AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

© 2024 American Psychological Association
ISSN: 0278-7393

2024, Vol. 50, No. 11, 1791–1810
https://doi.org/10.1037/xlm0001382

A Grain of Truth in the Grain Size Effect: Retrieval Practice Is More Effective When Interspersed During Learning

Hilary J. Don[1], Shaun Boustani[1], Chunliang Yang[2], and David R. Shanks[1]
[1] Division of Psychology and Language Sciences, University College London
[2] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University

Retrieval practice is a powerful method for consolidating long-term learning. When learning takes place over an extended period, how should tests be scheduled to obtain the maximal benefit? In an end-test schedule, all material is studied prior to a large practice test on all studied material, whereas in an interim test schedule, learning is divided into multiple study/test cycles in which each test is smaller and only assesses material from the preceding study block. Past investigations have generally found a difference between these schedules during practice but not during a final assessment, although they may have been underpowered. Five experiments confirmed that final assessment performance was better in students taught using interim than end tests in list (Experiments 1, 2, and 5) and paired associate (Experiments 3 and 4) learning, with a meta-analysis of all available studies ($k = 19$) yielding a small- to medium-sized effect, $g = 0.25$, 95% confidence interval [0.09, 0.42]. Experiment 5 finds that the higher level of practice retrieval success in interim tests contributes to the grain size effect, but the effect is eliminated if these tests are too easy. Additional analyses also suggest that the forward testing effect, in which tests promote subsequent learning, may be a major cause of the grain size effect. The practical and theoretical implications of these demonstrations of robust grain size effects are discussed.

Keywords: grain size, testing effect, retrieval practice, desirable difficulty

Supplemental materials: https://doi.org/10.1037/xlm0001382.supp

Testing through retrieval practice is a potentially powerful educational tool. Research has found that attempting to retrieve previously studied information can enhance long-term memory, an effect known as the *testing effect* (Adesope et al., 2017; Rowland, 2014; Yang et al., 2021), as well as the learning of new information, an effect known as the *forward testing effect* (FTE) or *test-potentiated new learning* (Chan et al., 2018). Despite decades of research, however, remarkably little is known about a key issue: What is the optimal placement of tests during a learning episode?

Imagine students are taking a 1-hr class. One option is to have them study all the information and then attempt to retrieve as much of that information as possible, which we will call an *end-test* schedule (see Figure 1 for a schematic illustration). Another option is to segment the information (e.g., into 15-min sections) and have students attempt to recall information after studying each section, which we call an *interim test* schedule. The amount of content covered during the test is known as the *grain size* of recall practice, which is small in the interim test condition and large in the end-test condition. A limited amount of research, reviewed below, has examined the potential differences between these schedules on both the practice tests (i.e., aggregate performance during the interim tests compared to the end test) and a final criterial assessment (i.e., a test assessing all learned sections). The difference in performance in these criterial tests is known as the *grain size effect* (Wissman & Rawson, 2015).

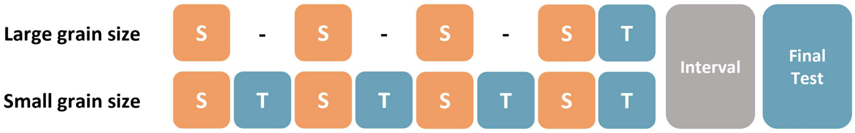
The grain size hypothesis states that interim tests of smaller segments of information throughout learning should be more beneficial for long-term retention compared to testing large segments at the end of learning. Retrieval success during practice

Jonathan Tullis served as action editor.
Hilary J. Don https://orcid.org/0000-0003-2961-5212
Shaun Boustani https://orcid.org/0000-0002-2976-3699
Chunliang Yang https://orcid.org/0000-0002-4995-7300
David R. Shanks https://orcid.org/0000-0002-4600-6323
The data, analysis scripts, and materials for all experiments are available at the Open Science Framework and are accessible at https://osf.io/8wuny/.
The authors have no known conflicts of interest to disclose. This research was funded by Grant ES/S014616/1 from the United Kingdom Economic and Social Research Council awarded to David R. Shanks. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.
Hilary J. Don played a lead role in data curation, formal analysis, validation, and visualization and an equal role in conceptualization, investigation, methodology, software, writing–original draft, and writing–review and editing. Shaun Boustani played a supporting role in formal analysis and an equal role in investigation, methodology, software, writing–original draft, and writing–review and editing. Chunliang Yang played an equal role in funding acquisition and writing–review and editing. David R. Shanks played a lead role in project administration and supervision and an equal role in conceptualization, funding acquisition, methodology, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Hilary J. Don, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. Email: h.don@ucl.ac.uk

Figure 1

Schematic Illustration of the Grain Size of Study/Test Cycles



Note. In the large grain-size condition, practice retrieval of all studied material occurs in a single end test. In the small grain-size condition, there is a retrieval practice opportunity following each successive chunk of studied material. The final test is administered either shortly after the study phase (immediate test) or after a delay (delayed test). S = study; T = practice test. See the online article for the color version of this figure.

tests is fundamental to long-term retention (Pyc & Rawson, 2009, but see Chan et al., 2024), and a meta-analysis confirms that the testing effect tends to be correlated with practice retrieval success (Rowland, 2014). Interim tests are likely to lead to higher practice retrieval success than end tests, due to testing smaller chunks of information after shorter intervals on average, and should therefore facilitate better retention (Lavigne & Risko, 2018; Uner & Roediger, 2018; Weinstein et al., 2016; Wissman & Rawson, 2015). Indeed this is simply a “list length” effect, whereby recall probability is greater in short than long lists (Underwood, 1978). There are also further benefits of interim tests that should intuitively lead to better long-term learning than end tests. Interim testing during a lecture can decrease the amount of task-irrelevant mind wandering that students experience, and it can increase their employment of positive study behaviors, such as note taking (Jing et al., 2016; Szpunar et al., 2013, 2014). Similarly, the *cognitive antidote principle* suggests that making a monotonous task more difficult can increase attention and performance (Kole et al., 2008), which should promote learning (Healy et al., 2017). Interim testing can reduce overconfidence (Szpunar et al., 2014) and improve integration of content across sections (Jing et al., 2016).

In addition, as Figure 1 illustrates, the practice tests are distributed when the grain size is small but massed when it is large. Research on the testing effect shows that there is a considerable advantage to spacing out retrieval practice attempts (for a meta-analysis, see Latimier et al., 2021). This research explores the effects of the timing of repeated retrieval attempts after a *single* study episode (e.g., Karpicke & Roediger, 2007) and so does not examine the grain size of study/test cycles but nonetheless points to a potential reason why interim tests might be more advantageous for long-term learning. Finally, and perhaps most critically, interim testing also potentiates new learning of subsequent information (see Chan et al., 2018). In an interim test schedule, each test carries the potential to facilitate learning of the next section of information, which could result in substantially greater overall learning.

Against this theoretical backdrop, it is surprising that most efforts to demonstrate grain size effects in final test recall have failed. The first investigation of the grain size effect was by Duchastel and Nungester (1984). They presented student participants with a 1,700-word history text consisting of 12 unrelated paragraphs and compared three groups: interim test, end test, and a control treatment. In the interim test group, students were asked a short-answer question after studying each paragraph, whereas the end-test group answered all 12 questions after the entire text had been studied. The control group studied the entire text with no practice testing. After a 2-week retention interval, all groups completed a final assessment that included the same 12 short-answer questions as in the practice stage. Duchastel and Nungester found that although both testing schedules benefited learning of the old questions (a testing effect), there was no difference between interim and end testing, that is, there was no grain size effect in final recall.

Since then, investigations of the grain size effect have found similar results using varied methods and materials. Primarily, although interim tests lead to a substantial boost to practice test performance, this effect is fragile and does not appear to translate to improved final test performance. This lack of a grain size effect has been found in replications using texts on different topics (Wissman & Rawson, 2015), studies using multimedia PowerPoint presentations (Weinstein et al., 2016), textbook chapters (Uner & Roediger, 2018), and a short online course (Latimier et al., 2020).

Research has also suggested that the grain size effect does not become stronger as the length of time between the practice test and the final assessment increases. For example, the grain size effect was absent in studies using 15- and 20-min (Wissman & Rawson, 2015), 48-hr (Uner & Roediger, 2018; Wissman & Rawson, 2015), 1-week (Latimier et al., 2020; Weinstein et al., 2016), and 2 week intervals (Duchastel & Nungester, 1984), and even retention intervals of longer than a month (Weinstein et al., 2016), despite these studies observing robust effects during practice. The lack of any influence of retention interval on the grain size effect in final recall differentiates it from the testing effect, which meta-analyses have suggested emerges and grows with longer retention intervals (Adesope et al., 2017; Rowland, 2014; Yang et al., 2021). This confirms that the absence of a grain size effect in the literature is unlikely to be due to insufficiently long retention intervals.

Similarly, research has also suggested that the delay between study and practice test does not significantly influence the grain size effect. Wissman and Rawson (2015), in their Experiments 1 and 2, demonstrated that a grain size effect was absent after both no delay and a 2-min delay. This is important as it rules out the possibility that participants in the interim test group were recalling information from short-term rather than long-term memory, which would result in poorer retention. It is interesting to compare this result to meta-analyses on the FTE, which have suggested that delay duration is negatively associated with FTE magnitude (Chan et al., 2018). In addition, studies by Wissman and Rawson (2015) suggested that the absence of a grain size effect is not due to increased recall of unimportant details in interim tests or disruption of integration across sentences.

More recently, one study has demonstrated a significant grain size effect in an immediate test using simpler study materials. Healy et al. (2017) had participants study eight lists of artificial facts and showed that interim tests improved final assessment performance and increased self-reported engagement compared to an end test. This difference emerged to a greater extent for later lists.

Supplementing this narrative review, we report a meta-analysis of all research on the grain size effect after describing a set of new experiments.

Why Has Past Research Failed to Observe Strong Grain Size Effects?

It is useful to consider some potential reasons why past studies failed to observe a grain size effect or only obtained small effects. One possibility is that the interim tests may not have required enough effortful retrieval to enhance long-term learning, in line with the *desirable difficulties* framework (Bjork, 1994). Because of the short lag between study and retrieval in the interim test conditions and smaller memory loads of items to be recalled, retrieval will not involve as much effort compared to recalling all material at the end of study. Thus, although the ease of interim tests leads to high initial practice performance, yielding a clear effect in practice, there may not have been sufficient effortful retrieval to facilitate long term retrieval at the immediate test (Weinstein et al., 2016; Wissman & Rawson, 2015).

A second possibility is that when text or complex materials are used (as is the case in 13/14 of the individual experiments included in the studies reviewed above), frequent breaks in the learning episode might interfere with the formation of a coherent mnemonic representation (Duchastel & Nungester, 1984; Healy et al., 2017; Latimier et al., 2020). As highlighted by Latimier et al. (2020), end tests may facilitate better understanding of the materials as a cohesive whole, in comparison to interim tests which might interrupt the flow of learning and induce switch costs (Pashler, 2000). In this case, the direction of the grain size effect might be dependent on the relationship between content across lists. Disjointed facts not requiring whole-text comprehension may not suffer from the impediment of relational processing that interim testing might cause, resulting in a benefit of interim testing (such as that seen in Healy et al., 2017). In contrast, materials requiring integration across lists may benefit less from interim tests. Wissman and Rawson (2015; Experiment 7) stated that they found little evidence for this hypothesis, although instructions encouraging participants to make connections between sections tended to reduce the benefit of interim tests.

The final possibility concerns test-potentiated new learning. As noted earlier, when interim tests are administered, encoding of each new set of materials might improve, relative to the equivalent set in the end-test group (the well-established FTE). The FTE is typically studied in experiments that compare an interim test group (equivalent to the small grain-size group in Figure 1) with an otherwise identical group that engages in some other nonretrieval activity (often restudy) in place of the interim tests. The key outcome—the FTE—refers to the finding that final test recall of the last chunk of material is enhanced in the interim test group (Chan et al., 2018). Thus interim tests facilitate the learning and retention of subsequent information. Several nonexclusive mechanisms to explain the FTE have been proposed and evaluated (see Chan et al., 2018; Shanks et al., 2023; Yang et al., 2018, 2022). First, interim tests insulate new material against the buildup of proactive interference from preceding materials via an enhancement in list discrimination and reduction in prior-list intrusions. Second, experience of retrieval failures during retrieval practice may induce participants to adopt more efficient strategies for encoding and retrieval of subsequent material. Finally, interim tests may serve to maintain motivation and concentration and reduce fatigue and mind wandering.

Given the importance of the FTE in potentially explaining the grain size effect, surprisingly few studies have established that their methods were sufficient to produce an FTE. Wissman and Rawson (2015, Experiment 4) found better recall of the final section in the interim than end-test group during both practice recall and final recall, indicative of an FTE, although their study lacked an appropriate exposure-matched control group. Their failure to observe a grain size effect is therefore a challenge to the idea that the FTE is a necessary and sufficient precondition for obtaining a grain size effect. However, other studies may have employed experimental conditions that were not conducive to obtaining an FTE, and this, in turn, might explain the absence of a grain size effect.

