Journal of Memory and Language 144 (2025) 104664

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Individual differences in working memory and the benefit of retrieval practice☆

Andy L. Fordyce *, Thomas S. Redick, Joseph P. Bedwell, Jeffrey D. Karpicke

Department of Psychological Sciences, Purdue University, USA

ARTICLE INFO

Keywords:
Retrieval practice
Working memory
Individual differences

ABSTRACT

Previous research on the association between individual differences in working memory and the benefit of retrieval practice has yielded mixed results, with various studies showing no differential retrieval practice benefit as a function of working memory ability, and others finding either more or less retrieval practice benefit for individuals lower in working memory. The current studies addressed how (a) variations in the learning task procedure and (b) measurement of working memory might influence the presence and/or strength of the relationship between working memory and retrieval practice. To ensure high initial retrieval success in Experiments 1 and 2, we used a learning-to-criterion procedure which had not been used in previous retrieval practice studies that examined individual differences. Experiment 3 extended the results of Experiments 1 and 2 to different learning task materials, while attempting to replicate a previous study that had shown a specific retrieval practice benefit for individuals with lower working memory. Additionally, separate analyses were conducted using partial and absolute scoring methods for operation span to address variability in previous research. Across all three experiments, retrieval practice outperformed restudying, and this benefit held regardless of individual differences in working memory ability.

Individual differences in working memory and the benefit of retrieval practice

Repeated retrieval is an effective tool for enhancing learning and long-term retention. This benefit has been observed in numerous studies (for review, see Dunlosky et al., 2013; Karpicke, 2017; or McDermott, 2021). However, few studies have examined the role of individual differences in this retrieval practice benefit. Of particular interest to the current study, past research has reported working memory (WM) as a significant predictor of various measures of academic success such as reading comprehension, fluid intelligence, and SAT scores (for review, see Unsworth & Redick, 2017). However, limited evidence has been provided for its effect on retrieval practice. In addition, past research investigating the relationship between WM and retrieval practice has yielded mixed results. The current research sought to clarify this relationship in multiple, large-sample experiments, and to examine the influence of variations in task procedure and measurement.

Retrieval practice

In a typical within-subjects retrieval practice experiment, subjects begin by learning a set of to-be-remembered items, such as word pairs or scientific facts. They are then asked to complete a learning activity that requires them to actively retrieve some of these items. Subjects are also asked to complete a learning activity that does not require retrieval, such as restudying other items. Following these activities, subjects are given a final assessment involving retrieval of all to-be-remembered items.

One methodological concern in retrieval practice studies is how to examine the direct effects of retrieval without the influence of potentially confounding or mediating effects. For example, providing feedback following retrieval may influence how individuals study and engage with the material on subsequent trials. However, in the absence of feedback, there may be differences in re-exposure to the materials between repeated study and repeated testing conditions. During learning, subjects are not typically successful in recalling all to-be-remembered items during test trials, and will therefore only be re-

☆ This article is part of a special issue entitled: 'Individual differences in memory' published in Journal of Memory and Language.

* Corresponding author at: Department of Psychological Sciences, 703 Third Street, West Lafayette, IN 47907, USA.
E-mail address: afordyc@purdue.edu (A.L. Fordyce).

https://doi.org/10.1016/j.jml.2025.104664
Received 11 January 2024; Received in revised form 7 May 2025; Accepted 7 June 2025
Available online 12 June 2025

exposed to the items they successfully recalled. In contrast, subjects are re-exposed to all items during study trials. This creates a bias toward study trials and may influence the magnitude or presence of a retrieval practice effect (Karpicke, 2017; Vaughn & Rawson, 2011). This issue is also particularly relevant to individual differences research conducted with groups that may vary in initial retrieval success. Groups with higher levels of initial retrieval success will be exposed to more items during retrieval trials than groups with lower levels of initial retrieval success. Indeed, prior retrieval practice studies have shown evidence that individuals with higher cognitive abilities (e.g., WM, fluid intelligence) have greater initial retrieval success (Agarwal et al., 2017; Minear et al., 2018). One proposed way to alleviate this exposure bias is to create conditions that afford a high level of initial retrieval success during the learning phase. To accomplish this, several studies have used a learning-to-criterion procedure (Grimaldi & Karpicke, 2014; Karpicke, 2009; Karpicke & Roediger, 2007, 2008; Pyc & Rawson, 2009). In this procedure, prior to repeated study and repeated retrieval manipulations, subjects are required to learn all items to the criterion of one correct recall per item. This procedure improves initial retrieval success in the following learning activity phase. By ensuring high initial retrieval success, without making retrieval trivially easy, the concern of differential re-exposure to the items is mitigated and the benefit solely due to retrieval practice can be better isolated.

Individual differences in working memory

Though the benefit of retrieval practice has been clearly demonstrated across several studies, fewer studies have examined the role of individual differences in the benefit students gain from utilizing this strategy. However, the role of individual differences in WM has been demonstrated as an important learner characteristic in relation to higher-order cognitive functioning. For example, individual differences in WM are strongly predictive of fluid intelligence and reading comprehension, as well as measures of academic performance like SAT scores (Engle & Kane, 2004). Additionally, WM has been implicated in the ability to conduct a cue-driven controlled search of long-term memory (Unsworth & Engle, 2007b) and engage attentional control processes necessary for effortful retrieval and inhibition of distractions when retrieving information (Kane & Engle, 2000). Indeed, WM has been shown to be positively correlated with long-term memory performance, and in free and cued recall tasks, high WM individuals tend to recall more items, make fewer intrusions, and recall items faster than low WM individuals (Unsworth, 2019). Previous research has suggested that high WM individuals are better at narrowing their search set during recall and reducing proactive interference (e.g., Kane & Engle, 2000; Unsworth, 2019), both of which are cognitive processes that have also been implicated in retrieval practice (e.g., Lehman et al., 2014; Szpunar et al., 2008). Due to these reasons, prior research (e.g., Brewer & Unsworth, 2012; Minear et al., 2018) has proposed that WM ability may influence the magnitude of the retrieval practice effect observed for students of different ability ranges (akin to an aptitude-by-treatment interaction; Cronbach & Snow, 1977). This may have important implications for the use of retrieval practice in applied settings because it could either exacerbate or reduce the gap in learning across students at varying levels of WM ability. Alternative study strategies may need to be considered for certain populations.

However, the direction of this WM influence remains unclear. Various explanations have been proposed to support a greater benefit of retrieval practice for low ability students, for high ability students, as well as an equal benefit for low and high ability students. Evidence for all of these outcomes have been found in various studies. The rationale for each outcome is as follows:

1. *Greater benefit for high WM students.* High WM students can more effectively utilize cognitive processes and abilities, such as attentional control or use of effective retrieval cues, relative to low WM students. Consistent with the "rich-get-richer" hypothesis, this enables high ability students to reap more benefits from retrieval practice.
2. *Greater benefit for low WM students.* Repeated retrieval may facilitate more effective use of cognitive processes and abilities that are not normally utilized by low WM students. In contrast, high WM students already optimally use these cognitive abilities, so repeated retrieval does not grant them additional benefit.
3. *Equal benefit for low and high WM students.* The benefit of retrieval practice is independent of WM ability. For all students, learning will be enhanced equally following retrieval practice.

Previous research

Of the studies that have examined the relationship between WM ability and retrieval practice, results have been mixed, with some reporting a greater benefit of retrieval practice for low ability students (Agarwal et al., 2017; Yang et al., 2020), others reporting a greater benefit for high ability students (Harrison et al., 2012; Tse & Pu, 2012), and others reporting an equal benefit for low and high ability students (Bertilsson et al., 2021; Brewer & Unsworth, 2012; Jonsson et al., 2021; Minear et al., 2018; Wiklund-Hornqvist et al., 2014).

Synthesizing the results across these studies proves difficult due to differences in task, materials, procedure, scoring, and sample (see Table 1 for summary). Regarding task, various studies have used a paired-associates task to measure retrieval practice (Bertilsson et al., 2021; Brewer & Unsworth, 2012; Jonsson et al., 2021; Minear et al., 2018; Tse & Pu, 2012), though there has been variation in procedure for factors such as retention interval, feedback presence, the number of study-test opportunities, and between- versus within- subject manipulations. Additionally, other studies have used alternative retrieval practice tasks and materials such as general knowledge questions (Agarwal et al., 2017), passage reading (Harrison et al., 2012), lecture learning (Wiklund-Hornqvist et al., 2014), and multilist free recall (Yang et al., 2020).

Another source of variation across studies has been the scoring method utilized for the WM span tasks. Although all studies previously mentioned administered the operation span task, half reported using a partial scoring method (Bertilsson et al., 2021; Brewer & Unsworth, 2012; Minear et al., 2018; Wiklund-Hornqvist et al., 2014), whereas the other half reported using an absolute scoring method (Agarwal et al., 2017; Jonsson et al., 2021; Tse & Pu, 2012; Yang et al., 2020). Partial scoring is quantified as the sum of all items recalled in the correct serial position. In contrast, absolute scoring is measured as the sum of all trials in which *all* items were recalled in the correct serial order. Past research has provided evidence that partial scoring has better psychometric properties than absolute scoring (Conway et al., 2005; Friedman & Miyake, 2005). Additionally, the scoring method has been shown to influence the strength or presence of a correlation between WM and higher-order cognitive abilities such as reading comprehension (Friedman & Miyake, 2005) and fluid intelligence (Unsworth & Engle, 2007a).

One advantage of the bulk of the prior research on this topic using the same version of the operation span task (Unsworth et al., 2005) is that it facilitates a more direct comparison of the WM ability range of each sample across studies. However, one drawback of solely using operation span to measure WM is that such estimates are limited in construct variance, which is better represented in composites or latent factors composed of multiple measures (for similar discussions see Redick et al., 2016; Shipstead et al., 2012). Substantial variation in WM ability has been observed among past studies, complicating generalizability of results. As a means of comparison, Redick et al. (2012) provided normative data for various complex span tasks (see Supplemental Materials for descriptive statistics). Most relevant to the current research, the results for operation span were established using a sample of 6,236 subjects recruited from multiple universities and a non-student

Table 1
Summary of past research on working memory and retrieval practice.

Citation	Sample Size (N =)	Design	Task/Materials	WM Scoring	WM Descriptives	Comparison to Normative WM	RP Benefit
Harrison et al. (2012)	48	Within	Passage reading	Partial	NR	NR	High WM
Tse & Pu (2012)	160	Within	Paired-associates	Absolute	51.54 (13.75)	High	High WM
Yang et al. (2020)[a]	1075	Between	Multi-list free recall	Absolute	Test: 51.56 (13.75)Study: 50.60 (14.76)	High	Low WM
Agarwal et al. (2017)[b]	156	Within/ Between	General knowledge	Absolute	60.30 (14.30)	High	Low WM
Jonsson et al. (2021)	324	Within	Paired-associates	Absolute	32.27 (16.90)	Low	Equal
Wiklund-Hornqvist et al. (2014)	83	Between	Lecture learning	Partial	Test: 40.81 (16.61)Study: 36.50 (19.18)	Low	Equal
Minear et al. (2018)	343	Within	Paired-associates	Partial	41.20 (18.50)	Low	Equal
Bertilsson et al. (2021)	196	Within	Paired-associates	Partial	53.17 (12.44)	Low	Equal
Brewer & Unsworth (2012)	107	Within	Paired-associates	Partial	60.51 (12.49)	Normal	Equal

Note. Overview of past research examining the relationship between WM and retrieval practice. Observations per cell: [a] n = 433 (test condition) and n = 404 (study condition), [b] n = 39. Design: Refers to manipulation of retrieval practice conditions. Agarwal et al. (2017) had additional between subject manipulations of feedback and retention interval. Task/Materials: Task and materials used for the retrieval practice manipulation. NR: Not reported. WM Descriptives: Mean WM score is reported. Standard deviation is reported in parentheses. Comparison to Normative WM: See Redick et al. (2012) or the supplemental materials for normative WM data. RP Benefit: Refers to the group that showed a larger benefit of retrieval practice. See also Hinze & Rapp, 2014, who report using operation span and science texts and "did not observe any effects of WMC [working memory capacity] on quiz or final test performance" (p. 602). No further information regarding WM was reported.

community sample. When comparing these established operation span norms to prior WM and retrieval practice studies, the sample in Brewer and Unsworth (2012) exhibited operation span descriptive statistics most similar to the normative data of Redick et al. (2012), and Brewer and Unsworth found no relationship between WM and retrieval practice. Other WM and retrieval practice studies either reported mean operation span scores in a high ability range above the 66.6 percentile (Agarwal et al., 2017; Tse & Pu, 2012; Yang et al., 2020), or a low ability range below the 33.3 percentile (Bertilsson et al., 2021; Jonsson et al., 2021; Minear et al., 2018; Wiklund-Hornqvist et al., 2014) of the normative data in Redick et al. (2012). Samples clustered at the high or low ends of WM ability may not be truly representative of the population, and the observed correlations may be distorted by restriction of range (for similar discussion in the n-back literature, see Redick & Lindsey, 2013).

Finally, as succinctly stated by Salthouse et al. (2006), "it is important to consider two prerequisites for the meaningful interpretation of correlations; moderately large sample size, and adequate reliability of the critical measures" (p. 107). Studies investigating the relationship between WM ability and retrieval practice are not immune to these considerations. Regarding sample size, many, but not all, of the studies listed in Table 1 have moderate-to-large sample sizes. However, even some of the studies with larger samples used a between-subjects manipulation of retrieval practice in their research design, meaning the specific sample size used to assess the relationship between WM and retrieval practice was smaller for the critical analyses. As the sample size in individual differences studies can greatly influence the precision of the correlation (Schönbrodt & Perugini, 2013), this too may be a source of the variability across studies in Table 1.

Regarding reliability, the complex span measures of WM most often used in this line of research possess high levels of reliability across a variety of dimensions, including internal consistency (Redick et al., 2012). In addition, the retrieval practice effect is extremely robust, with an enormous number of studies in which retrieval practice produces a significant learning effect relative to a condition without retrieval practice (Adesope et al., 2017; Karpicke, 2017; Rowland, 2014). However, it is this robustness of the retrieval practice effect in the experimental approach, and in particular the difference score calculated to measure the retrieval practice effect, that can contribute to its lower reliability as an individual differences measure (Draheim et al., 2019; Hedge et al., 2018; Salthouse et al., 2006; Unsworth, 2010). Whether the cognitive task is Stroop (congruent vs. incongruent), task-switching (switch vs. non-switch), or retrieval practice (retrieval vs. study), the reliability of the difference score calculated reflects the reliability of the individual conditions and the magnitude of the correlation between the conditions. As noted by Salthouse et al. (2006), "contrary to what is sometimes assumed, this is not because difference scores are intrinsically unreliable (e.g., Rogosa & Willett, 1983), but instead it occurs because the original variables are often so highly correlated with one another that there is little variability in the difference" (p. 122). Indeed, Brewer and Unsworth (2012) discussed this exact situation in one of the first individual differences studies of the retrieval practice effect. Although the separate memory conditions to calculate the retrieval practice effect difference score produced high reliabilities, because the tested and non-tested memory conditions were highly correlated with each other, the reliability of the difference score was poor.

Another related issue is that difference scores do not completely remove the influence of the control condition. The score in the experimental condition is dependent on the score in the control condition. When a difference score is correlated with another measure, it is not then reflecting solely the variance due to the difference between the control and experimental conditions as typically interpreted, but also the variance due to the initial value of the control. This issue is irrespective of reliability, but may be exacerbated with unreliable measures. Cohen et al. (2002) discuss this issue in greater depth and describe partial correlations as a potential solution. Partial correlations calculate the correlation between a dependent measure and performance in the experimental condition, partialing out performance in the control condition. This removes the correlation of the control condition with other variables. Another approach that avoids the use of difference scores altogether and still allows for an examination of the relationship between measures is an ANCOVA. Given these statistical issues, although it may be straightforward to use the retrieval practice difference score in individual differences studies to calculate a zero-order correlation, it is important to also consider alternative methods such as ANCOVA and partial correlations to assess the relationship between working memory and retrieval practice.

The present experiments

The current research addressed how variations in the learning task procedure and measurement of WM influence the strength and/or direction of the relationship between WM and retrieval practice. In Experiments 1 and 2, we implemented a learning-to-criterion procedure for the paired-associates learning task based on Karpicke and Smith (2012). This procedure has not been used in the previous studies investigating the association between WM and retrieval practice. As

initial retrieval success has been shown to positively influence the magnitude of a retrieval practice effect (Rowland, 2014), this factor may be especially important to address when conducting individual differences research. Initial performance for high ability individuals may be greater than performance for low ability individuals, potentially introducing moderating effects on the size of a retrieval practice effect. In Experiment 3, we aimed to replicate and generalize the results of Experiments 1 and 2 to different learning materials by using a general knowledge question learning task from Agarwal et al. (2017). In addition, analyses of results were conducted utilizing both partial and absolute scoring methods for the WM tasks. WM was measured with both operation span, to facilitate comparisons to previous research, and symmetry span, to permit the calculation of a WM composite instead of relying on a single measure of the WM construct. Finally, as previously discussed, relying solely on the retrieval practice difference score may be problematic in an individual differences design (Draheim et al., 2019). Therefore, we also used additional analysis approaches to thoroughly investigate the relationship between individual differences in WM and the retrieval practice effect.

Experiment 1

Data availability

Materials, data, and analysis files for this and subsequent experiments are available on OSF (https://osf.io/7jf9m/?view_only = f1317737f1a9430f96e4bc057066780b).

Method

Subjects

One hundred and thirty-one Purdue University undergraduates participated in person in exchange for either course credit or for monetary compensation ($10/hour). The subjects recruited to participate in this study were contacted because they had already completed measures of WM in a previous, unrelated session and had indicated that they would like to be notified about future research studies for which they might qualify. With a final sample size of $N = 131$, a zero-order correlation of $r = +/-.172$ is statistically significant (alpha = .05, two-tailed). Prior work reported correlations between WM and the retrieval practice effect ranging from $r = .01$ to $r = .42$, so a precise effect size estimate was unclear. However, this indicated that a small correlation would still be statistically significant with our sample size.

Design & materials

Learning condition (repeated retrieval, restudy, or no practice) was manipulated within-subject. A set of 36 Swahili–English word pairs (e.g., mashua – boat) were selected from the norms of Nelson and Dunlosky (1994) for this experiment. Twelve word pairs were assigned to each learning condition. This assignment was counterbalanced across conditions and subjects.

Procedure

An overview of the procedure is provided in Fig. 1. For all sessions of the experiment, subjects were tested individually, and instructions were presented on the computer. All procedures were performed in compliance with relevant laws and institutional guidelines and were approved by the Purdue University Institutional Review Board (Protocol Numbers: #1311014187, #1007009512; Approval Dates: November 26, 2013, July 30, 2010). The computerized tasks in Session 1 were programmed in JavaScript. Session 1 of the experiment began with subjects signing an informed consent form and then learning each of the 36 Swahili–English word pairs to the criterion of one correct recall per item. Specifically, subjects alternated between study and recall blocks until the English translation of each Swahili word had been recalled correctly. If the correct English translation had not been recalled following 6 recall blocks, the program would progress to the next phase of the task. Throughout this experiment, a recall attempt was scored as correct if the first three letters of the response matched the target word (Karpicke & Smith, 2012). For study trials, both the Swahili word and its English translation were presented in the center of the screen with the translation appearing directly beneath the Swahili word. Each word pair was shown on the screen for a total of 3 s with a 1 s inter-stimulus-interval, and the order in which the words were presented was randomized at the block's onset. During recall blocks, subjects were shown a Swahili word and asked to type its English translation into a short answer box. Recall trials were self-paced, and if a subject recalled the item correctly, it was removed from subsequent study and recall blocks.

After the English translation of each of the 36 Swahili words had been correctly recalled or 6 recall trials had occurred, the learning activity manipulation began. Twelve of the words in the repeated retrieval condition received two additional test trials, 12 word pairs in the restudy condition received two additional study trials, and the remaining 12 word pairs in the no practice condition were dropped and not presented again in the learning activity. Words were presented blocked by trial type, and the order of the blocks was randomized across subjects. Once this phase of the experiment was complete, subjects were finished with Session 1 and thanked for their participation.

Session 2 occurred 24 h after Session 1, and all subjects were given a

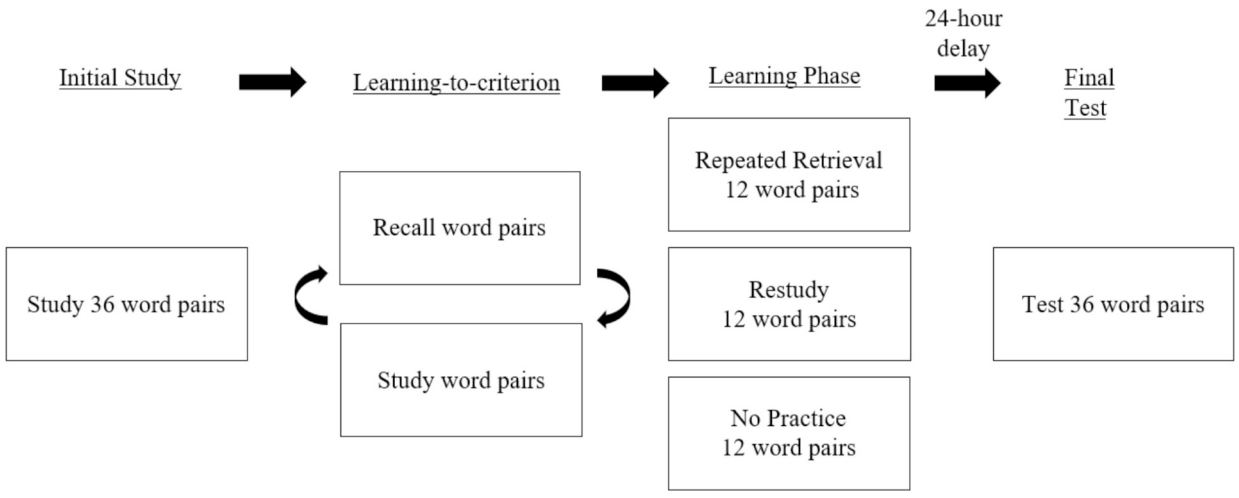

Fig. 1. Procedure for the paired-associates task of Experiment 1.

test in which they were asked to recall the English translation of all 36 Swahili words. The format for this test was identical to the repeated retrieval trials in Session 1. After subjects completed the final test, they were debriefed and thanked for their participation.

Working memory tasks

Operation Span. Subjects completed this task using E-Prime software. During each trial, subjects solved a series of math problems while trying to remember a set of unrelated letters (taken from the set F, H, J, K, L, N, P, Q, R, S, T, Y). In the math portion of the task, subjects were shown a math operation followed by a number. They then completed a true/false judgment task indicating if the number was the correct answer to the previous operation. The presentation duration for math operations varied across subjects and was determined based on how quickly subjects had solved the math operations in the practice phase of the task. For experimental trials, the true/false judgment for the math operation would be skipped if this time limit was exceeded. Following each math operation, subjects were presented with a letter for 1 s. Immediately afterward, the next problem was presented and solved, and then another letter was presented. After this alternating sequence occurred three to seven times, subjects were asked to recall the presented letters in order (i.e., list lengths ranged from 3 to 7). Three trials of each list length were presented, for a total possible of 75 letters correctly recalled. The order in which the different list lengths were presented varied randomly. At recall, subjects reported the list as they remembered it by clicking on the appropriate letters (Redick et al., 2012).

Symmetry Span. Subjects completed this task using E-Prime software. On each trial in this task, subjects performed a symmetry-judgment task while trying to remember a sequence of locations in which red squares were presented within a matrix. Subjects viewed an 8 × 8 matrix with some squares filled in black and reported whether the design was symmetrical about its vertical axis. The design was symmetrical for 50 % of trials. Subjects were then presented with a 4 × 4 matrix with one of the cells filled in red for 650 ms. They then performed the judgment task again and were presented with another matrix with a red square. After this alternating sequence occurred two to five times, subjects were asked to report the sequence of red-square locations in the preceding displays (i.e., list lengths ranged from 2 to 5), in the order in which they had appeared, by clicking on the cells of an empty matrix. There were three trials of each list length, for a possible total of 42 locations correctly recalled (Redick et al., 2012).

Analyses

To examine the differences in recall performance among the conditions, we conducted a repeated measures analysis of variance (ANOVA) with learning condition as a within-subjects variable with three levels: repeated retrieval, restudy, and no practice. Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated for learning condition, $\chi^2(2) = 7.07$, $p = .029$. However, the pattern of results (for all experiments) did not differ when sphericity was assumed and when the Greenhouse-Geisser correction was used. For simplicity, results assuming no violation of the assumption of sphericity are reported.

Next, we examined associations with WM. First, total scores using the partial scoring method for operation span and symmetry span were standardized and averaged together to create a WM composite *z*-score. To examine if learning rate during the learning-to-criterion phase differed as a function of WM ability, we first conducted an ANCOVA with recall trial as a within-subjects variable and WM composite *z*-score as a continuous covariate. Additionally, we divided the sample into low, mid-range, and high WM groups based on each subject's WM composite *z*-score and conducted a mixed design ANOVA with recall trial as a within-subjects variable and WM group as a between-subjects variable.

To investigate the effect of retrieval practice on final recall controlling for the effect of WM, we conducted an ANCOVA with learning condition (repeated retrieval, restudy, and no practice) as a within-subjects variable and WM composite *z*-score as a continuous covariate. Note the use of ANCOVA allows examination of a role for potential interaction of WM with learning condition, without concern about potential low reliability or other statistical issues inherent to difference scores when using a correlational approach.

Finally, we examined the correlation between WM and the retrieval practice difference score, despite potential concerns about the reliability of difference scores. This analysis was conducted to facilitate comparison with previous research. To quantify the retrieval practice effect, we calculated a difference score by subtracting the proportion correct in the restudy condition from the proportion correct in the repeated retrieval condition. We also conducted a partial correlation analysis, where the correlation between the WM composite *z*-score and final test performance in the repeated retrieval condition was calculated after first partialing out restudy condition performance. Bayes Factors were provided where appropriate to show degree of evidence in favor of the null hypothesis for correlation results. Partial-eta squared (η_p^2) is provided as an index of effect size. All analyses were two-tailed and conducted using $\alpha = 0.05$, and 95 % confidence intervals are reported.

Results

Learning-to-criterion performance

The cumulative learning curves from the learning-to-criterion phase of the experiment can be found in Fig. 2. This curve shows the proportion of items that had been correctly recalled in each recall trial period. Given that this phase occurred before the learning activity manipulation, no differences among conditions were expected.

Final test performance

Descriptive statistics and reliabilities are reported in Table 2 and the Supplemental Materials, respectively. As can be seen, performance was highest in the repeated retrieval condition and lowest in the no practice condition, with restudy performance in between these conditions.

A repeated-measures ANOVA with learning condition as a within-subjects variable confirmed that there was a significant main effect of condition, $F(2, 260) = 209.54$, $MSE = .02$, $p < .001$, $\eta_p^2 = .62$. Subsequent pairwise comparisons showed significantly greater performance in the repeated retrieval condition relative to the restudy condition, $t(130) = 8.20$, $p < .001$, $d = 0.70$, CI [0.50, 0.89], and the no practice condition, $t(130) = 19.50$, $p < .001$, $d = 1.74$ [1.47, 2.01]. In addition, the restudy

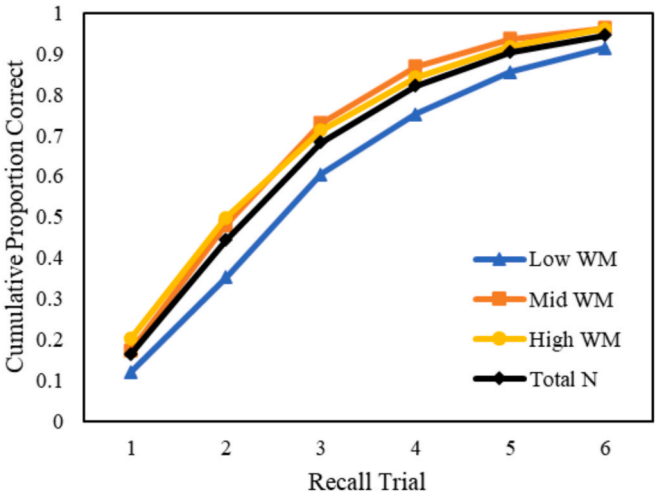

Fig. 2. Cumulative recall of items across recall trials in the learning-to-criterion phase of Experiment 1.

Table 2
Experiment 1 descriptive statistics.

Variables	Mean	SD	Min	Max	Skew	Kurtosis
Repeated Retrieval	.80	.18	.08	1.00	−0.95	1.11
Restudy	.68	.23	.08	1.00	−0.44	−0.83
No Practice	.45	.22	.00	1.00	0.20	−0.54
RP Difference Score	.12	.18	−.25	.67	0.67	0.79
Operation Span Partial	61.37	10.65	22.00	75.00	−1.17	1.52
Symmetry Span Partial	31.54	6.53	13.00	42.00	−0.58	−0.28

Note: RP Difference Score: Repeated Retrieval – Restudy; Min: Minimum; Max: Maximum.

condition outperformed the no practice condition, $t(130) = 12.00$, $p < .001$, $d = 1.05$ [0.84, 1.26].

Associations with working memory

The mean operation span and symmetry span scores (see Table 2) were consistent with the normative data of Redick et al. (2012) and previous studies using this university sample (e.g., Wiemers & Redick, 2019), as was the correlation between operation span and symmetry span, $r(129) = .46$, $p < .001$, CI [.31,.59].

Learning-to-Criterion Performance. To examine whether learning-to-criterion phase performance differed as a function of WM ability, we first conducted an ANCOVA with recall trial as a within-subjects variable and WM composite z-score as a continuous covariate. There was a significant main effect of recall trial, $F(5, 645) = 1410.79$, $MSE = .01$, $p < .001$, $\eta_p^2 = .92$, as well as a significant main effect of WM, $F(1, 129) = 7.47$, $MSE = .14$, $p = .007$, $\eta_p^2 = .06$. Critically, the interaction was also significant, $F(5, 645) = 4.68$, $MSE = .01$, $p < .001$, $\eta_p^2 = .04$.

To further examine this interaction, we divided the sample into three groups based on their WM ability (for a similar approach, see Poole & Kane, 2009). Participants below the 33rd percentile were classified as low WM ($n = 43$), participants between the 33rd and 66th percentile were classified as mid-range WM ($n = 46$), and participants between the 66th and 100th percentile were classified as high WM ($n = 42$). We then conducted a mixed-design ANOVA with recall trial as a within-subjects variable and WM group as a between-subjects variable. Mirroring the ANCOVA results, there was a significant main effect of recall trial, $F(5, 640) = 1401.30$, $MSE = .01$, $p < .001$, $\eta_p^2 = .92$, a significant main effect of WM group, $F(2, 128) = 5.22$, $MSE = .14$, $p = .007$, $\eta_p^2 = .08$, and a significant interaction, $F(5, 640) = 2.60$, $MSE = .01$, $p = .004$, $\eta_p^2 = .04$. As illustrated in Fig. 2, post-hoc comparisons showed that while participants in the low-WM group performed worse than individuals with higher WM on trials 1–4 of the learning-to-criterion phase, there were no significant WM group differences on trial 6.

Final Test Performance. We conducted an ANCOVA with learning condition (repeated retrieval, restudy, and no practice) as a within-subjects variable and WM composite z-score as a continuous covariate. Mirroring the ANOVA results above, there was a main effect of learning condition, $F(2, 258) = 208.95$, $MSE = .02$, $p < .001$, $\eta_p^2 = .62$. The main effect of WM was also significant, $F(1, 129) = 7.87$, $MSE = .09$, $p = .006$, $\eta_p^2 = .06$, indicating that individuals with higher WM scores performed better across learning activity conditions. However, the critical interaction between learning condition and WM was not significant, $F(2, 258) = 0.64$, $MSE = .02$, $p = .531$, $\eta_p^2 < .01$.

To further examine the relationship between the retrieval practice effect and WM, we computed the correlation between the retrieval practice difference score (performance in the repeated retrieval condition minus performance in the restudy condition) and the WM composite z-score. Visualization of this analysis is shown in Fig. 3, and a correlation matrix for all variables of interest can be found in Table 3. Consistent with the ANCOVA results, the magnitude of the retrieval practice effect did not differ as a function of WM ability. The correlation was not significant, $r(129) = -.10$, $p = .223$, CI: [−.27,.07]. Using JASP, this non-significant correlation had a $BF_{01} = 4.79$, which is classified as 'moderate evidence' in favor of the null hypothesis (Wagenmakers et al., 2018).

Additionally, we computed the correlation between performance in the repeated retrieval condition and the WM composite z-score, partialing out performance in the restudy condition. Consistent with the other methods of analyzing the relationship between WM and retrieval

Table 3
Experiment 1 correlation matrix.

Variables	1	2	3	4	5	6
1. Repeated Retrieval	–					
2. Restudy	**.66**	–				
3. No Practice	**.51**	**.55**	–			
4. RP Difference Score	.13	**−.66**	**−.21**	–		
5. Operation Span Partial	.14	**.19**	**.17**	−.11	–	
6. Symmetry Span Partial	.17	**.17**	**.18**	−.06	**.46**	–

Note: RP Difference Score: Repeated Retrieval – Restudy; Bolded values indicate a significant correlation ($p < .05$).

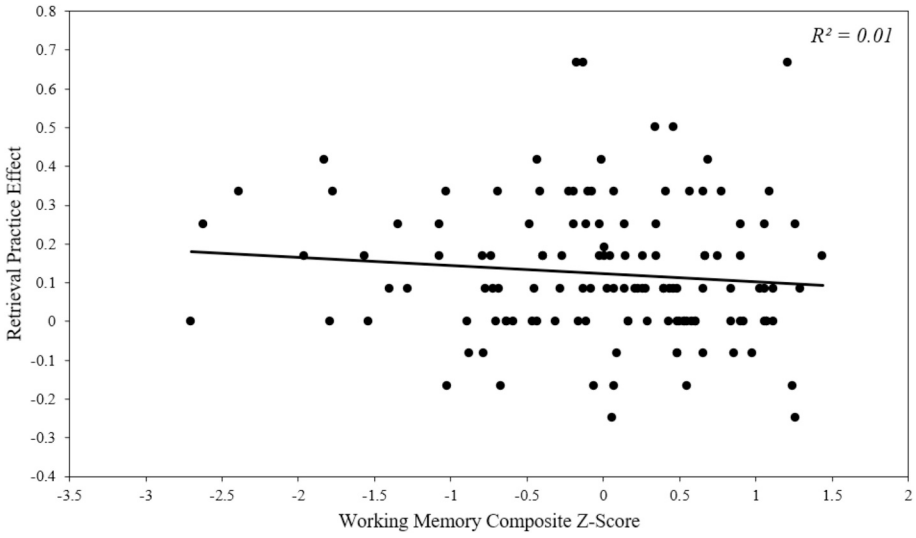

Fig. 3. Magnitude of the retrieval practice effect (*repeated retrieval − restudy*) as a function of WM composite z-score in Experiment 1.

practice, the partial correlation was not significant, $pr(128) = .05$, $p = .588$, CI [−.11,.21].[1]

Discussion

Experiment 1 is the first retrieval practice study that used a learning-to-criterion procedure while assessing the relationship with individual differences in WM. As noted by Karpicke (2017), using a learning-to-criterion procedure helps ensure items are initially learned well enough such that, at final test, differential amounts of forgetting due to retrieval practice, restudy, or no practice learning conditions can be observed. Given that previous studies have shown individuals higher in WM perform better than those lower in WM on paired-associates learning tasks (Martin et al., 2020; Rosen & Engle, 1998; Unsworth, 2009b) in general, we wanted subjects across the distribution of WM scores to start with similar levels of performance. Our analysis of learning-to-criterion performance as a function of WM supported these ideas. While there were WM-related differences on the early recall trials of the learning-to-criterion procedure, these differences were no longer significant by the final trial.

The results of Experiment 1 indicated a robust retrieval practice effect, as seen by significantly higher final recall performance for items that had been repeatedly tested relative to items that had been restudied or not practiced from the study phase. However, with a large sample size and a distribution of WM scores that was very similar to normative data, Experiment 1 provided no evidence of a differential benefit of retrieval practice dependent on WM ability. Thus, our results are very similar to those of others (Bertilsson et al., 2021; Brewer & Unsworth, 2012; Jonsson et al., 2021; Minear et al., 2018) who have also failed to observe a significant relationship between individual differences in WM and the retrieval practice effect when using word pairs.

Experiment 2

Experiment 2 aimed to replicate the results of Experiment 1, with a few modifications to the procedure. Due to 40 subjects in Experiment 1 failing to learn all items to a criterion of one correct recall, the learning-to-criterion procedure for Experiment 2 was modified so that the program would not progress to the learning activity phase before correct recall of all items once or after 8 recall blocks. We also removed the no practice condition and administered only the repeated retrieval and restudy conditions, and we increased the retention interval from 24 h to 48 h. In addition, in Experiment 1 we used a relatively large sample size ($N = 131$), particularly for a within-subjects manipulation of learning condition. The 95 % confidence interval around the observed nonsignificant correlation between WM and the retrieval practice difference score ($r = -.10$) is [−.27,.07], and the Bayes Factor showed 'moderate evidence' in favor of the null hypothesis. Given this, we increased the sample size in Experiment 2 in order to gain greater precision surrounding the outcome. Finally, in Experiment 2, analyses are reported utilizing both absolute and partial scoring methods for the WM span tasks to examine the potential role of scoring method on the association with the retrieval practice effect.

Method

Subjects

Two hundred and twenty-eight Purdue undergraduates participated in person in exchange for course credit in an introductory psychology course. From this sample, 213 subjects completed both sessions of the study, and 3 additional subjects were excluded due to an error in data collection, for a final sample of $N = 210$ subjects (108 female; 37 non-native English speakers;$M_{age} = 18.82$,$SD_{age} = 0.98$). We used a larger sample size than Experiment 1 to provide more statistical power to potentially detect even a small association between WM and retrieval practice. With a final sample size of $N = 210$, a zero-order correlation of $r = +/-.136$ is statistically significant (alpha = .05, two-tailed).

Design & materials

The overall design of Experiment 2 was very similar to Experiment 1. A within-subjects design was again used, and subjects engaged in two different learning activities (repeated retrieval, restudy). A set of 24 Swahili–English word pairs (e.g., mashua – boat) were selected from the norms of Nelson and Dunlosky (1994) for this experiment. Twelve word pairs were assigned to each learning activity and presented in a randomized order. This assignment was counterbalanced across conditions and subjects.

Procedure

An overview of the procedure is provided in Fig. 4. For all sessions of the experiment, subjects were tested individually, and instructions were presented on the computer. All procedures were performed in compliance with relevant laws and institutional guidelines and were approved by the Purdue University Institutional Review Board (Protocol Number: #1007009512; Approval Date: July 30, 2010). Session 1 of the experiment began with subjects signing an informed consent form and then answering various demographics questions (age, gender, first language). Following this, in the learning-to-criterion phase, subjects learned each of the 24 Swahili–English word pairs to the criterion of one correct recall per item. Word pairs were presented in a randomized order. Subjects alternated between study and test blocks until the English translation of each Swahili word had been recalled correctly. If the correct English translation had not been recalled following eight recall blocks, the program would progress to the next phase of the task. Throughout this experiment, a recall attempt was scored as correct if the first three letters of the response matched the target word (Karpicke & Smith, 2012). For study trials, both the Swahili word and its English translation were presented in the center of the screen with the translation appearing directly beneath the Swahili word. Each word pair was shown on the screen for a total of 5 s with a 500 ms inter-stimulus-interval. During test blocks subjects were shown a Swahili word and asked to type its translation into a short answer box. Subjects were given 8 s to type their response before the program progressed to the next item. If a subject recalled the item correctly, they were given feedback informing them that the answer was correct and the item was removed from subsequent study and test blocks. Then, the learning activity manipulation began. Twelve of the words in the repeated retrieval condition received two additional test trials, and 12 word pairs in the restudy condition received two additional study trials.

Session 2 occurred 48 h after Session 1, and all subjects were given a test in which they were asked to recall the English translation of all 24 Swahili words. The format for this assessment was similar to the repeated retrieval trials in Session 1. Subjects were shown a Swahili word in the center of the screen and asked to type in the English translation. Subjects were given 12 s to type their response before the program progressed to the next item. Following the final test, subjects completed operation span and symmetry span tasks. In contrast to Experiment 1, both span tasks were programmed in JavaScript. However, the procedure for these tasks was modeled from the E-Prime versions used in Experiment 1.

Analyses

To examine the differences in performance across learning activity conditions, we conducted a paired-samples *t*-test comparing

[1] Forty subjects did not learn all items to criterion after 6 repetitions. The results of the analyses with these subjects excluded are reported in the Supplemental Materials. These results did not differ from the pattern of results for the full sample. Analyses were also conducted utilizing only operation span scores for greater ease of comparison to past studies. Results did not differ from the reported analyses using both the full sample and the reduced sample.

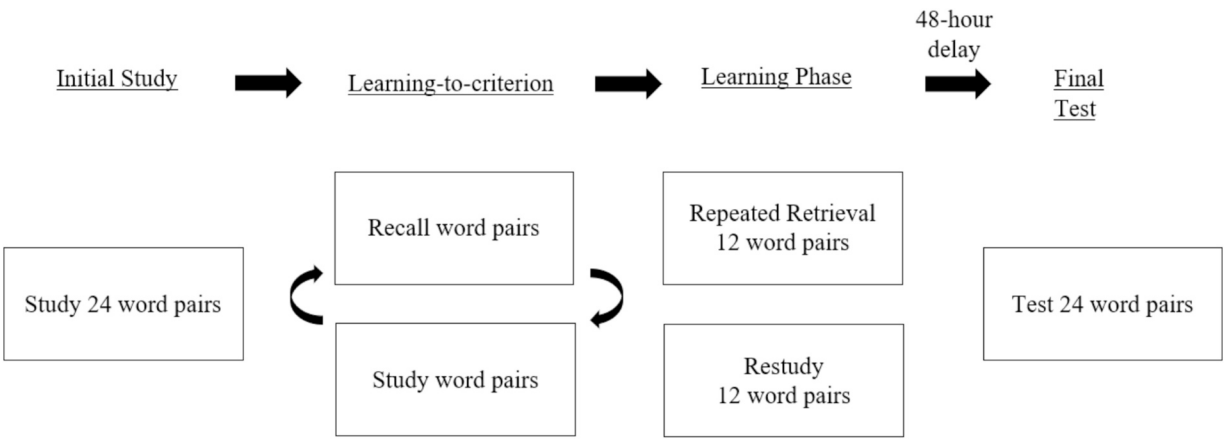

Fig. 4. Procedure for the paired-associates task in Experiment 2.

performance in the repeated retrieval condition to performance in the restudy condition. Next, we examined associations with WM. Scores for operation span and symmetry span were standardized and averaged together to create a WM composite *z*-score. Note that span scores were calculated utilizing both partial and absolute scoring methods. To examine if learning rate during the learning-to-criterion phase differed as a function of WM ability, we conducted an ANCOVA with recall trial as a within-subjects variable and WM composite *z*-score as a continuous covariate. Additionally, we divided the sample into low, mid-range, and high WM groups and conducted a mixed-design ANOVA with recall trial as a within-subjects variable and WM group as a between-subjects variable. Following this analysis, we examined final test performance and conducted an ANCOVA with learning condition (repeated retrieval, restudy) as a within-subjects variable and WM composite *z*-score as a continuous covariate. Finally, we examined the association between WM and the retrieval practice difference score. To quantify the retrieval practice effect, we calculated a difference score by subtracting the proportion correct in the restudy condition from the proportion correct in the repeated retrieval condition. We also conducted a partial correlation analysis, where the correlation between the WM composite *z*-score and final test performance in the repeated retrieval condition was calculated after first partialing out restudy condition performance. Bayes Factors were provided where appropriate to show degree of evidence in favor of the null hypothesis for correlation results. All analyses were two-tailed and conducted using $\alpha = 0.05$, and 95 % confidence intervals are reported.

Results

Learning-to-criterion performance

The cumulative learning curve from the learning-to-criterion phase of the experiment can be found in Fig. 5. This curve shows the proportion of items that had been correctly recalled in each recall trial period. Given that this phase occurred before the learning activity manipulation, no differences between conditions were expected.

Learning phase performance

Performance during the learning phase increased with each repetition. A paired-samples *t*-test confirmed that recall in repetition 2 of the learning phase was higher than recall in repetition 1, $t(209) = 4.21$, $p < .001$, $d = .30$ [.16,.44]. Additionally, learning phase performance (Repetition 1: $M = .75$, $SD = .19$; Repetition 2: $M = .78$, $SD = .19$) met the recommended 75 % accuracy suggested by Rowland's (2014) meta-analysis.

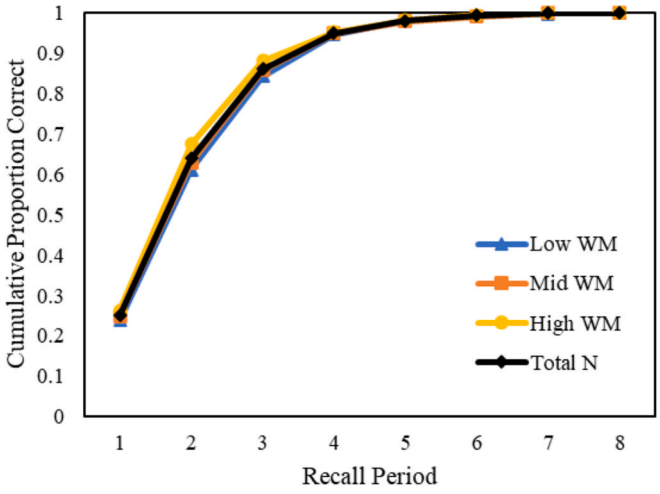

Fig. 5. Cumulative recall of items across recall trials in the learning-to-criterion phase of Experiment 2.

Final test performance

Descriptive statistics are reported in Table 4, and reliabilities are reported in the Supplemental Materials. Performance in the repeated retrieval condition was higher than the restudy condition, indicating a retrieval practice effect. This was confirmed by a paired-samples *t*-test, $t(209) = 6.47$, $p < .001$, $d = .45$ [.30,.59].

Associations with working memory

The mean operation span and symmetry span scores were consistent

Table 4
Experiment 2 descriptive statistics.

Variables	Mean	SD	Min	Max	Skew	Kurtosis
Repeated Retrieval	.70	.21	0.00	1.00	−0.80	0.32
Restudy	.63	.23	0.00	1.00	−0.59	−0.24
RP Difference Score	.07	.17	−.33	.50	0.07	−0.21
Operation Span Absolute	40.27	17.71	0.00	75.00	−0.12	−0.71
Symmetry Span Absolute	20.94	9.09	0.00	42.00	0.05	−0.44
Operation Span Partial	56.92	12.78	2.00	75.00	−1.07	1.34
Symmetry Span Partial	30.02	7.68	4.00	42.00	−0.77	−0.05

Note: RP Difference Score: Repeated Retrieval – Restudy; Min: Minimum; Max: Maximum.

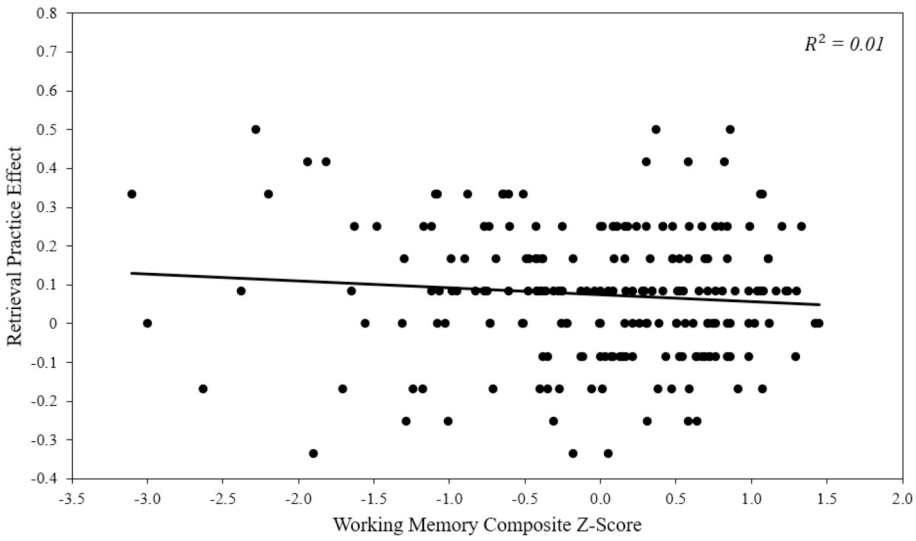

Fig. 6. Magnitude of the retrieval practice effect (*repeated retrieval − restudy*) as a function of WM composite *z*-score (partial scoring) in Experiment 2.

with the normative data of Redick et al. (2012) and Experiment 1, as was the correlation between operation span and symmetry span with partial scoring, $r(208) = .53$, $p < .001$, and absolute scoring, $r(208) = .44$, $p < .001$.

Learning-to-Criterion Performance. To examine whether learning-to-criterion phase performance differed as a function of WM ability, we first conducted an ANCOVA with recall trial as a within-subjects variable and WM composite *z*-score (partial scoring) as a continuous covariate. There was a significant main effect of trial, $F(7, 1421) = 1724.06$, $MSE = .01$, $p < .001$, $\eta_p^2 = .90$. However, the main effect of WM was not significant, $F(1, 203) = 3.20$, $MSE = .05$, $p = .075$, $\eta_p^2 = .02$. The interaction between trial and WM was significant, $F(7, 1421) = 3.47$, $MSE = .01$, $p = .001$, $\eta_p^2 = .02$. To further examine this interaction, we divided the sample into three groups based on their WM ability. Participants below the 33rd percentile were classified as low WM ($n = 67$), participants between the 33rd and 66th percentile were classified as mid-range WM ($n = 68$), and participants between the 66th and 100th percentile were classified as high WM ($n = 70$). We then conducted a mixed-design ANOVA with recall trial as a within-subjects variable and WM group as a between-subjects variable. There was a significant main effect of trial, $F(7, 1414) = 1707.28$, $MSE = .01$, $p < .001$, $\eta_p^2 = .89$. However, the main effect of WM group was not significant, $F(1, 202) = .832$, $MSE = .05$, $p = .437$, $\eta_p^2 = .01$. In contrast to the ANCOVA results, the interaction between trial and WM group was also not significant, $F(7, 1414) = 1.18$, $MSE = .01$, $p = .284$, $\eta_p^2 = .01$.

Final Test Performance. We first conducted an ANCOVA with learning condition (repeated retrieval vs. restudy) as a within-subjects variable and the WM composite *z*-score as a continuous covariate. Reported first is the analysis utilizing the partial scoring method for operation span. Mirroring the paired-samples *t*-test results above, there was a significant main effect of learning condition, $F(1, 208) = 41.93$, $MSE = .01$, $p < .001$, $\eta_p^2 = .17$. Performance was better in the repeated retrieval condition than in the restudy condition. The main effect of WM was also significant, $F(1, 208) = 5.08$, $MSE = .08$, $p = .025$, $\eta_p^2 = .02$. Individuals with higher WM ability performed better across conditions. However, the critical interaction between learning condition and WM was not significant, $F(1, 208) = 1.77$, $MSE = .01$, $p = .185$, $\eta_p^2 = .01$.

Next, we conducted the same analysis using the absolute scoring method for the span tasks and found virtually identical results. There was a significant main effect of learning condition, $F(1, 208) = 41.67$, $MSE = .01$, $p < .001$, $\eta_p^2 = .17$, reflecting the retrieval practice effect. The main effect of WM ability was also significant, $F(1, 208) = 5.57$, $MSE = .08$, $p = .019$, $\eta_p^2 = .03$, indicating that individuals with higher WM ability performed better across conditions. But again, the critical interaction between learning condition and WM was not significant, $F(2, 208) = 0.43$, $MSE = .01$, $p = .512$, $\eta_p^2 < .01$.

To further examine the relationship between WM and the retrieval practice effect, we computed the correlation between the retrieval practice difference score (performance in the repeated retrieval condition minus performance in the restudy condition) and the WM composite *z*-score. