariance Information Used in Single Paper Meta-Analysis (SPM)

Study	Factor 1	Factor 2	Study	Factor 1	Factor 2	Covariance
1	Semantic	Retrieval practice	1	Semantic	Cued restudy	0.003
1	Phonemic	Retrieval practice	1	Phonemic	Cued restudy	−0.006
2	Semantic	Retrieval practice	2	Semantic	Cued restudy	0.006
2	Semantic	Retrieval practice	2	Phonemic	Retrieval practice	0.003
2	Semantic	Retrieval practice	2	Phonemic	Cued restudy	−0.004
2	Semantic	Cued restudy	2	Phonemic	Retrieval practice	−0.001
2	Semantic	Cued restudy	2	Phonemic	Cued restudy	0.003
2	Phonemic	Retrieval practice	2	Phonemic	Cued restudy	0.002
3	Semantic	Retrieval practice	3	Semantic	Cued restudy	0.001
3	Semantic	Retrieval practice	3	Phonemic	Retrieval practice	0.007
3	Semantic	Retrieval practice	3	Phonemic	Cued restudy	0.004
3	Semantic	Cued restudy	3	Phonemic	Retrieval practice	−0.002
3	Semantic	Cued restudy	3	Phonemic	Cued restudy	0.001
3	Phonemic	Retrieval practice	3	Phonemic	Cued restudy	0.007
4	Semantic	Retrieval practice	4	Semantic	Cued restudy	0.000
4	Semantic	Retrieval practice	4	Phonemic	Retrieval practice	0.001
4	Semantic	Retrieval practice	4	Phonemic	Cued restudy	0.001
4	Semantic	Cued restudy	4	Phonemic	Retrieval practice	0.000
4	Semantic	Cued restudy	4	Phonemic	Cued restudy	0.000
4	Phonemic	Retrieval practice	4	Phonemic	Cued restudy	0.001

typical effect of levels on memory encoding, the effect of retrieval on later memory, as isolated by the testing effect, did not vary across the semantic and phonemic conditions. Significantly, these results were obtained in the face of robust testing effects. The pattern was found regardless of whether levels was manipulated between subjects (Experiment 1) or within subjects (Experiments 2–4), and when the main effect of levels was obtained (Experiments 3–4) or not (Experiments 1–2).

Earlier research on the mnemonic effects of retrieval (predating the current interest in the testing effect) produced mixed results but is limited by two methodological concerns (e.g., Bartlett, 1977; Bartlett & Tulving, 1974; McDaniel et al., 1989; McDaniel & Masson, 1985; Whitten, 1978). First, the mnemonic effects of retrieval were assessed by directly comparing the semantic and nonsemantic retrieval conditions, confounding the mnemonic effects of retrieval with the effects of reexperience. Thus, any difference in the retrieval conditions could be due to the mnemonic effect of retrieval, per se, or the effect of additional processing in the company of a semantic versus nonsemantic cue. Isolating the effects of retrieval on subsequent memory requires comparison to an appropriate restudy condition. The importance of this issue can be observed in the current final recall results (see Table 2). Examining only the retrieval conditions shows that semantic retrieval generally produced higher final recall than nonsemantic retrieval—indeed the numerical averages are higher in four of four experiments. This could well be taken as evidence that semantic retrieval enhances memory more than nonsemantic retrieval. Likewise, if the two retrieval conditions were compared with a common baseline (restudy) condition, the same conclusion would seem warranted (Veltre et al., 2015). However, the cued restudy conditions indicate that likewise, semantic restudy generally produced higher final recall than nonsemantic restudy—the numerical averages are again higher in four of four experiments. The appropriate comparison instead compares the testing effects (retrieval—restudy) for the semantic and nonsemantic conditions, each measured relative to its appropriate restudy condition. This comparison results in little difference—and the conclusion that under the present circumstances, the level of retrieval has little effect on the mnemonic benefits of retrieval.

The second concern about the bulk of the earlier studies is that initial retrieval was often greater in the semantic than nonsemantic condition, potentially confounding retrieval difficulty (or effort) with type of retrieval (cf. Veltre et al., 2015). Via pilot testing, we tried to equate retrieval difficulty and generally succeeded. That is, the two indicators of initial retrieval difficulty (accuracy and retrieval speed) provide little evidence of differential retrieval difficulty across the four experiments. In two experiments, initial retrieval accuracy was greater in the semantic condition but in the other two, accuracy did not differ (and numerically favored the nonsemantic condition). With regard to retrieval speed, the nonsemantic condition produced numerically quicker retrieval in all four experiments, significantly so in one of the experiments. Assessing across experiments indicates that retrieval difficulty did not systematically differ across the semantic and phonemic conditions.[7]

The Veltre et al. (2015) study requires additional discussion. As reviewed in the introduction, this experiment instituted retrieval practice with semantic or nonsemantic cues along with a single (uncued) restudy condition, and found a significant effect of type of retrieval on a final semantic-cued-recall test (another condition using a nonsemantic final test is not relevant for present purposes). These results imply an effect of retrieval level on the testing effect and thus conflict with the current results. Why the discrepant results? There are two salient differences between the Veltre et al. (2015) study and the current experiments. First, as noted in the introduction, a common restudy condition may be problematic if reexperience in the presence of

[7] Although it should be noted that even though initial retrieval success and difficulty have been shown to influence learning in the absence of feedback (e.g., Pyc & Rawson, 2009), recent studies have found a much smaller effect (if any) when feedback is provided during retrieval practice (e.g., Kornell, Klein, & Rawson, 2015; Vaughn & Kornell, 2019), as was the case in the current study and in Veltre et al. (2015).

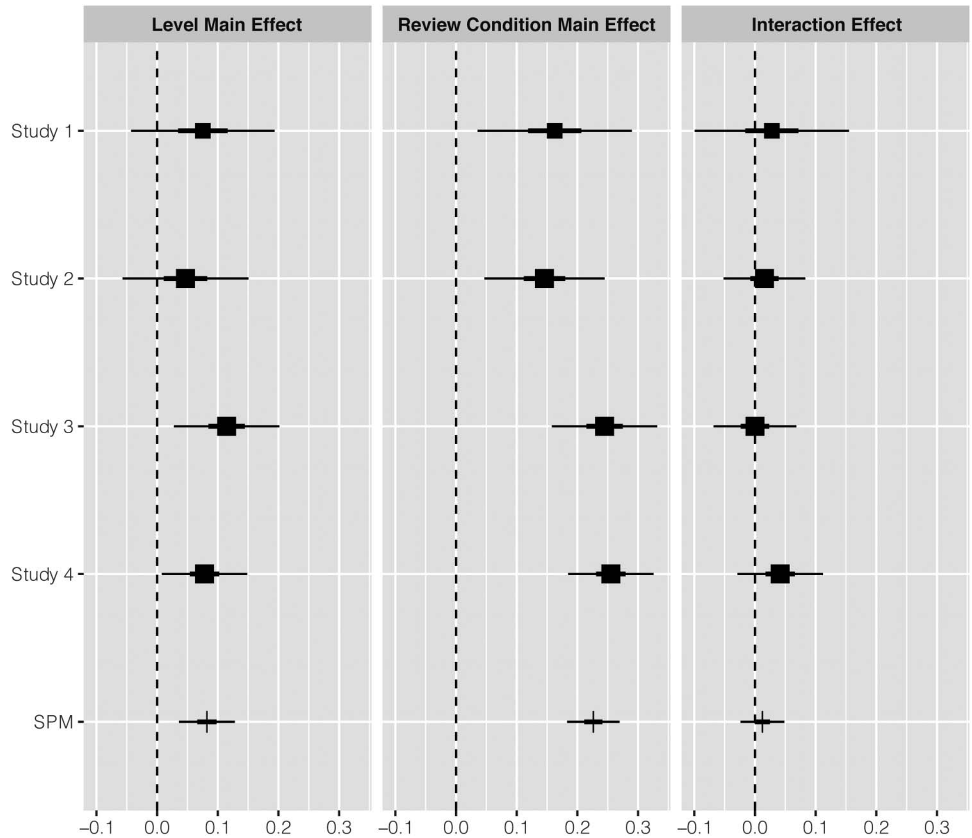

Figure 2. Results of single paper meta-analysis (SPM). Effect estimates for single experiment (study) and SPM are given by the squares; 50% and 95% intervals are given by the thick and thin lines, respectively (McShane & Böckenholt, 2017).

different cues influences later recall. Restudy with the same cues used in the retrieval condition seems a better way to equate the effects of reexperience. Second, the final test in Veltre et al. (2015) was semantic-cued recall whereas the final test in the present experiments was free recall. With regard to the first difference, the bulk of our results indicate that the nature of the restudy condition is important—as described above, overall there appears to be a difference in restudy with a semantic versus a nonsemantic cue, indicating that the testing effect needs to be assessed relative to a matched restudy condition. However, one aspect of the results of Experiment 1 seems inconsistent with this possibility—the cued and uncued restudy conditions in that experiment did not significantly differ. One would expect at least some difference, perhaps with regard to the semantic restudy condition producing higher final recall than the uncued restudy condition. This did not occur in Experiment 1, raising a question as to whether the use of multiple restudy conditions is the critical issue. We argue that the use of matched restudy conditions is still most appropriate going forward as it eliminates at least a potential confound, and as noted, the overall results indicate a difference between the semantic and phonemic restudy conditions. That this did not occur in Experiment 1 may be due to the weaker effect of levels in that experiment, an effect that was more robust in later experiments.

The other difference is the final test. It is possible that the semantic cued-recall test used by Veltre et al. (2015) is more sensitive to a potential levels-of-retrieval effect than is a free recall test as used in the present experiments, perhaps because cued recall more completely evokes differences in prior retrieval processes during retrieval practice. However, it is also possible that semantic cued recall is more sensitive to differences in prior restudy conditions, specifically differences in semantic versus nonsemantic restudy conditions. Had multiple restudy conditions been implemented in Veltre et al. (2015), it is possible that differences between them would be more easily detected with cued recall than free recall (again, because cued recall, especially with identical cues, might more precisely reinstate the prior restudy experience). Adjudicating these possibilities is an important issue for subsequent research.

In the introduction, we discussed what theories of the testing effect might lead us to expect about the effect of levels of retrieval. The elaborative retrieval account (Carpenter, 2009, 2011; Carpenter & Yeung, 2017) suggests that semantic retrieval should be more potent than nonsemantic retrieval. According to this account, retrieval activates information semantically related to the cue, which in turn increases the number of retrieval routes to the targets (Pyc & Rawson, 2010, propose a similar, mediator-effectiveness hypothesis; see Rowland, 2014, for discussion). Retrieval with semantic cues should provide greater opportunities for semantic elaboration than retrieval from nonsemantic cues which suggests that a deep retrieval condition should enhance later memory more than a shallow retrieval condition. For similar reasons, traditional theories of memory encoding (e.g., the levels-of-processing account) predict a levels-of-retrieval effect. To further align the expectations from the elaborative retrieval account and traditional theories of memory encoding, it should be noted that semantic elaboration at retrieval is often equated with the elaborative processes that operate during encoding (Han, O'Connor, Eslick, & Dobbins, 2012; Raposo, Han, & Dobbins, 2009; Wing, Marsh, & Cabeza, 2013; see Lehman & Karpicke, 2016). However, there was no evidence of a larger testing effect in the semantic condition, contrary to these expectations.

Despite the lack of an effect of levels on the testing effect, the semantic retrieval condition actually does produce an advantage in later memory relative to the nonsemantic retrieval condition but this advantage is not actually a result of retrieval, per se. Rather, it is a result of related processing of the stimulus and cue that occurs as or after the item is retrieved (or when the target is presented as feedback)—related processing that is matched by the processing occurring in the appropriate restudy condition. As reviewed earlier, the comparison of the semantic and nonsemantic retrieval conditions (rather than the comparison of the testing effects across conditions) points in this direction, as do the results of Veltre et al. (2015). This possibility is compatible with other accounts of the testing effect. For example, the retrieval effort hypothesis argues that the testing effect reflects more effortful processing of the target in the retrieval than restudy condition, with the corollary that more effortful retrieval produces a greater testing effect than less effortful retrieval (e.g., Bjork, 1975; Endres & Renkl, 2015; Halamish & Bjork, 2011; Pyc & Rawson, 2009). Given that the current semantic and nonsemantic retrieval conditions were approximately matched on difficulty (that is, retrieval effort), the resulting equality of the testing effect sits comfortably in this view. Likewise, the results are compatible with the episodic context account which proposes that retrieval updates the contextual representation of targets by combining features from both the original study context and the context prevalent during retrieval practice (e.g., Karpicke et al., 2014; Whiffen & Karpicke, 2017). If one assumes that semantic and nonsemantic retrieval both provide the requisite contextual updating, then equivalent testing effects should result.[8]

The present experiments provide evidence that the level of retrieval does not modify the testing effect—that is, that deep retrieval does not enhance memory more than shallow retrieval, in marked contrast to the effects of deep versus shallow encoding tasks. If subsequent research concurs, this would mark an important difference between the encoding consequences of retrieval and other forms of initial stimulus encoding. But it should be noted that this work is preliminary and requires follow up, in several ways. First, we adapted the methods of Whitten (1978) for two important purposes: (a) to produce high initial retrieval and (b) to equate retrieval difficulty for the semantic and nonsemantic conditions. The current methods were largely successful in these two regards. Despite this, the current paradigm is a little unusual in presenting participants with many short lists, each quickly followed by retrieval or restudy trials. This area of research more commonly uses a long list or block of study materials followed by a separate retrieval/restudy phase. It will be important to explore the effects of levels-of-retrieval in the context of this more typical paradigm but the challenges that motivated the current experimental design may reemerge. In particular, long study lists followed by a separate retrieval phase commonly produce much lower levels of initial retrieval success, impairing our ability to observe a robust testing effect. Second, in designs using long study lists, nonsemantic cues often produce quite low performance compared to semantic cues (Morris et al., 1977; Mulligan & Picklesimer, 2012). See Veltre et al. (2015) for procedures to alleviate some of these concerns.

Second, the current experiments used a brief retention interval (of 5 min) prior to the final recall test. This is certainly adequate for the present goal of exploring whether retrieval-based encoding produces effects analogous to the levels-of-processing effect, given that the vast majority of experiments on levels-of-processing used similar, brief retention intervals. However, because testing effects may increase with longer (e.g., multiday) retention intervals, it is important to determine if the approximate equality of testing effects for semantic and nonsemantic retrieval persists. Third, to eliminate the potential for a reexperience confound between the restudy and retrieval conditions, the present experiments used feedback during retrieval practice, another detail requiring evaluation in subsequent research.

Finally, if effects of levels-of-retrieval are ultimately documented, it will become important to determine whether these effects transfer to other types of final tests. Despite the well-known advantage of semantic encoding on later memory, it is also widely known that the effect has limitations, occurring for conceptually driven tests but reversed on perceptually driven memory tests (e.g., Morris et al., 1977; Roediger, 1990). Indeed, this was a central issue in Veltre et al.'s (2015) study. The present experiments find no evidence of a levels-of-retrieval effect when the final test is free recall, a commonly used conceptual memory test. However, it is unknown if a levels-of-retrieval effect (possibly favoring a nonsemantic retrieval condition) would occur on a perceptually driven final test. The results of Veltre et al. (2015) suggest not, but this issue may well require additional research.

[8] However, if evidence emerges that semantic and non-semantic retrieval do not produce equivalent contextual updating, then the implications of the present results for this account will have to be revisited. Currently, to our knowledge, there is no clear evidence on this point.

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,* 659–701. http://dx.doi.org/10.3102/0034654316689306

Bartlett, J. C. (1977). Effects of immediate testing on delayed retrieval: Search and recovery operations with four types of cue. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 719–732. http://dx.doi.org/10.1037/0278-7393.3.6.719

Bartlett, J. C., & Tulving, E. (1974). Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval. *Journal of Memory and Language, 13,* 297–309.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Buchin, Z. L., & Mulligan, N. W. (2017). The testing effect under divided attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1934–1947. http://dx.doi.org/10.1037/xlm0000427

Buchin, Z. L., & Mulligan, N. W. (2019). Divided attention and the encoding effects of retrieval. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 72,* 2474–2494. http://dx.doi.org/10.1177/1747021819847141

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

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *The European Journal of Cognitive Psychology, 19,* 514–527. http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569. http://dx.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,* 1547–1552. http://dx.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,* 268–276. http://dx.doi.org/10.3758/BF03193405

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826–830. http://dx.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. http://dx.doi.org/10.1016/j.jml.2016.06.008

Cermak, L. S., & Craik, F. I. (1979). *Levels of processing in human memory*. Hillsdale, NJ: Erlbaum.

Chan, J. C., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. http://dx.doi.org/10.1037/0096-3445.135.4.553

Cho, K. W., & Neely, J. H. (2017). The roles of encoding strategies and retrieval practice in test-expectancy effects. *Memory, 25,* 626–635. http://dx.doi.org/10.1080/09658211.2016.1202983

Craik, F. I. (2002). Levels of processing: Past, present. and future? *Memory, 10,* 305–318. http://dx.doi.org/10.1080/09658210244000135

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125,* 159–180. http://dx.doi.org/10.1037/0096-3445.125.2.159

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Memory and Language, 11,* 671–684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*. Advance online publication. http://dx.doi.org/10.3389/fpsyg.2015.01054

Fisher, R. P., & Craik, F. I. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 701–711. http://dx.doi.org/10.1037/0278-7393.3.6.701

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812. http://dx.doi.org/10.1037/a0023219

Han, S., O'Connor, A. R., Eslick, A. N., & Dobbins, I. G. (2012). The role of left ventrolateral prefrontal cortex during episodic decisions: Semantic elaboration or resolution of episodic interference? *Journal of Cognitive Neuroscience, 24,* 223–234. http://dx.doi.org/10.1162/jocn_a_00133

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12,* 852–857. http://dx.doi.org/10.3758/BF03196776

JASP Team. (2019). JASP (Version 0.11.1) [Computer software]. Retrieved from https://jasp-stats.org/

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

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (2nd ed., pp. 487–514). Cambridge, MA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61,* 237–284. http://dx.doi.org/10.1016/B978-0-12-800283-4.00007-1

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. http://dx.doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 283–294. http://dx.doi.org/10.1037/a0037850

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology, 109,* 451–464. http://dx.doi.org/10.2307/1423016

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1573–1591. http://dx.doi.org/10.1037/xlm0000267

Lockhart, R. S. (1975). The facilitation of recognition by recall. *Journal of Memory and Language, 14,* 253–258.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 671–685. http://dx.doi.org/10.1037/a0018785

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *The European Journal of Cognitive Psychology, 19*(4–5), 494–513. http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition, 17,* 423–434. http://dx.doi.org/10.3758/BF03202614

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. http://dx.doi.org/10.1037/0278-7393.11.2.371

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *The Jour-*

nal of Consumer Research, 43, 1048–1063. http://dx.doi.org/10.1093/jcr/ucw085

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 609–622. http://dx.doi.org/10.1037/0278-7393.2.5.609

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Memory and Language, 16,* 519–533.

Moscovitch, M., & Craik, F. I. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Memory and Language, 15,* 447–458.

Mulligan, N. W. (2008). Attention and memory. *Learning and Memory: A Comprehensive Reference, 2,* 7–22.

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 859–871. http://dx.doi.org/10.1037/xlm0000056

Mulligan, N. W., & Picklesimer, M. (2012). Levels of processing and the cue-dependent nature of recollection. *Journal of Memory and Language, 66,* 79–92. http://dx.doi.org/10.1016/j.jml.2011.10.001

Mulligan, N. W., & Picklesimer, M. (2016). Attention and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 938–950. http://dx.doi.org/10.1037/xlm0000227

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407. http://dx.doi.org/10.3758/BF03195588

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,* 437–447. http://dx.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,* 335. http://dx.doi.org/10.1126/science.1191465

Raposo, A., Han, S., & Dobbins, I. G. (2009). Ventrolateral prefrontal cortex and self-initiated semantic elaboration during memory retrieval. *Neuropsychologia, 47,* 2261–2271. http://dx.doi.org/10.1016/j.neuropsychologia.2008.10.024

Roediger, H. L., III. (1990). Implicit memory. Retention without remembering. *American Psychologist, 45,* 1043–1056. http://dx.doi.org/10.1037/0003-066X.45.9.1043

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

Rose, N. S., & Craik, F. I. (2012). A processing approach to the working memory/long-term memory distinction: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1019–1029. http://dx.doi.org/10.1037/a0026976

Rose, N. S., Myerson, J., Roediger, H. L., III, & Hale, S. (2010). Similarities and differences between working memory and long-term memory: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 471–483. http://dx.doi.org/10.1037/a0018405

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22,* 304–321. http://dx.doi.org/10.1037/met0000057

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. http://dx.doi.org/10.1037/a0037559

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory, 23,* 403–419. http://dx.doi.org/10.1080/09658211.2014.889710

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications, 4,* 35. http://dx.doi.org/10.1186/s41235-019-0187-y

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory, 23,* 1229–1237. http://dx.doi.org/10.1080/09658211.2014.970196

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25,* 58–76. http://dx.doi.org/10.3758/s13423-017-1323-7

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1036–1046. http://dx.doi.org/10.1037/xlm0000379

Whitten, W. B. (1978). Initial-retrieval "depth" and the negative recency effect. *Memory & Cognition, 6,* 590–598. http://dx.doi.org/10.3758/BF03198248

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia, 51,* 2360–2370. http://dx.doi.org/10.1016/j.neuropsychologia.2013.04.004

(*Appendices follow*)

Appendix A

Results From the Postexperiment Questionnaire

Participants answered a brief postexperiment questionnaire after the final free recall test. This included: (a) whether they read the cue first when restudying with a cue; (b) whether they said the correct word aloud if they had not retrieved it during practice; and (c) which condition they thought produced the best final recall performance. Question 1 examined whether participants utilized the cues during restudy. Question 2 was used to see if there was a production difference between semantic and phonemic retrieval. Because saying words aloud can produce better memory than reading them silently (i.e., the production effect; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010), it was possible that the effects of retrieval level would be confounded with the effects of production. Question 3 assessed participants' metamemory.

Question 1: Whether Participants Read the Cue First When Restudying With a Cue

In Experiment 1, six participants (three per level) claimed that they didn't read the cue first in the cued restudy condition. Removing their data and rerunning the 2 (level: semantic vs. phonemic) × 3 (review condition: retrieval vs. cued restudy vs. restudy) ANOVA on final free recall performance did not change the results. In each of the remaining experiments, one participant in said they did not read the cue first. Not surprisingly, removing their data and rerunning the 2 (level: semantic vs. phonemic) × 2 (condition: retrieval vs. cued restudy) ANOVA on final free recall performance did not change the results.

Question 2: Whether Participants Said the Correct Word Aloud if They had Missed It

In Experiment 1, seven participants indicated that they did not read the word aloud (two in the semantic level and five in the phonemic level). As above, rerunning the 2 (level) × 3 (review condition) ANOVA on final free recall performance without their data did not change the results. In Experiments 2–4, we recorded retrieval practice responses instead of overtly asking this question. On average, the number of missing responses, 0.56 and 0.47 (Experiment 2), 0.41 and 0.44 (Experiment 3), and 0.84 and 0.97 (Experiment 4) for the semantic and phonemic retrieval conditions, respectively, did not significantly differ, $ps > .05$.

Question 3: Which Condition Produced the Best Final Recall Performance

In Experiment 1, a 2 (level) × 3 (review condition) table indicated that two cells of the restudy condition had an expected count (two) less than five, so we deleted the restudy condition row. A 2 (level) × 2 (review condition) Fisher's exact test revealed no significant relation between participant's prediction and their condition, $p = .151$. Combing across levels, participants predicted that retrieval practice (frequency = 27) would benefit final recall performance more than cued restudy (frequency = 11), $\chi^2(1) = 6.737$, $p < .05$.

In Experiments 2–4, participants were asked to rank each condition in terms of predicted final recall performance. In Experiment 2, a related-samples Friedman's two-way ANOVA by ranks rejected the null hypothesis, $\chi^2(3) = 15.094$, $p < .01$. Pairwise comparisons using Bonferroni adjusted alphas revealed a difference in ranking between semantic retrieval practice (Mean rank = 2.88) and semantic restudy (Mean rank = 2), between phonemic retrieval practice (Mean rank = 3) and phonemic restudy (Mean rank = 2.12), and between phonemic retrieval practice and semantic restudy, $ps < .05$. In Experiment 3, the same analysis failed to reject the null hypothesis, $\chi^2(3) = 1.396$, $p = .706$. In Experiment 4, the analysis rejected the null hypothesis, $\chi^2(3) = 8.753$, $p < .05$. However, pairwise comparisons indicated no significant differences between mean rankings (semantic retrieval practice = 2.02, semantic restudy = 2.36, phonemic retrieval practice = 2.80, and phonemic restudy = 2.83; $ps > .05$).

(*Appendices continue*)

Appendix B

Nontargets Final Free Recall Proportion Correct: *M* (*SD*)

Experiment and review condition	Semantic	Phonemic
Experiment 1		
Restudy	0.17 (0.14)	0.14 (0.12)
Cued restudy	0.24 (0.19)	0.19 (0.17)
Retrieval practice	0.11 (0.13)	0.13 (0.16)
Experiment 2		
Cued restudy	0.17 (0.18)	0.16 (0.17)
Retrieval practice	0.10 (0.14)	0.15 (0.18)
Experiment 3		
Cued restudy	0.09 (0.16)	0.11 (0.12)
Retrieval practice	0.11 (0.14)	0.08 (0.11)
Experiment 4		
Cued restudy	0.05 (0.07)	0.06 (0.08)
Retrieval practice	0.04 (0.06)	0.06 (0.08)

Appendix C

JASP Bayesian Repeated Measures ANOVA Output

Table C1
Experiment 1: Main Effects Only Model Versus Main Effects and Interaction Model

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. Review Condition, Level, subject)	0.500	0.749	2.987	1.000	
Interaction (Review Condition × Level)	0.500	0.251	0.335	2.987	3.376

Note. All models include review condition, level, subject.

Table C2
Experiment 2: Main Effects Only Model Versus Main Effects and Interaction Model

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. Level, Review Condition, subject)	0.500	0.792	3.797	1.000	
Interaction (Level × Review Condition)	0.500	0.208	0.263	3.797	2.693

Note. All models include level, review condition, subject.

(*Appendices continue*)

Table C3
Experiment 3: Main Effects Only Model Versus Main Effects and Interaction Model

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. Level, Review Condition, subject)	0.500	0.797	3.923	1.000	
Interaction (Level × Review Condition)	0.500	0.203	0.255	3.923	3.282

Note. All models include level, review condition, subject.

Table C4
Experiment 4: Main Effects Only Model Versus Main Effects and Interaction Model

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. Level, Review Condition, subject)	0.500	0.685	2.178	1.000	
Interaction (Level × Review Condition)	0.500	0.315	0.459	2.178	4.040

Table C5
Experiment 1: All Possible Models

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject)	0.200	0.083	0.363	1.000	
Review condition	0.200	0.560	5.099	0.149	1.112
Level	0.200	0.035	0.145	2.375	1.240
Review condition + Level	0.200	0.242	1.275	0.344	1.650
Review condition + Level + Interaction	0.200	0.080	0.346	1.046	2.483

Note. All models include subject.

Table C6
Experiment 2: All Possible Models

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject)	0.200	0.034	0.142	1.000	
Level	0.200	0.010	0.042	3.320	1.802
Review condition	0.200	0.684	8.668	0.050	0.952
Level + Review condition	0.200	0.213	1.085	0.161	2.052
Level + Review condition + Interaction	0.200	0.058	0.245	0.594	3.312

Note. All models include subject.

(*Appendices continue*)

Table C7
Experiment 3: All Possible Models

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject)	0.200	3.718e−7	1.487e−6	1.000	
Level	0.200	1.083e−6	4.333e−6	0.343	1.544
Review condition	0.200	0.104	0.466	3.564e−6	1.646
Level + Review condition	0.200	0.718	10.194	5.176e−7	3.481
Level + Review condition + Interaction	0.200	0.177	0.863	2.095e−6	3.554

Note. All models include subject.

Table C8
Experiment 4: All Possible Models

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject)		1.476e−9	5.904e−9	1.000	
Level	0.200	1.343e−9	5.371e−9	1.099	1.032
Review condition	0.200	0.280	1.556	5.269e−9	2.148
Level + Review condition	0.200	0.491	3.858	3.006e−9	1.938
Level + Review condition + Interaction	0.200	0.229	1.188	6.448e−9	2.024

Note. All models include subject.

Received December 13, 2019
Revision received June 28, 2020
Accepted August 3, 2020 ■