The Present Study

Our literature review revealed several gaps in the grain size effect literature, which we address in the present research. Notably, only one study has used simple materials: Healy et al. (2017) presented participants with artificial facts about plants and found a significant grain size effect. The use of novel unrelated materials could decrease potential negative impacts of interim testing, such as interfering with the formation of a coherent, whole-text, mnemonic representation (Duchastel & Nungester, 1984; Healy et al., 2017; Latimier et al., 2020), and allow the benefits of interim testing to be more evident. Using simple materials such as word lists (Experiments 1, 2, and 5) and paired associates (Experiments 3 and 4) also allows for easier manipulation of relatedness within and between lists. To further test the longevity of potential grain size effects, the majority of the current experiments also included a second delayed cumulative assessment at a time point following the immediate test (24 hr, 48 hr, or 1 week later).

Few studies have assessed whether the grain size effect is specific to recall. Latimier et al. (2020) compared test and restudy in small, medium, and large grain sizes and found a benefit of testing over restudy for large and medium grain sizes but not for small ones due to elevated performance in the restudy group. This suggests that it may be the grain size of interim tasks that is beneficial, regardless of the type of task (test or restudy). To assess this further, Experiment 1 included interim and end-restudy conditions.

Another potential explanation for the absence of a grain size effect at immediate test is that the format of recall changes from practice to the final assessment for the interim test group but not for the end-test group. In all previous research on the grain size effect, the final assessment has been a cumulative test where participants are asked to recall all content from the learning phase (we will refer to this as a "whole" format). This is an exact repeat of what is required in the end-test group when they take the practice test but constitutes a change in format for the interim test participants, who are required to restrict recall to a particular study section (we will refer to this as a “list” format). This mismatch in format could create a situation where the strategies adopted during practice testing to support recall are more useful in the final assessment for the end-test group than the interim test group. One way of assessing this claim is to measure the grain size effect when the final assessment is either in whole or list format (Experiment 1). If worse performance in the interim test group is due to a mismatch in format, then the grain size effect should be larger when the final assessment requires list recall.

The additional aim of the present research was to test theories that may explain the pattern of results observed in prior (and current) investigations of the grain size effect. Experiment 5 tested whether desirable difficulty can explain the poor retention of correctly recalled information in the interim test group and whether retrieval success during practice transfers to improved immediate test recall. Experiment 2 and combined experimental analyses investigated the contribution of the FTE to the grain size effect. Experiment 2 included a control group to assess whether the FTE was evident in the interim test condition, and we separately assessed recall for earlier versus later lists, across studies.

The experiments reported in this study were designed to fill these gaps in the literature and to investigate the impact of retrieval success and desirable difficulty on the grain size effect and accelerated forgetting following interim testing. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Data and analysis scripts are available on Open Science Framework at https://osf.io/8wuny/.

Experiment 1

Experiments 1 and 2 were preregistered (Don et al., 2024) and assessed the grain size effect using lists of related and unrelated words, respectively. Both were experimenter-paced. A standard task was used where participants were tested either after studying each list—an interim test schedule—or after studying all lists—an end-test schedule (see Figure 2). We expected to observe a significant effect of the interim or end-test manipulation in practice and a small grain-size effect in the immediate test. We collected data from a large sample in order to detect a small effect size.

Experiment 1 also assessed whether the effect of grain size was dependent on the test format and if it was recall-specific by using a 2 × 2 × 2 between-subjects design crossing interval task (restudy vs. test), immediate test format (whole vs. list), and task grain size (interim vs. end). If the effect is recall specific, then there should only be an impact of grain size on test conditions and not restudy conditions. If the grain size effect is sensitive to the match between practice and immediate test format, then the impact of grain size should be larger when using a list format in the immediate test.

Method

Participants

Participants were first year undergraduate psychology students at the University of Sydney participating as part of a tutorial class activity. All enrolled students were eligible to participate. We collected data from 680 participants. One participant was excluded for incomplete data from Session 1 and completing the experiment a second time and another for a nonserious attempt. This left 678 participants in Session 1 ($M_{\text{age}} = 19.7$, $SD = 4.2$; 450 identified as female, 219 male, three nonbinary, six undisclosed). Four hundred twenty-six participants completed the delayed test in Session 2. Two participants completed Session 2 twice, and we kept only the data from the first attempt. Allocation of participants to each condition was random (see Supplemental Materials for *n*s by condition for each session). In the Supplemental Materials, we report analyses that found no significant differences in participants who remained in the study and those who did not.

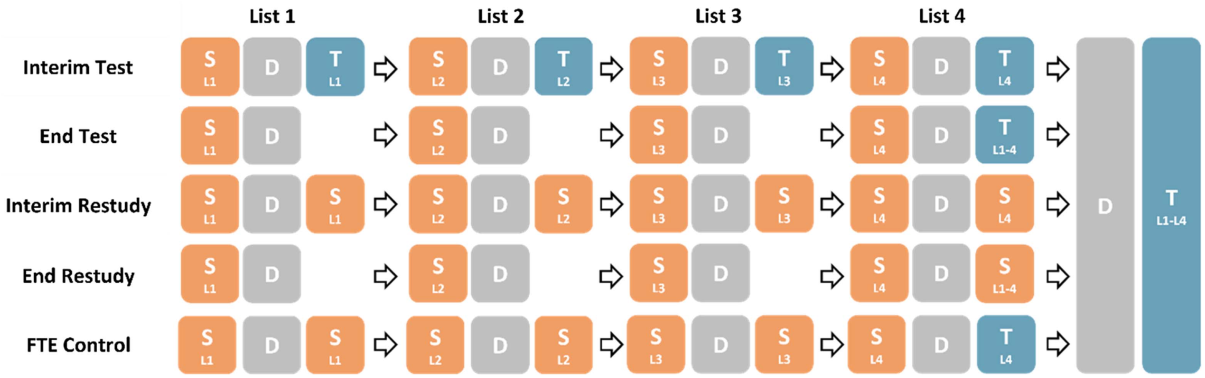
Sample size was limited by the number of students enrolled in the course. However, a post hoc sensitivity analysis using G*Power 3.1 (Faul et al., 2009) indicated that a sample size of 426 is sufficient to detect a small effect size of $f = 0.136$ for a main effect of grain size with $\alpha = .05$ and power = .80 in a 2 × 2 × 2 factorial analysis of variance (ANOVA).

Materials

We created four lists of 16 words, with each list containing four exemplars from four categories, taken from Van Overschelde

Figure 2

Schematic of Conditions Used in the Current Experiments


List 1
List 2
List 3
List 4
Interim Test
S L1 D T L1 S L2 D T L2 S L3 D T L3 S L4 D T L4
End Test
S L1 D S L2 D S L3 D S L4 D T L1-4
Interim Restudy
S L1 D S L1 S L2 D S L2 S L3 D S L3 S L4 D S L4
End Restudy
S L1 D S L2 D S L3 D S L4 D S L1-4
FTE Control
S L1 D S L1 S L2 D S L2 S L3 D S L3 S L4 D T L4
D
T L1-L4

Note. S = study phase; D = distractor task; T = test; L1–L4 = List 1–4; FTE = forward testing effect. See the online article for the color version of this figure.

et al. (2004; see Supplemental Materials). The four categories were reading materials, animals, fabrics, and kitchen utensils. The average taxonomic frequencies did not differ between the categories, $M_{\text{materials}} = .25$, $SD = .31$, $M_{\text{animals}} = .36$, $SD = .27$, $M_{\text{fabrics}} = .27$, $SD = .27$, $M_{\text{utensils}} = .30$, $SD = .34$, $F(3, 60) = 0.39$, $p = .76$, $BF_{01} = 7.96$. They also did not differ across the four lists (range = .26–.33), $F(3, 60) = 0.20$, $p = .90$, $BF_{01} = 9.63$. Each list was distinguished by a different color (blue, red, green, and yellow) in order to facilitate list format recall. Word order within a list and list order were randomized.

Procedure

Ethical approval was obtained from the University of Sydney ethics committee (Project Number 2018/930). Experiment 1 was programmed and run in Qualtrics. Prior to beginning the experiment, informed consent was obtained and participants completed a demographic questionnaire. After the conclusion of the experiment, consent to use the data was also obtained. The study was run in the first tutorial of the university semester, with attendance split between online and in-person participation. At the beginning of the experiment, participants were informed that they would study several word lists and that, after each list, they would complete a task.

For all conditions, the study list began with a heading stating the color of the list for 2 s (i.e., the BLUE list, the RED list, the GREEN list, the YELLOW list), followed by a fixation cross for 1 s. After this, words appeared on the screen one at a time for 4 s, in the corresponding color text, followed by a 500-ms interstimulus interval, progressing automatically. The subsequent distractor task was 30 s of simple arithmetic (e.g., 85 + 29 = ?).

In the interim conditions, four interval tasks were completed, each following directly after the distractor task. The interim restudy group restudied the previous list. In the interim test group, participants were instructed to retrieve all words from the previous list. Words were typed and remained on the screen. The test lasted 64 s before automatically progressing.

In the end conditions, one interval task was completed following directly after the distractor task of List 4. The end-restudy group restudied all words from all four lists, with words presented in a random order. In the end-test group, participants were instructed to retrieve all words from the previous lists. Words were typed and remained on the screen. The test lasted 4 min and 16 s before automatically progressing.

Following completion of the final interval task, all participants completed a 2-min distractor task, which served as the immediate retention interval. After this, they completed a cumulative test. In the whole-test format conditions, participants were asked to recall all words from all the previous lists in any order and were given 4 min and 16 s to do so. In the list test format, participants were told to recall words from a particular list selected in random order and were given 64 s to recall words in that list. Lists were denoted by the list color, for example, “recall all words from the BLUE list.” In both formats, participants were allowed to proceed once half the allotted time had elapsed (i.e., 128 s in the whole format and 34 s in the list format). A second delayed cumulative test was administered after a 1-week retention interval in the following week’s tutorial, in whole format as previously described.

Data Analysis

Free recall was scored using the *amatch* function in R. Any entries with a maximum Levenstein distance of 2 to the closest matching word, but which were not an exact match, were checked and scored manually. In the list format condition, we were interested in whether the match between study and final test would assist recall of the studied words rather than whether participants could correctly recall which list words were from per se. Therefore, in the immediate test, we scored any studied word recalled at any point in the immediate test as correct, regardless of whether it was recalled in the appropriate cued list test. This also allows for a fairer comparison to the whole-test format, which had no constraints on when words could be recalled. Duplicated recalls were counted only once.

All statistical analyses were conducted in R (R Core Team, 2024). We report p values as well as Bayes factors to assess the strength of evidence for the alternative hypothesis (BF_{10}) or null hypothesis (BF_{01}). Bayes factors were computed via Bayesian ANOVA or t tests with default priors. Results are reported according to the preregistration plan. Exploratory analyses are reported as such.

Results

The number of correctly recalled words in the practice, immediate, and delayed tests is shown in Figure 3.

Practice Test

An independent samples t test found significantly better practice test recall in the interim test group (summed across the four interim tests; $M = 34.65$, $SD = 8.26$) compared to the end test group ($M = 21.00$, $SD = 9.39$), $t(358) = 14.66$, $p < .001$, $d = 1.55$, $BF_{10} = 9.56 \times 10^{34}$. This indicates an effect of the grain size manipulation in practice recall.

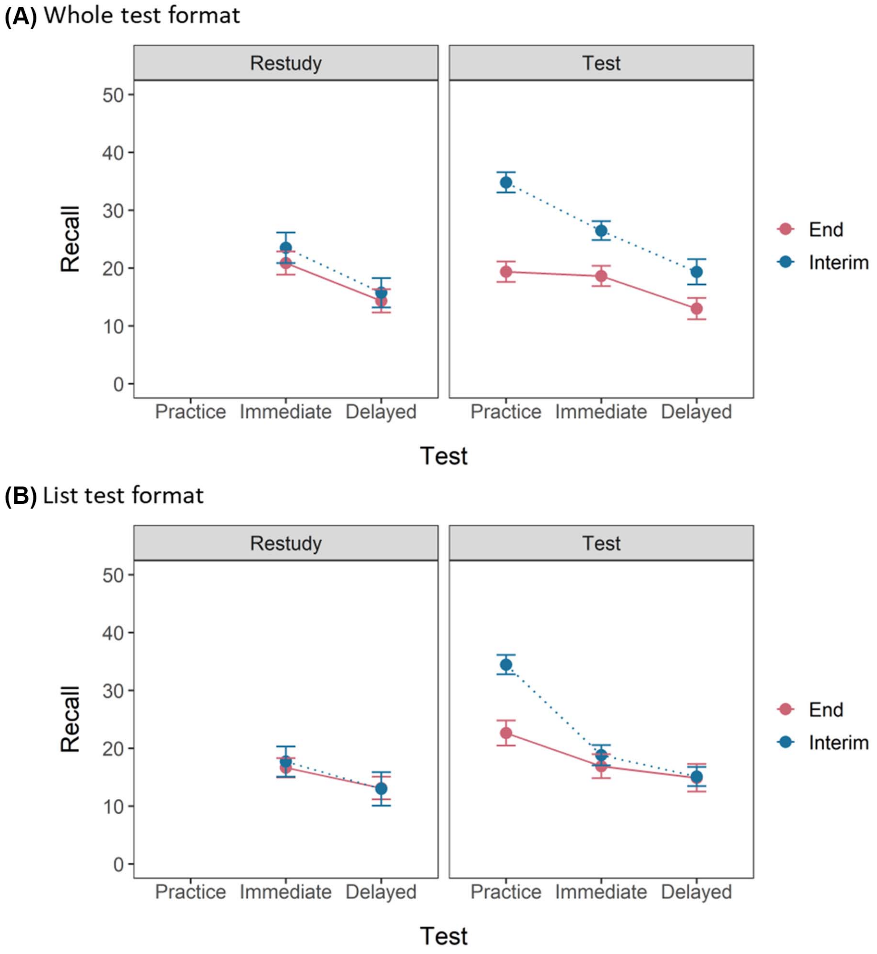
Immediate Test

To preempt our results, we did not observe the influence of test format that we expected. Instead, list recall test format produced very poor recall. For brevity, we therefore report the comparison of this factor in the Supplemental Materials.

Our preregistered analysis plan included separate ANOVAs to determine whether there was a significant grain size effect within each test format. There was no significant effect of task, grain size, or interaction between task and grain size in the list test condition, largest $F(1, 328) = 2.15$, $p = .144$, $\eta_p^2 = .006$, $BF_{01} = 2.52$. In the whole-test format, there was no main effect of task, $F(1, 342) = 0.13$, $p = .719$, $\eta_p^2 < .001$, $BF_{01} = 6.96$, but there was a significant main effect of grain size, $F(1, 342) = 27.50$, $p < .001$, $\eta_p^2 = .074$, $BF_{10} = 105{,}276.47$, and an interaction between task and grain size, $F(1, 342) = 6.70$, $p = .010$, $\eta_p^2 = .019$, $BF_{\text{incl}} = 3.78$. Independent samples t tests comparing end and interim conditions within the whole-test format showed a grain size effect for the test conditions, $t(181) = 6.50$, $p < .001$, $d = 0.96$, but not the restudy conditions, $t(161) = 1.63$, $p = .106$, $d = 0.26$. We therefore only saw effects of grain size when participants were allowed free recall of all lists in the immediate test and not when they were asked to recall words by list.

Figure 3

Experiment 1: Number of Correctly Recalled Words in the Practice, Immediate, and Delayed Tests in the Whole (A) and List (B) Test Format Groups



Note. See the online article for the color version of this figure.

Delayed Test

We analyzed the delayed test with a similar 2 × 2 × 2 ANOVA, but in this case the test format (list vs. whole) factor refers to how each group was tested in the immediate test—all participants were tested under a whole format in the delayed test. The ANOVA found a significant main effect of task, with greater recall in the test (M = 15.60, SD = 7.93) than restudy (M = 14.04, SD = 8.32) conditions, $F(1, 418) = 3.92$, $p = .048$, $\eta_p^2 = .009$, $BF_{10} = 0.71$. There was again a significant main effect of grain size, with greater recall in the interim (M = 15.94, SD = 8.41) than end (M = 13.89, SD = 7.82) conditions persisting in the delayed test, $F(1, 418) = 6.41$, $p = .012$, $\eta_p^2 = .015$, $BF_{10} = 2.74$. However, the interaction between task and grain size was no longer significant, $F(1, 418) = 2.91$, $p = .089$, $\eta_p^2 = .007$, $BF_{\text{excl}} = 1.68$. Nevertheless, independent samples t tests comparing the end and interim conditions showed a significant grain size effect for test conditions, $t(218) = 3.15$, $p = .002$, $d = 0.43$, $BF_{10} = 14.76$, but not restudy conditions, $t(204) = 0.47$, $p = .639$, $d = 0.07$, $BF_{01} = 5.91$.

Participants tested via the whole-test format in the immediate test (M = 15.67, SD = 8.14) showed better delayed recall than those tested in the list format (M = 14.10, SD = 8.12), $F(1, 418) = 4.06$, $p = .044$, $\eta_p^2 = .010$, $BF_{10} = 0.57$. This was driven primarily by a benefit in the whole–interim conditions, as indicated by an interaction between grain size and format, $F(1, 418) = 6.09$, $p = .014$, $\eta_p^2 = .014$, $BF_{\text{incl}} = 2.91$. There was no three-way interaction between task, grain size, and format, $F(1, 418) = 2.14$, $p = .144$, $\eta_p^2 = .005$, $BF_{\text{excl}} = 1.93$.

We again ran two separate ANOVAs for each immediate test format. There was no significant effect of task, grain size, or interaction between task and grain size in the list test condition, largest $F(1, 205) = 2.99$, $p = .085$, $\eta_p^2 = .014$, $BF_{01} = 1.63$. In the whole-test format, there was no main effect of task, $F(1, 213) = 1.11$, $p = .293$, $\eta_p^2 = .005$, $BF_{01} = 3.61$, but similarly to the immediate test, there was a significant main effect of grain size, $F(1, 213) = 13.15$, $p < .001$, $\eta_p^2 = .058$, $BF_{10} = 87.81$, and an interaction between task and grain size, $F(1, 213) = 5.28$, $p = .023$, $\eta_p^2 = .023$, $BF_{\text{incl}} = 2.10$. An independent samples t test comparing the end and interim conditions within the whole-test format only showed a grain size effect for tests, $t(107) = 4.43$, $p < .001$, $d = 0.85$, but not for restudy, $t(106) = 0.89$, $p = .375$, $d = 0.17$.

Retention Across Tests

In addition to the preregistered analysis plan, we compared recall across the practice and immediate tests for the test groups. Details are given in the Supplemental Materials. Confirming the pattern shown in Figure 3, the interim test groups forgot an average of 11.92 ($SD = 6.18$) words, while the end-test groups forgot an average of 3.2 ($SD = 5.04$) words.

We also analyzed recall across the immediate and delayed tests for all groups (details in the Supplemental Materials). This revealed a greater benefit of testing over restudy in the delayed than the immediate test as well as a greater difference between interim and end conditions in the immediate than the delayed test. The interim test groups forgot an average of 5.71 ($SD = 5.56$) words, and the end-test groups forgot an average of 3.41 ($SD = 6.32$) words.

Testing Effect

In the immediate test, there was no significant main effect of task, suggesting an absence of a testing effect (although a testing effect was clearly evident in the delayed test). To investigate this further, we compared test and restudy in each of the grain size conditions separately. The effect of task reached significance in the interim test condition, $t(162) = 1.98$, $p = .049$, $d = 0.31$, but not the end-test condition, $t(180) = 1.66$, $p = .099$, $d = 0.25$. Similarly, although there was an overall effect of task in the delayed test, this was primarily driven by the interim test group, $t(98) = 2.17$, $p = .033$, $d = 0.44$, and not the end-test group, $t(115) = 0.96$, $p = .34$, $d = 0.18$.

Discussion

Experiment 1 demonstrated a grain size effect (under the whole-test format) in which both immediate and delayed recall benefitted from a smaller (interim test) compared to a larger (end test) grain size. Despite this, the grain size manipulation (interim/end test) had a much larger impact in practice than in the immediate test, consistent with the findings of earlier studies, and was also specific to tests, as the benefit of a small grain size only occurred for the test and not restudy conditions. In sum, interim testing yields better immediate recall than an end test and better recall than interim restudy opportunities (a conventional testing effect).

These effects were moderated by test format but not in the way we anticipated. We found no evidence that the grain size effect is sensitive to the match between practice and immediate test format: The difference between the interim and end tests was smaller, not larger, when using a list format in the immediate test. In fact, the grain size effect in immediate and delayed recall was only observed under the whole-test format. It is possible that requiring participants to recall items from one list at a time (in the list format) in an experimenter-determined order may have interfered with recall strategies, reducing overall recall levels. It is also possible that the repetition of categories across lists made recall particularly difficult for participants in the list format groups. Other methods might reveal a benefit of matching grain size.

Experiment 2

Experiment 2 included an FTE control group where participants restudied all but the final list, for which they completed a criterial test. This group was included to confirm that our methods produced an FTE, where interim tests boost learning and recall for subsequent lists. As discussed, one potential causal factor of the grain size effect is that interim tests potentiate new learning (Chan et al., 2018). However, only Wissman and Rawson (2015) have demonstrated the absence of a grain size effect in the presence of a robust FTE (albeit without an exposure-matched control). We aimed to replicate the FTE using our word list materials and an appropriate control group.

Method

Participants

Participants were enrolled using the online pool Prolific (https://www.prolific.co/) in return for monetary compensation. They were eligible to participate if they were fluent in English, had a Prolific score >90, were between 18 and 60 years old, and did not previously participate in any related studies run by this research group. Sample size was determined by a power analysis using G*Power 3.1 (Faul et al., 2009), which indicated that a minimum total sample size of 49 participants per group was needed to detect a medium- to large-sized effect (Cohen's $d = 0.60$) with $\alpha = .05$ and power = .90 in an independent samples t test. This effect size was chosen as a recent meta-analysis found a medium to large FTE (Hedges' $g = 0.61$; Chan et al., 2018) in studies using a standard procedure and a restudy control. This effect size also provides ample power to detect a difference between the end-test and interim test groups, which we found to be very large when using whole-format cumulative assessments in Experiment 1 (immediate test: $d = 0.96$; delayed test: $d = 0.85$).

We collected data from 166 participants, allocated randomly to the interim test, end-test, and FTE control groups. Five reported taking notes and 12 had incomplete data sets and accordingly were removed from the analyses. This left 149 participants in Session 1 ($M_{\text{age}} = 37.95$, $SD = 9.81$; 105 identified as female, 44 male). One hundred thirty-seven participants completed the delayed test in Session 2 (sample sizes for each condition in each session are reported in the Supplemental Materials).

Materials

Four new lists of 16 medium-frequency words randomly selected from the SUBTLEXus database (Brysbaert & New, 2009) were created (see Supplemental Materials). Words were between four and eight letters and medium frequency as defined by a Zipf number between 2 and 4 (van Heuven et al., 2014). The words had a lexical decision accuracy of greater than 90% according to the English Lexicon Project database (Yarkoni et al., 2008). Word order within a list and list order were randomized. The change from related (Experiment 1) to unrelated (Experiment 2) word lists was intended to extend the generality of the results.

Procedure

Ethical approval was obtained from the Ethics Committee of the University College London (UCL) Research Department of

Experimental Psychology (ID No: EP/2020/007). Experiment 2 was programmed and run in Qualtrics. The design and procedure for the interim and end-test groups were nearly identical to the whole-format test conditions used in Experiment 1. The only difference was that the option to terminate the immediate test halfway through was removed.

In the FTE control group (see Figure 2), four interval tasks were completed, each following directly after the distractor task. Participants restudied Lists 1–3 and completed a test after List 4, following the same procedure as in the interim test group.

Results

The number of correctly recalled words in the practice, immediate test, and delayed test for each group is shown in Figure 4. The following analyses were preregistered (unless otherwise stated).

Practice Test

An independent samples *t* test found significantly better recall in the interim test group (summed across the four interim tests; M = 29.67, SD = 9.67) compared to the end-test group (M = 16.47, SD = 10.97), $t(98) = 6.39$, $p < .001$, $d = 1.28$, $BF_{10} = 1.68 \times 10^6$, indicating an effect of the grain size manipulation in practice recall.

Immediate Test

Replicating the grain size effect in immediate recall, an independent samples *t* test found significantly better recall in the interim test (M = 20.12, SD = 9.40) than the end test (M = 14.94, SD = 10.35) group, $t(98) = 2.62$, $p = .01$, $d = 0.52$, $BF_{10} = 4.21$.The FTE control group recalled a mean of 15.74 (SD = 12.08) words in the immediate test.

Delayed Test

An independent samples *t* test found no difference in recall between the interim (M = 5.15, SD = 4.98) and end (M = 5.43, SD = 7.58) tests, $t(90) = 0.21$, $p = .83$, $d = 0.04$, $BF_{01} = 4.48$. Thus unlike in Experiment 1, the grain size effect did not persist across a delay. While this fails to replicate the equivalent effect from Experiment 1, a later analysis aggregating data across experiments does confirm a group effect in delayed tests, suggesting that the present null result is due to sampling error. The FTE control group recalled a mean of 4.13 (SD = 5.83) words in the delayed test.

Retention Across Tests

We again compared recall across tests in the interim and end-test conditions (see Supplemental Materials for details). This confirmed the pattern in Figure 4 of a greater difference between interim and end-test groups in the practice than immediate test. The interim test group forgot an average of 9.55 (SD = 4.97) words and the end-test group an average of 1.53 (SD = 3.33) words. A comparison of forgetting from the immediate to delayed test found a greater benefit of interim over end grain sizes in the immediate than the delayed test, as shown in Figure 4. The interim test group forgot an average of 15.70 (SD = 6.53) words and the end-test group an average of 9.63 (SD = 6.96) words.

FTE

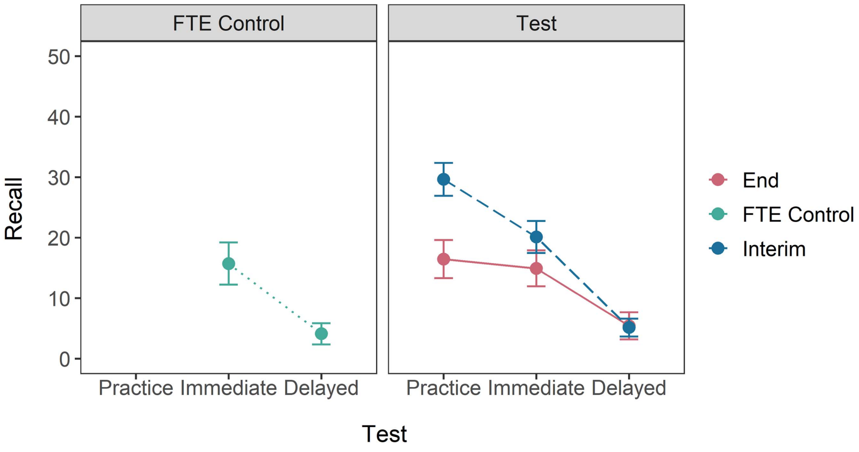
According to the preregistration plan, a one-tailed independent samples *t* test found significantly better recall in the criterial test (the List 4 test) for the interim test group (M = 7.71, SD = 3.21) compared to the FTE control group (M = 5.04, SD = 3.43), $t(96) = 3.97$, $p < .001$, $d = 0.80$, $BF_{10} = 330.98$, revealing a robust FTE. The number of correctly recalled words in the criterial test for each group is shown in Figure 5.

Discussion

Experiments 1 and 2 found a significant effect of study/test schedule in practice and a significant grain size effect in the immediate test. The reduction in effect size from practice to immediate test appeared to be due to greater forgetting in the interim test group.

Figure 4

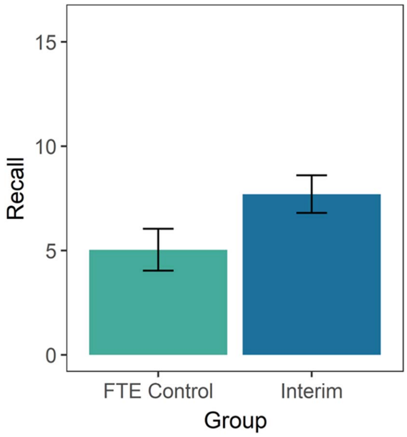
Experiment 2: Mean Correct Recall in the Practice, Immediate Test, and Delayed Tests for Each Condition



Note. FTE = forward testing effect. See the online article for the color version of this figure.

Figure 5

Experiment 2: List 4 Recall in the FTE Control and Interim Test Group, Demonstrating a Forward Testing Effect (FTE)



Note. See the online article for the color version of this figure.

The presence of a grain size effect in immediate test recall in both experiments suggests that the relatedness of simple materials (categorized word lists in Experiment 1, unrelated words in Experiment 2) does not have a large impact, although this conclusion rests on a between-experiment comparison, and the experiments were run in different contexts (lab-based vs. online).

The results from Experiment 2 confirmed that the interim test group produced a reliable FTE, which could be a mechanism of the grain size effect. We later present an evaluation of immediate test recall by list where an FTE account predicts a larger difference between groups for later lists.

Experiment 3

The primary aim of Experiments 3 and 4 was to ascertain whether the grain size effect found in Experiments 1 and 2 generalizes to a different form of learning, paired associates. Participants learned foreign language word pairs without (Experiment 3) and with (Experiment 4) feedback in the practice tests (the feedback manipulation is more fully elaborated in the introduction to Experiment 4). In addition to now requiring participants to learn word associations rather than single items, these materials allow us to explore the grain size effect in a different type of memory test, cued recall rather than free recall. We compared interim test, end-test, and restudy groups.

Method

Participants

Participants were first year psychology students at UCL who participated as part of a laboratory class activity. The sample size was therefore limited to the number of students enrolled in the course. One hundred twenty-five students participated. Participants were allocated to conditions sequentially according to the order of enrolment. Five participants had incomplete data sets that were excluded from the analyses. This left a total of 120 participants (103 female, 14 male, one nonbinary, and two not reported, $M_{\text{age}} = 18.8$, $SD = 0.71$), with 41 in the interim test group, 39 in the end-test group, and 40 in the restudy group. A post hoc sensitivity analysis indicated that this sample was sufficient to detect an effect size of $d = 0.56$ with $\alpha = .05$ and power = .80 on a one-tailed independent samples t test.

Materials

Study materials were 36 Euskara–English translation word pairs (e.g, *hodei–cloud*; see Supplemental Materials for full word lists), divided into three sets. These sets were randomly allocated to study lists.

Procedure

Ethical approval was provided by the UCL Department of Experimental Psychology ethics committee. The experiment was programmed in PsychoPy and run online via Pavlovia. Prior to beginning the experiment, consent was obtained, and participants completed a demographic questionnaire. They studied three lists of 12 Euskara–English translation word pairs, self-paced. For each study list, each word pair was presented on screen until the participant clicked the continue button (or for a maximum of 15 s). After each list, participants undertook a 1-min distractor task in which they completed jigsaw puzzles on the screen. We switched from a numerical to a visuospatial distractor task to avoid any involvement of retrieval processes during the distractor phase.

Directly after the distractor task, participants in the restudy group were asked to restudy the previous list. In the interim test group, participants completed a test of the previous list. In this test, Euskara words were presented sequentially in random order, and participants were prompted to type in the corresponding translation of each one. There was no time limit on recall. Participants in the end-test group proceeded to the next list. However, after studying all three lists, participants in this group were tested on all items from all lists. Again, Euskara words were presented sequentially in random order, and participants were prompted to type in the English translations. All participants then undertook a 5-min jigsaw puzzle distractor task before completing an immediate test of all studied items. In the immediate test, each Euskara word was presented, and participants typed in the English translation. They were then asked to indicate whether or not they took notes or recordings to assist their learning or recall during the task. There was no delayed test in Experiment 3.

Results

Practice

There was significantly better recall across interim tests (M = 21.30, SD = 7.78) than in the end test (M = 14.15, SD = 8.26), $t(77) = 3.96$, $p < .001$, $d = 0.89$, $BF_{10} = 142.74$.

Immediate Test

There was better recall in the interim ($M = 17.38$, $SD = 8.76$) than end-test ($M = 14.08$, $SD = 8.27$) group, $t(77) = 1.72$, $p = .045$ one-tailed, $BF_{10} = 1.58$, $d = 0.39$. A two-tailed t test showed no significant difference between restudy ($M = 17.34$, $SD = 7.97$) and end-test groups, $t(78) = 1.80$, $p = .076$, $d = 0.40$, $BF_{01} = 1.07$. There was also no significant difference between the restudy and interim test groups, $t(79) = 0.02$, $p = .986$, $d = 0.004$, $BF_{01} = 4.33$.

Retention Across Tests

Analysis of the amount forgotten from the practice to immediate test (see Supplemental Materials) confirms the interaction that can be seen in Figure 6 indicating a larger decline in the interim test group. The interim test group forgot an average of 3.93 (SD = 2.77) translations, while the end-test group forgot an average of 0.08 (SD = 1.53).

Experiment 4

Experiment 3 generalizes the grain size effect to paired-associate learning, with better recall in the interim than the end-test group in the immediate test. The provision of feedback in Experiment 4 allows us to ask whether the effect generalizes in another important aspect: In educational settings, feedback is invariably provided in practice tests, so any applied relevance of the grain size effect rests on establishing that it extends to situations that include feedback. In addition, the provision of feedback permits us to assess the role of retrieval success to the grain size effect. If the effect is due to increased reexposure to materials during practice due to elevated retrieval success in the interim groups, we should see better recall in the immediate test and a reduced (or eliminated) grain size effect when feedback is provided.

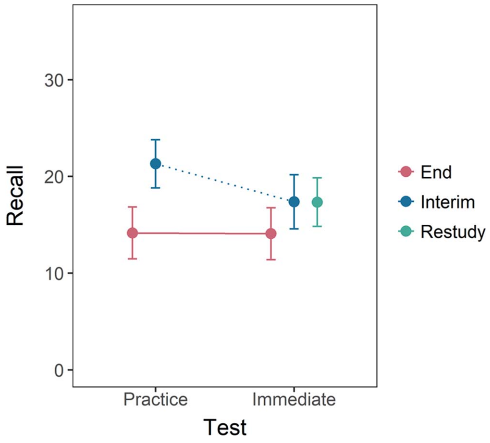
Method

Participants

118 undergraduate students at UCL participated as part of a laboratory class activity. Participants were allocated to conditions sequentially according to the order of enrolment. One had incomplete data, and one reported that they had made notes or recordings; hence, the data from these participants were excluded from the analyses. This left 116 participants (99 female, 16 male, one nonbinary, M_{age} = 18.4, SD = 1.8), with 36 in the interim test group, 40 in the end-test group, and 40 in the restudy group. A post hoc sensitivity analysis indicated that this sample was sufficient to detect an effect size of d = .58 with α = .05 and power = .80 on a one-tailed independent samples t test.

Figure 6

Mean Recall in Practice and Immediate Tests in Each Group in Experiment 3



Note. See the online article for the color version of this figure.

Materials

Experiment 4 was preregistered (Don et al., 2024). The 36 word pairs were identical to those used in Experiment 3.

Procedure

Ethical approval was obtained from UCL as before. The experiment was programmed in PsychoPy and run online via Pavlovia. The procedure was similar to that in Experiment 3; however, the 36 word pairs were studied in four lists of nine word pairs, and corrective feedback was provided in the interim and end tests. The change from three to four lists was aimed at increasing the magnitude of any FTE, which increases with number of interim tests (Chan et al., 2018). After typing and entering a translation, the correct word pair was presented in green if the typed translation was correct or in red if the typed translation was incorrect. Feedback was self-paced.

Results

Practice

The results are shown in Figure 7. According to the preregistration plan, a one-way ANOVA showed a significant effect of practice schedule, with better recall across the interim tests (M = 24.39, SD = 9.13) compared to the end test (M = 10.83, SD = 7.79), $F(1, 74) = 48.83$, $p < .001$, $\eta_p^2 = .398$, $BF_{10} = 7.93 \times 10^6$.[1]

Immediate Test

There was a significant grain size effect in the immediate test, with better recall in the interim test (M = 21.31, SD = 9.93) than end test (M = 15.53, SD = 9.06) group, $F(1, 74) = 7.04$, $p = .010$, $\eta_p^2 =$ .087, $BF_{10} = 4.62$.[2] A two-tailed t test showed no significant difference between restudy (M = 18.58, SD = 8.65) and end test, $F(1, 78) = 2.37$, $p = .128$, $\eta_p^2 = .029$, $BF_{01} = 1.55$.[3] There was also no significant difference between the restudy and interim test groups, $F(1, 74) = 1.64$, $p = .204$, $\eta_p^2 = .022$, $BF_{01} = 2.07$.[4]

Retention Across Tests

Analysis of the amount forgotten from the practice to immediate test (see Supplemental Materials) confirms the interaction that can be seen in Figure 7. Recall decreased for the interim test group, who forgot an average of 3.08 (SD = 3.13) translations but increased for the end-test group, who recalled an additional 4.70 (SD = 2.91) translations. This increase in recall in the end-test group was not observed in prior experiments and is likely due to the provision of feedback, as discussed below.

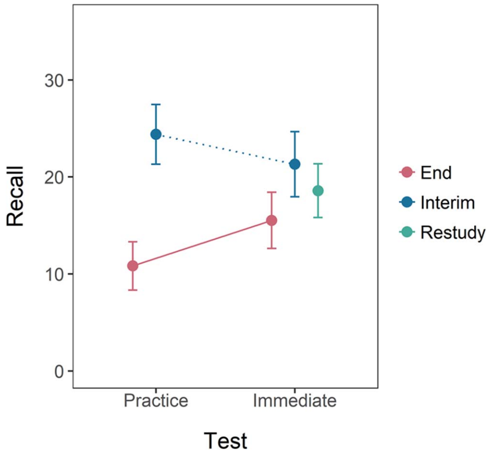
[1] Our preregistration plan specified ANOVAs; however, we also report t tests here for consistency with previous experiments; $t(1,74) = 7.00$, $p <$ .001, $d = 1.61$, $BF_{10} = 7.93 \times 10^6$.

[2] $t(74) = 2.65$, $p = .004$, $d = 0.61$, $BF_{10} = 4.62$.

[3] $t(78) = 1.54$, $p = .128$, $d = 0.34$, $BF_{01} = 1.55$.

[4] $t(74) = 1.28$, $p = .204$, $d = 0.29$, $BF_{01} = 2.07$.

Figure 7
Mean Recall in Practice and Immediate Tests in Each Group in Experiment 4



Note. Error bars reflect 95% confidence intervals. See the online article for the color version of this figure.

Discussion

Experiments 3 and 4 demonstrated better recall in interim than end tests in paired-associate learning, in both practice and, more importantly, immediate tests. Interestingly, we did not observe a testing effect in immediate test recall. This may be due to the short delay between study and immediate test (e.g., Roediger & Karpicke, 2006). Nevertheless, these studies provide further evidence for a benefit of interim over end tests in immediate recall, generalizing this grain size effect to a new type of material (paired associates) and to situations in which feedback is provided in the practice tests.

Despite a weaker grain size effect in Experiment 3 than we have previously observed in the word list experiments, this experiment demonstrates the same pattern of results as prior experiments: better practice performance in the interim than end test but better retention (or less forgetting) of items correctly recalled at practice in the end test than interim test condition in the immediate test. The increase in recall in the end-test group's immediate recall in Experiment 4 (M = 15.53) compared to practice (M = 10.83) is likely due to the provision of feedback during practice, as participants were reexposed to the correct translations for each item. This reexposure appears to only elevate recall in the end-test group but not the interim test group. This could be due to the shorter interval between practice and immediate tests in the end-test group or may be due to the nature of the test itself. The role of retrieval success in the grain size effect is further investigated in Experiment 5.

Experiment 5

The results from the previous four experiments suggest that interim testing does result in better performance in immediate tests in comparison to end tests. One potential reason for this could be increased retrieval success during the practice phase, which should induce greater reexposure to study materials and enhance long-term retention. The previous experiments have confirmed that aggregate recall across the interim practice tests is far superior (more than double as confirmed later) to that in an end-practice test. However, there appears to be a precipitous drop in retention from practice to immediate test. This could be due to a lack of desirable difficulty in the interim test group (Bjork, 1994), in which study items are relatively easy to recall in the practice tests due to the proximity to study and the smaller memory load (number of items to recall). Although this would result in good initial recall in practice, a lack of desirable difficulty would lead to shallow encoding and, therefore, poorer long-term retention, reducing the benefit of interim tests in immediate and delayed recall.

Experiment 5 assessed these explanations by manipulating the ease of practice tests. To achieve this, we included a three-letter word stem for each word. We used a 2 × 2 between-subjects design with grain size (interim test vs. end test) and stem (stem vs. no stem) varied orthogonally. The word stem should increase retrieval success in both interim and end conditions and decrease the difficulty of the practice test. If successful retrieval is important for long-term retention, including a stem in the practice tests should increase successful recall and carry over to superior immediate test recall (without stems). On the other hand, making the practice test easier might have the opposite effect of reducing desirable difficulty and, therefore, decrease retention even further in the immediate test.

Method

Experiment 5 was preregistered (Don et al., 2024).

Participants

Ethical approval was obtained from UCL as before. Sample size was determined by a power analysis using G*Power 3.1 (Faul et al., 2009), which indicated that a minimum total sample size of 70 participants per group was needed to detect a medium-sized effect (d = 0.50) with α = .05 and power = .90 in an independent samples t test. This effect size is based on the magnitude of the grain size effect in the previous experiments.

In total, 292 participants were recruited from Prolific, with the same eligibility requirements as Experiment 2, allocated randomly to conditions. Eight participants were excluded for reporting that they had taken notes during the experiment, and one was excluded for incomplete data, leaving 283 participants in Session 1 (71 per group in the interim-stem, interim-no stem, and end-stem groups and 70 in the end-no stem group; M_{age} = 36.13, SD = 10.76; 149 identified as female, 134 identified as male). Two hundred seventy-five participants completed Session 2. Following Session 1 exclusions, there were data for 266 participants in Session 2.

Materials

Four new lists of 16 words were created. Words were sourced from a word frequency list for the British National Corpus World Edition (https://www.kilgarriff.co.uk/bnc-readme.html). We took the 300 most frequent nouns between five and eight letters and selected 64 words that had unique three-letter stems (see Supplemental Materials). Words were randomly allocated to word lists. Word order within a list and list order were randomized.

Procedure

The experiment was programmed and run in Qualtrics. The procedure was identical to those for the interim and end-test groups in Experiment 2 with the following exception. In the stem conditions, each practice recall trial presented a list of response boxes, each accompanied by the first three letters of a studied word. Participants were prompted to type the remainder of the word, with each response remaining on the screen. In the no-stem conditions, each response box was blank with a prompt to recall studied words. No stems were provided in the immediate or delayed tests.

Results

Practice, immediate, and delayed recall in each test in all conditions are shown in Figure 8.

Practice Test

A 2 (grain size condition: interim vs. end) × 2 (stem condition: stem vs. no stem) between-subjects ANOVA revealed a significant main effect of study/test schedule, with better recall in the interim (M = 34.93, SD = 12.25) than end (M = 21.13, SD = 11.94) conditions, $F(1, 279) = 134.34$, $p < .001$, $\eta_p^2 = 0.325$, $BF_{10} = 8.66 \times 10^{15}$. There was also a main effect of stem, with better recall when a stem was provided ($M = 34.76$, $SD = 11.98$) than no stem ($M = 21.30$, $SD = 12.41$), $F(1, 279) = 127.86$, $p < .001$, $\eta_p^2 = 0.314$, $BF_{10} = 9.80 \times 10^{14}$. There was no significant interaction, $F(1, 279) = 0.12$, $p = .729$, $\eta_p^2 < .001$, $BF_{\text{excl}} = 5.10$.

Immediate Test

A 2 × 2 between-subjects ANOVA revealed a significant main effect of grain size, $F(1, 279) = 4.56$, $p = .034$, $\eta_p^2 = .016$, $BF_{10} = 1.05$, indicating better recall in the interim test ($M = 17.35$, $SD = 9.47$) than end-test ($M = 14.90$, $SD = 10.19$) conditions. There was no significant main effect of stem (which here refers to how the item was presented in the practice test), with no clear difference between stem ($M = 15.92$, $SD = 9.14$) and no stem ($M = 16.34$, $SD = 10.63$) conditions, $F(1, 279) = 0.12$, $p = .731$, $\eta_p^2 < .001$, $BF_{01} = 7.21$. However, this was qualified by a significant interaction between grain size and stem conditions, $F(1, 279) = 9.77$, $p = .002$, $\eta_p^2 = .034$, $BF_{10} = 15.77$. There was a significant grain size effect in the no-stem condition, $F(1, 139) = 12.42$, $p < .001$, $\eta_p^2 = .082$, $BF_{10} = 45.27$, but no equivalent effect in the stem condition, $F(1, 140) = 0.55$, $p = .459$, $\eta_p^2 = .004$, $BF_{01} = 4.31$. Figure 8 indicates that recall was better for stem than no-stem conditions in the end-test groups, but no stem was better than stem in the interim test groups. Analysis of simple effects indicated that the effect of stem was significant in the interim test conditions, $F(1, 140) = 6.58$, $p = .011$, $\eta_p^2 = .045$, $BF_{10} = 3.51$, but did not quite reach significance in the end-test conditions, $F(1, 139) = 3.56$, $p = .061$, $\eta_p^2 = .025$, $BF_{10} = 1.10$.

Delayed Test

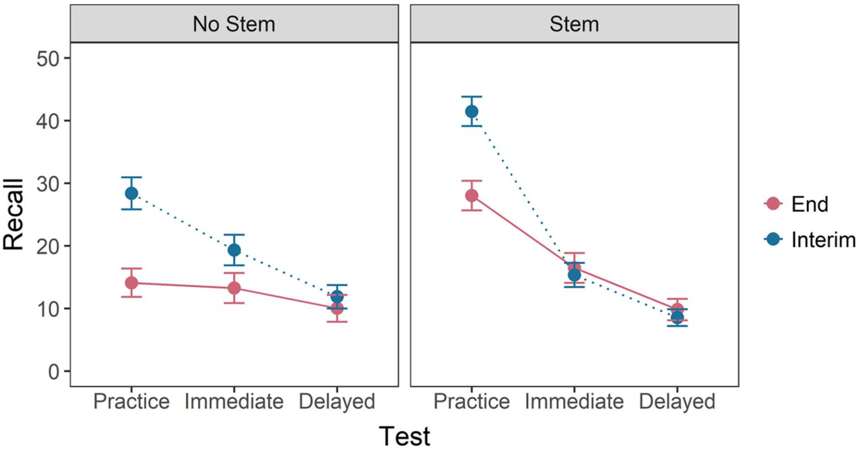
The delayed test was analyzed with a 2 × 2 ANOVA. There was no significant main effect of grain size condition (interim: $M = 10.27$, $SD = 6.95$; end: $M = 9.93$, $SD = 7.73$), $F(1, 262) = 0.10$, $p = .750$, $\eta_p^2 < .001$, $BF_{01} = 6.94$. Although recall was numerically better in the no-stem ($M = 11.03$, $SD = 8.10$) compared to the stem condition ($M = 9.22$, $SD = 6.42$), the main effect of stem did not reach significance, $F(1, 262) = 3.87$, $p = .050$, $\eta_p^2 = .015$, $BF_{01} = 1.07$, and there was no interaction, $F(1, 262) = 3.08$, $p = .081$, $\eta_p^2 = .012$, $BF_{\text{excl}} = 1.35$.

Retention Across Tests

Comparing practice and immediate tests (see Supplemental Materials for details) again found greater forgetting in the interim than the end-test group but also that this tendency was increased when stems were provided in the practice tests (as is evident in Figure 8). The interim stem group forgot an average of 26.13 ($SD = 9.46$) words, and the interim no-stem group forgot an average of 9.03 ($SD = 4.51$) words. The end stem group forgot an average of

Figure 8

Experiment 5: Mean Recall in Each Test for Interim and End-Test Conditions in the No-Stem (Left) and Stem (Right) Conditions



Note. See the online article for the color version of this figure.

11.55 (SD = 10.08) words, and the end no-stem group forgot an average of 0.83 (SD = 2.64) words.

An analysis of forgetting from the immediate to the delayed test (Figure 8; see Supplemental Materials for details) found that, while the differences were much smaller than between practice and immediate test, the provision of stems at practice lead to relatively more forgetting for the end group, whereas the interim group had similar levels of forgetting regardless of stem. The end stem group forgot an average of 6.83 (SD = 5.82) words compared to the end no-stem group who forgot an average of 3.85 (SD = 4.20) words. The interim stem group forgot an average of 7.05 (SD = 4.96) words, and the interim no-stem group forgot an average of 7.67 (SD = 6.70) words.

Discussion

Experiment 5 was a replication of the interim and end-test groups in Experiment 2, with the addition of a stem manipulation designed to assess whether increasing the ease of recall during practice resulted in enhanced or decreased retention in the immediate test. Experiment 5 replicated the major results, finding a benefit of interim tests on practice and immediate tests but which was no longer evident in a delayed test (Experiment 2 found the same result). Additionally, although we found that stems significantly improved practice recall for both interim and end tests to a similar extent, the impact on immediate recall was very different. Stems at practice improved immediate recall for the end-test condition but hindered immediate recall in the interim test condition, abolishing the grain size effect.

These results suggest that the benefits of retrieval success depend on test difficulty. Interim tests already improve retrieval success as a result of decreased memory load and short delay between study and recall, so increasing the ease of the tests even further may result in more superficial processing and poorer retention. Comparatively, end tests are more difficult, so increasing retrieval success through cues is beneficial to learning. This result therefore suggests that the poorer retention in the interim test groups in the immediate test could derive from a lack of desirable difficulty. Interim testing results in a higher likelihood of successful retrieval during the practice test, but given the relative ease of the test, provides only a short-term boost to retrieved items. Conversely, the comparable difficulty of end tests may provide a more durable memory boost to fewer items. This also provides an explanation for the selective improvement of the end-test group following practice test feedback in Experiment 4: Increased reexposure had a greater benefit for the more difficult end-test group compared to the relatively easier interim test group.

Although we have interpreted the effect of stems in terms of ease of recall, it must be acknowledged that there are other potential pathways by which they affected recall. For instance, the provision of stems disrupts the match between the interim and final test. Against this, Experiment 1 failed to obtain evidence that the grain size effect is sensitive to the match between practice and immediate test format. Another possibility is that participants engage in different encoding strategies when stems are and are not provided, given that their expectations of how they are going to be tested may be affected. It is not clear, however, how this could explain the finding that stems at practice led to improved immediate recall in one case (for the end-test condition) but worse recall in another (for the interim test condition).

Combined Analyses of Recall

Our results differ from the majority of previous research in reliably demonstrating an effect of schedule in both practice and immediate test recall, the latter constituting a grain size effect. In some conditions (related items in Experiment 1), this grain size effect persisted into a second delayed test. Nevertheless, the effect in immediate recall was much smaller than that in practice and smaller still in delayed tests. What is it that interim testing is benefiting? And where is the loss of long-term retention of these benefits coming from?

The following section attempts to examine these questions in more detail, specifically by decomposing the overall effect by list. This is potentially interesting as the only previous research that has found a grain size effect also revealed that the benefit of interim tests only emerged on later lists (Healy et al., 2017). Such a result suggests that the grain size effect might be partially caused by an FTE during encoding subsequent lists, such that each list is learned better after a preceding test. Experiment 2 indicated that our materials and methods produce a reliable FTE, and Chan et al.'s (2018) meta-analysis showed that the FTE might decay after longer retention intervals, which could explain why performance in the interim test conditions decreases over time. These results are consistent with the hypothesis that the FTE is a causal mechanism of the grain size effect.

One way of assessing this hypothesis is to analyze recall in each test by list. If the FTE is playing a role in the grain size effect, then the biggest difference in learning should be for later compared to earlier lists. In addition, if the reduction in the effect with time is due to decay of the FTE, then this benefit should be reduced with later tests. In this section, we report several analyses relevant to this issue, pooling data from the standard interim and end-test groups in the word recall experiments (Experiments 1 [whole format], 2, and the no stem groups in Experiment 5).

Recall by List

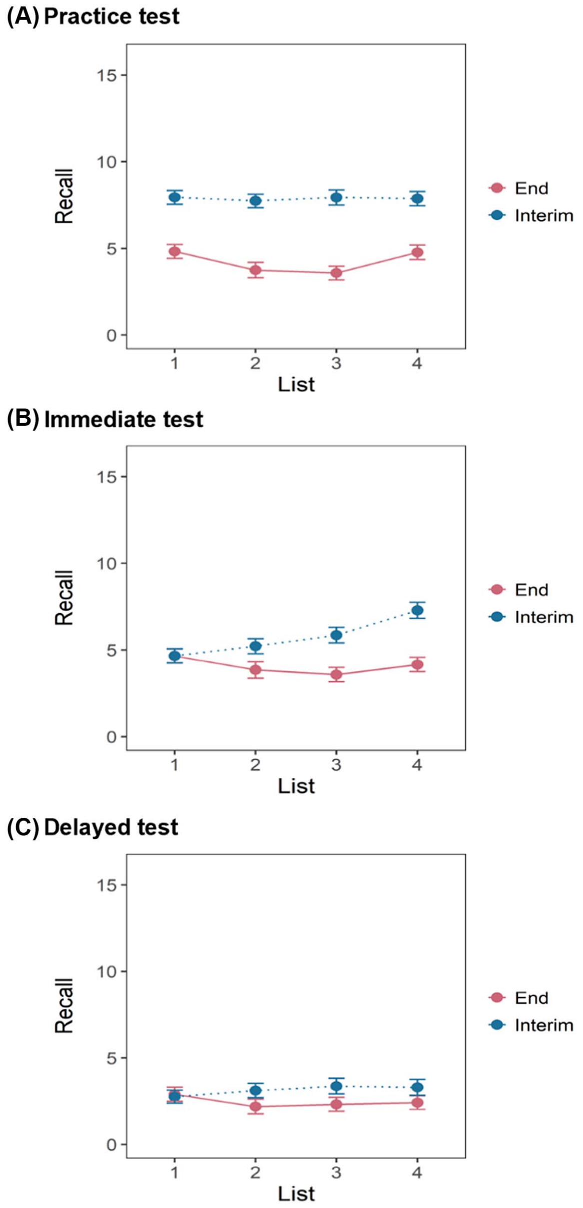
The numbers of items from each list recalled in practice, the immediate test, and the delayed test in the interim and end-test conditions are shown in Figure 9.

Practice Test

A 2 × 4 ANOVA with schedule (end vs. interim) and list (1–4) as factors indicated a significant main effect of schedule, with overall better aggregate recall in interim (M = 31.49, SD = 10.01) than end (M = 16.92, SD = 9.65) conditions (nearly double), $F(1, 422) = 232.51$, $p < .001$, $\eta_p^2 = .355$, $BF_{10} = 4.40 \times 10^{38}$. There was also a significant effect of list, $F(3, 1266) = 11.73$, $p < .001$, $\eta_p^2 = .027$, $BF_{10} = 11990.67$, and a significant interaction between list and schedule, $F(3, 1266) = 9.58$, $p < .001$, $\eta_p^2 = .022$, $BF_{10} = 2499.39$ (see Figure 9A). In the interim test group, recall did not change as a function of list, with no linear, $t(645) = 0.04$, $p = .970$, or quadratic, $t(645) = 0.52$, $p = .604$, trends. In the end-test group, polynomial contrasts indicated a significant quadratic trend, suggesting primacy and recency effects, $t(621) = 7.56$, $p < .001$.

Figure 9

Number of Items Recalled From Each List in Each Test Phase in the Interim and End-Test Conditions From Experiments 1, 2, and 5 Combined



Note. See the online article for the color version of this figure.

Immediate Test

A 2 × 4 ANOVA with schedule (end vs. interim) and list (1–4) as factors indicated a significant effect of schedule, with better aggregate recall in interim (M = 22.63, SD = 9.71) than end (M = 15.97, SD = 9.71) conditions, $F(1, 422) = 49.96$, $p < .001$, $\eta_p^2 = .106$, $BF_{10} = 1.08 \times 10^9$. This confirms the main finding of the individual experiments in revealing a robust grain size effect, with interim tests boosting immediate recall by over 40% compared to end tests (see Figure 9B).

There was also a significant effect of list, where recall was better for later than earlier lists, $F(3, 1266) = 24.25$, $p < .001$, $\eta_p^2 = .054$, $BF_{10} = 2.88 \times 10^{11}$, and this was qualified by a significant interaction, where the effect was greater for the interim than end-test group, $F(3, 1266) = 35.20$, $p < .001$, $\eta_p^2 = .077$, $BF_{\text{incl}} = 2499.39$. There was a linear and quadratic trend in both the interim test group, linear: $t(645) = 11.40$, $p < .001$; quadratic: $t(645) = 2.46$, $p = .014$, and the end-test group, linear: $t(621) = 2.52$, $p = .012$; quadratic: $t(621) = 4.84$, $p < .001$. This interaction confirms that the grain size effect is driven by superior later list learning in the interim test groups. Pairwise comparisons found no significant difference between end and interim tests for recall of List 1 words, interim: M = 4.58, SD = 2.97; end: M = 4.58, SD = 2.90, $t(422) = 0.02$, $p = .982$, $d = 0.002$, $BF_{01} = 9.29$, but a significant difference for List 4 words, interim: M = 7.13, SD = 3.37; end: M = 4.11, SD = 2.89, $t(422) = 9.91$, $p < .001$, $d = 0.96$, $BF_{10} = 6.49 \times 10^{17}$. Thus, under the present conditions, interim tests boost immediate recall overall but particularly for later lists consistent with the FTE. For the last of four lists, the boost is over 70%.

Delayed Test

The same analysis for the delayed test indicated a significant grain size effect, with better aggregate recall in interim (M = 12.47, SD = 9.06) than end (M = 9.72, SD = 8.17) conditions, $F(1, 321) = 7.82$, $p = .005$, $\eta_p^2 = .024$, $BF_{10} = 5.43$. Thus although the effect was not statistically significant in all experiments, overall the grain size effect persists across a delay and yields a boost compared to an end test of approximately 30%.

There was no significant main effect of list, $F(3, 963) = 0.99$, $p = .393$, $\eta_p^2 = .003$, $BF_{01} = 122.52$, but there was an interaction between list and grain size, where there was a greater increase in recall in later lists than earlier lists for the interim test group than the end-test group, $F(3, 963) = 7.24$, $p < .001$, $\eta_p^2 = .022$, $BF_{\text{incl}} = 121.55$ (see Figure 9C). In the interim test group, polynomial contrasts showed that there was a significant linear trend, $t(501) = 2.82$, $p = .005$. In the end-test group, there was both a linear trend, $t(462) = 2.15$, $p = .032$, and quadratic trend, $t(462) = 2.92$, $p = .004$. Pairwise comparisons showed no significant difference in end and interim test recall for List 1 words (interim: M = 2.80, SD = 2.48; end: M = 2.97, SD = 2.55), $t(321) = 0.59$, $p = .558$, $d = 0.07$, $BF_{01} = 6.91$, but a significant difference for List 4 words (interim: M = 3.34, SD = 3.02; end: M = 2.47, SD = 2.48), $t(321) = 2.83$, $p = .005$, $d = 0.32$, $BF_{10} = 5.52$.

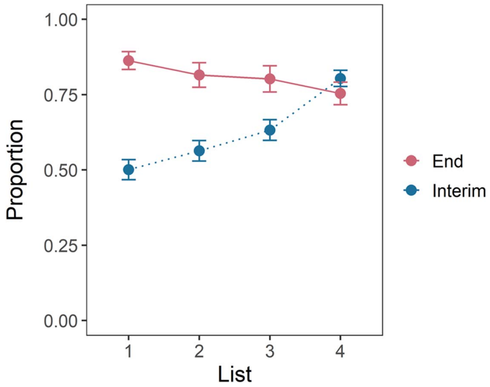
Practice test recall did not differ substantially by list in the interim test group, indicating a maintenance of recall performance across the task. Similar to Healy et al. (2017), immediate test recall was better for later lists than earlier lists for the interim test group, the benefit of interim testing over end tests increasing across lists. This benefit for later lists was also maintained in the delayed test, although to a lesser extent.

Retention of Previously Recalled Items

To examine what participants are retaining across test phases, we focused on the words recalled in the immediate test that had also been recalled in a practice test (meaning the end test or one of the interim tests). We examined this as a proportion of all words recalled in the practice test, as an index of retention (see Figure 10).

Figure 10

Proportion of Correctly Recalled Practice Test Words Retained in the Immediate Test From Experiments 1, 2, and 5 Combined



Note. See the online article for the color version of this figure.

Another 2 × 4 ANOVA with grain size (end vs. interim) and list (1–4) as factors found a main effect of grain size, $F(1, 366) = 132.86$, $p < .001$, $\eta_p^2 = .27$, $BF_{10} = 5.91 \times 10^{22}$, where the end-test group ($M = .81$, $SD = .17$) retained more words from the practice test than the interim test group ($M = .62$, $SD = .14$). There was also a main effect of list, $F(3, 1098) = 16.01$, $p < .001$, $\eta_p^2 = .04$, $BF_{10} = 5.49 \times 10^{11}$, arising from the fact that there was better retention for the later than earlier lists. Most interestingly, there was also an interaction between list and grain size, $F(3, 1098) = 49.71$, $p < .001$, $\eta_p^2 = .12$, $BF_{\text{incl}} = 1.98 \times 10^{27}$. This effect appears to be driven primarily by the interim test group in which there was a significant upward linear trend, $t(636) = 14.08$, $p < .001$, as well as quadratic trend, $t(636) = 3.63$, $p < .001$. In the end-test group, there was a significant downward linear trend, $t(462) = 3.35$, $p < .001$.

Thus items successfully recalled in the end test tended to remain recallable in the immediate test without much dependency on which list they originally appeared in (although, as Figure 10 shows, slightly more items proportionally were recalled from earlier lists). In contrast, many items successfully recalled in the early interim tests were not retained by the time of the immediate test. Both of these patterns seem reasonable given that the interval between the end and immediate tests was short, whereas the intervals were appreciably longer and filled with interfering learning and retrieval, in the case of the early lists with interim tests.

The same analysis for the delayed test (see Figure 11) showed a significant main effect of grain size, with more words retained in the end-test condition ($M = .50$, $SD = .24$) than interim test condition ($M = .34$, $SD = .19$), $F(1, 280) = 40.79$, $p < .001$, $\eta_p^2 = .13$, $BF_{10} = 1.21 \times 10^7$. There was no main effect of list, $F(3, 840) = 0.84$, $p = .471$, $\eta_p^2 = .003$, $BF_{01} = 244.39$. However, there was a significant interaction between grain size and list, $F(3, 840) = 4.82$, $p = .002$, $\eta_p^2 = .02$, $BF_{\text{incl}} = 1.98 \times 10^{27}$. Here, there was an upward linear effect in the interim test group, $t(478) = 1.95$, $p = .052$, but a significant downward linear trend in the end-test group, $t(342) = 2.45$, $p = .015$.

This benefit for later lists in immediate and delayed recall confirms a contribution of FTEs to the grain size effect, where encoding and retention of each successive list is facilitated by the preceding tests.

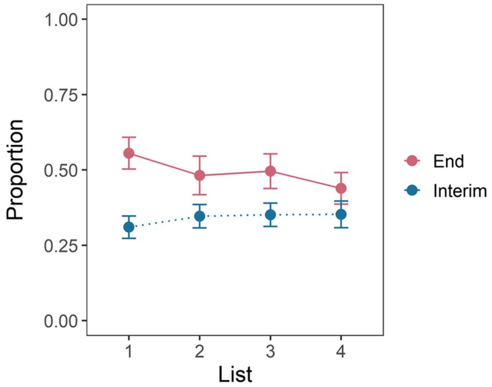
Meta-Analysis of Grain Size Experiments

We conducted a meta-analysis with the primary aim of estimating the magnitude of the grain size effect in immediate test performance across all available studies. As a secondary aim, we ran a comparable analysis to estimate the effect of study/test schedule in practice tests. Immediate test performance was calculated as the difference between the interim and end-test conditions in correct recall on the immediate test. Practice test performance was calculated as the difference in correct recall between the interim and end-test conditions on the practice tests (with recall across the interim tests being aggregated). Details of the study selection protocol and analysis method are provided in the Supplemental Materials.

All standardized differences were calculated based on Hedges' g (Hedges, 1981). We conducted our analysis on the data from the studies described in the Introduction section (five publications, one unpublished report, $k = 14$ effect sizes) together with our results for the standard interim and end-test conditions. Figure 12 provides a forest plot of these effects in immediate test performance, and an analogous plot for practice test performance is provided in the Supplemental Materials. With the full data set ($k = 19$), we found a significant and robust effect of schedule in practice tests, $g = 1.11$, 95% CI [0.87, 1.36], such that performance was better in interim than end tests, $t(17) = 9.57$, $p < .0001$. There was significant between-study heterogeneity, $\tau^2 = .19$, $Q = 85.91$, $p < .0001$, with a moderate–high percentage of variability not caused by sampling error ($I^2 = 80.2\%$). Most importantly, we also found a significant grain size effect in immediate tests, $g = 0.25$, 95% CI [0.09, 0.42], such that immediate test performance was better with interim than end tests, $t(18) = 3.24$, $p = .005$. There was significant between-study heterogeneity in this sample, $\tau^2 = .07$, $Q = 55.54$, $p < .0001$, with a moderate–high percentage of variability not caused by sampling error ($I^2 = 67.6\%$). The magnitude of the effect is small to medium, but it is quite robust: 15/19 experiments observed a numerical benefit of interim tests over end tests.

Figure 11

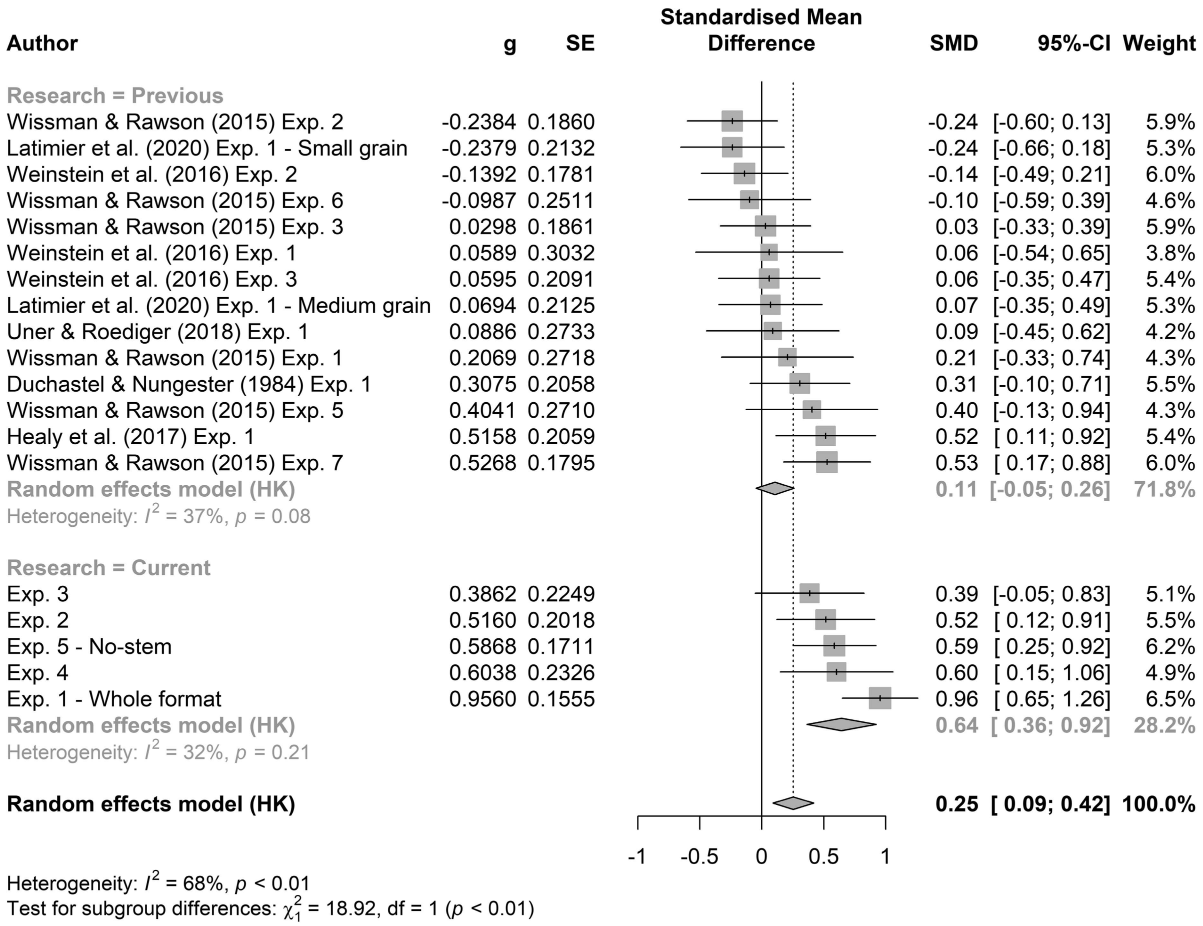
Proportion of Correctly Recalled Practice Test Words Retained in the Delayed Test From Experiments 1, 2, and 5 Combined



Note. See the online article for the color version of this figure.

Figure 12

Forest Plot of Effect Sizes for the Grain Size Effect in Immediate Tests in Previous Research and the Current Experiments 1–5



Author	g	SE	SMD	95%-CI	Weight
Research = Previous					
Wissman & Rawson (2015) Exp. 2	-0.2384	0.1860	-0.24	[-0.60; 0.13]	5.9%
Latimier et al. (2020) Exp. 1 - Small grain	-0.2379	0.2132	-0.24	[-0.66; 0.18]	5.3%
Weinstein et al. (2016) Exp. 2	-0.1392	0.1781	-0.14	[-0.49; 0.21]	6.0%
Wissman & Rawson (2015) Exp. 6	-0.0987	0.2511	-0.10	[-0.59; 0.39]	4.6%
Wissman & Rawson (2015) Exp. 3	0.0298	0.1861	0.03	[-0.33; 0.39]	5.9%
Weinstein et al. (2016) Exp. 1	0.0589	0.3032	0.06	[-0.54; 0.65]	3.8%
Weinstein et al. (2016) Exp. 3	0.0595	0.2091	0.06	[-0.35; 0.47]	5.4%
Latimier et al. (2020) Exp. 1 - Medium grain	0.0694	0.2125	0.07	[-0.35; 0.49]	5.3%
Uner & Roediger (2018) Exp. 1	0.0886	0.2733	0.09	[-0.45; 0.62]	4.2%
Wissman & Rawson (2015) Exp. 1	0.2069	0.2718	0.21	[-0.33; 0.74]	4.3%
Duchastel & Nungester (1984) Exp. 1	0.3075	0.2058	0.31	[-0.10; 0.71]	5.5%
Wissman & Rawson (2015) Exp. 5	0.4041	0.2710	0.40	[-0.13; 0.94]	4.3%
Healy et al. (2017) Exp. 1	0.5158	0.2059	0.52	[0.11; 0.92]	5.4%
Wissman & Rawson (2015) Exp. 7	0.5268	0.1795	0.53	[0.17; 0.88]	6.0%
Random effects model (HK)			**0.11**	**[-0.05; 0.26]**	**71.8%**
Heterogeneity: I^2 = 37%, p = 0.08					
Research = Current					
Exp. 3	0.3862	0.2249	0.39	[-0.05; 0.83]	5.1%
Exp. 2	0.5160	0.2018	0.52	[0.12; 0.91]	5.5%
Exp. 5 - No-stem	0.5868	0.1711	0.59	[0.25; 0.92]	6.2%
Exp. 4	0.6038	0.2326	0.60	[0.15; 1.06]	4.9%
Exp. 1 - Whole format	0.9560	0.1555	0.96	[0.65; 1.26]	6.5%
Random effects model (HK)			**0.64**	**[0.36; 0.92]**	**28.2%**
Heterogeneity: I^2 = 32%, p = 0.21					
Random effects model (HK)			**0.25**	**[0.09; 0.42]**	**100.0%**

Heterogeneity: I^2 = 68%, p < 0.01
Test for subgroup differences: χ^2_1 = 18.92, df = 1 (p < 0.01)

Note. Within each subset, studies are ordered by increasing size of the effect. Meta-analytic estimates for each subset and the combined data are also included. CI = confidence interval; Exp. = experiment; HK = Hartung–Knapp.

There was a small and nonsignificant grain size effect in previous studies, $g = 0.11$, 95% CI [−0.05, 0.26], and a significantly ($p < .01$) larger effect in our own experiments, g = 0.64, 95% CI [−0.36, 0.92]. In the Introduction section, we suggested that previous studies may have failed to obtain stronger evidence because of their use of complex text materials, insufficient demands on effortful retrieval in the practice tests, or employing experimental conditions that were not conducive to obtaining an FTE. The meta-analysis does not permit these (or other) possibilities to be evaluated, but it is nevertheless clear from the results that the methods employed in our experiments evoked a much stronger grain size effect on immediate recall. Also noteworthy is that, in future studies, if the meta-analytic effect size is the only available basis for a power calculation, then researchers will need to plan to test approximately 200 participants per group to achieve 80% power to detect a true grain size effect. On the other hand, if list materials are employed and the meta-analytic effect size from the present studies is a justifiable basis for a comparable power calculation, much smaller samples of 30 per group will be sufficient.

General Discussion

The grain size of recall hypothesis states that testing smaller chunks of information interspersed throughout learning should lead to better long-term retention than testing larger chunks at the end of learning. Previous research has, on the surface, failed to demonstrate a reliable grain size effect in immediate test recall, despite obvious benefits for retrieval success during practice (e.g., Wissman & Rawson, 2015). Our meta-analysis of prior research failed to demonstrate a grain size effect in immediate recall, although the studies had small sample sizes and the estimated effect could have been as large as $g = 0.26$ (the upper 95% CI). Over five experiments, we demonstrated a clear grain size effect in immediate tests. This was observed for both free recall of word lists (Experiments 1, 2, and 5) and cued recall of paired associates (Experiments 3 and 4) and both with (Experiment 4) and without (Experiments 1–3 and 5) corrective feedback in practice tests. In Experiment 1, this effect persisted into a delayed test after a 1-week retention interval, and we also observed a grain size effect in the delayed test when collapsing data from all word list experiments. Finally, a meta-analysis of all available studies ($k = 19$) revealed a significant and robust grain size effect in immediate tests, $g = 0.25$, 95% CI [0.09, 0.42] of small to medium effect size magnitude.

Experiment 1 suggested that the grain size effect may be specific to tests, as there was no benefit of smaller grain sizes on restudy. The effect was also not dependent on the format of the immediate test matching that in interim tests, with a robust grain size effect being observed in the whole-text format. Note that, here, we refer to test format as retrieval according to a specific list or free recall of all items. Future research is needed to assess whether the grain size effect also transfers to tests of very different formats compared to practice (e.g., free recall vs. recognition, short answer vs. multiple choice).

Potential Causes of the Grain Size Effect

One potential explanation for the stronger grain size effect in immediate recall we observed, compared to previous studies, could be the use of simple rather than complex materials. The only other study to find a significant grain size effect used simple unrelated facts as study materials (Healy et al., 2017). Here, we see a clear grain size effect using word lists and paired associates. One hypothesis for the absence of the grain size effect in previous research is that interim tests interfere with the formation of an integrated whole-text representation of complex related materials (Duchastel & Nungester, 1984; Healy et al., 2017; Latimier et al., 2020). On the other hand, a comparison of the effect sizes in Experiments 1 and 2 suggests that relatedness may not be an important factor in simple materials. This could suggest that interference alone is unlikely to be the sole cause for the absence of the effect in previous research. However, this is based on a between-experiments comparison, and moreover, breaks in learning may not interfere with learning words of the same category to the same extent as they do for learning related complex text passages. High-powered replications are therefore needed to compare the grain size effect for related and unrelated complex materials to adequately test this hypothesis.

Another possible cause of the grain size effect in our experiments might be the FTE. As discussed, the FTE is the finding that interim tests boost learning of subsequent materials (Chan et al., 2018). In Experiment 2, we showed that our methods and materials produced a robust FTE in the interim test condition relative to an exposure-matched control. Our analyses, combining results across experiments, also suggested that the FTE may be critical in producing a grain size effect. Across experiments, the interim test group showed a maintenance of test performance across lists during practice and better retention of items from later lists than earlier lists in the immediate test. This was also seen to a lesser but still significant extent in the delayed test. Similar results were observed by Healy et al. (2017), who found that the benefit of interim testing only emerged for later lists.

This suggests that the grain size effect may in fact be primarily driven by a later list encoding boost. In addition, the reduction in the size of the benefit for the delayed test is consistent with the finding that the benefits of the FTE tend to be short-lived. A recent meta-analysis showed that the FTE reduces over time (Chan et al., 2018), although few studies have involved a delay longer than 24 hr. In addition, the FTE might be a motivational phenomenon (Yang et al., 2018) and therefore might be more sensitive to changes in state, such as time delays. However, how interim testing potentiates new learning is unclear (Chan et al., 2018; Yang et al., 2018). Healy et al. (2017) proposed that this might be due to interim tests sustaining motivation across trials and based this hypothesis on both self-reported effort and on the finding that an interim (compared to end) test enhanced recall for unquizzed and quizzed facts. Interestingly, this latter finding has a parallel in the FTE: Interim testing (compared to interim restudying) boosts later recall of both tested and untested items (Don et al., 2023). This is consistent with the idea that interim tests sustain motivation for later learning. The grain size and FTEs are not due to item-specific processes—on this account—but to something more general such as motivation. An implication of this view is that models of the testing effect, which emphasize item-specific processes (e.g., Hopper & Huber, 2018), are unlikely to be readily extended to explain grain size effects, but more research is needed.

Whether the lack of a robust grain size effect in previous research is due to the absence of an FTE is also unclear. Most previous research has not examined whether their materials and methods produce a reliable and robust FTE, and future research should aim to address this gap. One immediate consideration regarding the role of the FTE in the grain size effect is that it suggests that within-subjects designs (such as those adopted by Uner & Roediger, 2018; Weinstein et al., 2016) might wrongly conclude that there is no grain size effect, as all participants will benefit from a forward testing benefit.

A final explanation of the grain size effect in our studies could be a difference in reexposure to materials caused by increased retrieval success during practice in the interim group. In that sense, the provision of corrective feedback during practice provides insight by equating reexposure. Only one experiment in the present study included feedback (Experiment 4), yet it still obtained a significant grain size effect. However, immediate test performance in the end-test group was improved relative to practice. Immediate test performance still decreased relative to practice in the interim test condition, although we cannot determine if this is to a lesser extent when no feedback is provided. In addition, interim restudy led to poorer recall than interim tests in Experiment 1. Based on these results, it appears unlikely that the grain size effect is simply due to reexposure to study materials alone. However, this is an underexplored area of research, and future experiments are needed to compare the magnitude of the grain size effect with and without feedback in a single experiment.

Boundary Conditions on the Grain Size Effect

An important detail relevant to the pedagogical application of this effect is that, despite observing a significant grain size effect overall, there was notably poorer relative retention in the immediate test following interim compared to end tests (Figures 10 and 11). Although recall levels were lower in practice for end-test conditions, this level of recall was retained at a higher relative level in the immediate test, while the interim test conditions showed substantial amounts of forgetting. We are, therefore, observing a boost in retrieval success that does not transfer well to longer term recall in immediate tests (and even more poorly after a substantial delay).

It is difficult to avoid the potential for item selection effects in these circumstances. As the interval between study and test was longer in the end-test group, and fewer items were recalled in the end test than interim tests, it could be the case that the items recalled in the end test were easier to remember than those recalled in the interim test and, therefore, are more likely to also be recalled in the immediate test. However, the results from Experiment 5 suggest that this is not the only explanation.

Experiment 5 implies that part of this loss is due to a lack of *desirable difficulty* (Bjork, 1994). Because of the short lag between study and retrieval in the interim test condition and smaller memory loads of items to be recalled, retrieval success is high in practice tests. However, this may also reduce the difficulty of the tests to an extent that leads to shallower encoding. Indeed, Experiment 5 showed that introducing a word stem to assist recall in interim and end tests enhanced retrieval success in the practice tests in both groups but selectively reduced immediate test recall in the interim test groups only. As the interim test was already relatively easy, increasing the ease of recall even further increased forgetting of successfully recalled items. In comparison, as the end test was already more cognitively demanding, with higher memory load of items to be recalled, and greater lag between study and recall of most items, improving retrieval success instead served to benefit immediate test recall.

Thus, retrieval success is important for boosting long-term recall, but only when the tests are sufficiently difficult. Similar results have been observed in other educationally relevant domains, such as attempts to improve relational learning. For instance, trial sequences that increase training performance are only beneficial for transfer of relational rules if the rule is sufficiently difficult (Don et al., 2020). These results suggest that for optimal learning, the interim test needs to be sufficiently difficult to engage deep encoding and to provide a lasting advantage and avoid large amounts of forgetting.

A related issue concerns the possible moderating role of list length, retrieval success, and lag to the final test. When practice retrieval is successful, long lags tend to yield larger testing gains on a later test (see Rowland, 2014). Thus there may be conditions (e.g., very short lists and, hence, easy practice recall in the interim test group combined with a long lag) where the grain size effect is eliminated or even reversed: despite poorer recall in the practice end test, the benefit for those items that are successfully recalled may be sufficient at a long lag to match or exceed the testing benefits in the interim test group. Future research is needed to explore a wider range of combinations of these factors than has been possible here.

Optimizing the Grain Size Effect for Learning

Previous research has shown that interim tests that are only partial (only testing some of the studied information) and distributed (testing items from prior lists as well as the immediately preceding list) provide just as much benefit for FTEs as testing all studied material and only studying material from the preceding list (Don et al., 2023). It would be interesting to determine whether this is also true for the grain size effect. Considering desirable difficulty may be necessary for long-lasting grain size effects, distributed tests in particular may increase the memory load requirement by requiring recall of items further in the past. This may provide both a forward test effect benefit and a more durable grain size effect.

It is well-established that tests provide a benefit to learning over restudy (see Rowland, 2014). Experiment 1 demonstrated a significant testing effect in delayed tests; however, in the immediate test, there was only a significant advantage of interim tests over interim restudy, and no advantage of end tests over end restudy. This suggests further benefits of using interim over end tests; however, it is interesting to consider why we observed no test benefit with a large grain size. Under short retention intervals, tests can be of lesser benefit, no benefit, or even be detrimental relative to restudy (e.g., Roediger & Karpicke, 2006). Although the retention interval between study and immediate tests is the same in all conditions in these studies, there was a smaller interval between the initial test and immediate test in the end-test group, which may interfere with observing a testing effect. While prior research has shown that repeated testing of the same material at short intervals has a hypermnesic effect (Wheeler & Roediger, 1992; Roediger & Challis, 1989), the optimal timing of successive tests should be considered to provide the most benefit over restudying.

Several studies have examined metacognitive awareness of the benefits of testing and self-regulated use of testing versus restudy strategies. While metacognitive awareness of the benefits of retrieval practice is sometimes poor (see Rivers, 2021, for a review), use of retrieval practice is often preferred when the likelihood of retrieval success is high, (i.e., when the tests or study materials are easy or the material has been studied recently; e.g., Persky, 2018; Toppino et al., 2018; Tullis et al., 2018; Vaughn & Kornell, 2019). In addition, prior research has shown that those who experience a benefit of testing effects are more likely to choose retrieval practice as a learning strategy in a new learning phase (Hui et al., 2021). Therefore, interim tests should be more likely than end tests to provide conditions that encourage self-regulated use of testing strategies. Nevertheless, further research is needed to investigate metacognitive awareness and control of grain size effects. That is, are students aware of the benefits of smaller grain sizes, and are they more likely to use tests in this way to benefit their learning?

To summarize, interspersing tests on smaller chunks of studied material throughout study leads to better recall in both practice and immediate tests than testing all learned information at the end of the study phase, a pattern that was consistently replicated across five experiments. There was also some evidence that this benefit persists to a degree in delayed tests. This study therefore provides the clearest evidence to date of a grain size effect in immediate tests. Higher levels of retrieval success in practice tests contribute to the grain size effect, but the effect is eliminated if these tests are too easy. Furthermore, the FTE, where interim tests facilitate subsequent learning, may be a contributing cause of the grain size effect.

Constraints on Generality Statement

Participants included both undergraduate students participating in classroom tutorials and the general population recruited online. Experiments 1, 3, and 4 were run with undergraduate students, a relatively homogenous sample. Therefore, there may be some constraints on the generalization of these results based on age and education level. Experiments 2 and 5 were limited to participants aged 18–60 but with no constraints on education level.

The experiments used laboratory study materials of word lists and word pairs. Prior research has used more complex and real-world materials but has generally found weaker effects. It is of theoretical and practical interest whether the results generalize to more complex study materials. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

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). MIT Press. https://doi.org/10.7551/mitpress/4561.003.0011

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chan, J. C. K., Davis, S. D., Yurtsever, A., & Myers, S. J. (2024). The magnitude of the testing effect is independent of retrieval practice performance. *Journal of Experimental Psychology: General*, *153*(7), 1816–1837. https://doi.org/10.1037/xge0001593

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Don, H., Boustani, S., & Shanks, D. (2024, April). *Grain size effects in retrieval practice*. https://osf.io/8wuny

Don, H. J., Goldwater, M. B., Greenaway, J. K., Hutchings, R., & Livesey, E. J. (2020). Relational rule discovery in complex discrimination learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(10), 1807–1827. https://doi.org/10.1037/xlm0000848

Don, H. J., Yang, C., Boustani, S., & Shanks, D. R. (2023). Do partial and distributed tests enhance new learning? *Journal of Experimental Psychology: Applied*, *29*(2), 358–373. https://doi.org/10.1037/xap0000440

*Duchastel, P. C., & Nungester, R. J. (1984). Adjunct question effects with review. *Contemporary Educational Psychology*, *9*(2), 97–103. https://doi.org/10.1016/0361-476X(84)90012-2

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

*Healy, A. F., Jones, M., Lalchandani, L. A., & Tack, L. A. (2017). Timing of quizzes during learning: Effects on motivation and retention. *Journal of Experimental Psychology: Applied*, *23*(2), 128–137. https://doi.org/10.1037/xap0000123

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, *102*, 1–15. https://doi.org/10.1016/j.jml.2018.04.005

Hui, L., de Bruin, A. B., Donkers, J., & van Merriënboer, J. J. (2021). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review*, *33*(4), 1835–1857. https://doi.org/10.1007/s10648-021-09604-x

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. https://doi.org/10.1037/xap0000087

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Kole, J. A., Healy, A. F., & Bourne, L. E., Jr. (2008). Cognitive complications moderate the speed–accuracy tradeoff in data entry: A cognitive antidote to inhibition. *Applied Cognitive Psychology*, *22*(7), 917–937. https://doi.org/10.1002/acp.1401

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

*Latimier, A., Riegert, A., Ly, S. T., & Ramus, F. (2020). *Retrieval practice promotes long-term retention irrespective of the placement*. PsyArXiv. https://doi.org/10.31234/osf.io/dk63q

*Lavigne, E., & Risko, E. F. (2018). Optimizing the use of interpolated tests: The influence of interpolated test lag. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 211–221. https://doi.org/10.1037/stl0000118

Pashler, H. (2000). Task switching and multitask performance. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII: Control of mental processes*. MIT Press.

Persky, A. M. (2018). A four year longitudinal study of student learning strategies. *Currents in Pharmacy Teaching & Learning*, *10*(11), 1496–1500. https://doi.org/10.1016/j.cptl.2018.08.012

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

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, *33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Roediger, H. L., III, & Challis, B. H. (1989). Hypermnesia: Improvements in recall with repeated testing. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree symposium on cognition* (pp. 175–199). Lawrence Erlbaum Associates.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Shanks, D., Don, H., Boustani, S., & Yang, C. (2023). Test-enhanced learning. *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.908

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, *3*(3), 161–164. https://doi.org/10.1016/j.jarmac.2014.02.001

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, *46*(7), 1164–1177. https://doi.org/10.3758/s13421-018-0828-2

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 540–552. https://doi.org/10.1037/xlm0000473

Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society*, *12*(2), 89–91. https://doi.org/10.3758/BF03329636

*Uner, O., & Roediger, H. L. (2018). The effect of question placement on learning from textbook chapters. *Journal of Applied Research in Memory and Cognition*, *7*(1), 116–122. https://doi.org/10.1016/j.jarmac.2017.09.002

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*, 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), Article 35. https://doi.org/10.1186/s41235-019-0187-y

*Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*(1), 72–84. https://doi.org/10.1037/xap0000071

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240–246. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

*Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 439–455. https://doi.org/10.1037/xlm0000047

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning*, *3*(1), Article 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(8), 1127–1143. https://doi.org/10.1037/xlm0001021

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Received October 5, 2023
Revision received April 23, 2024
Accepted May 7, 2024 ■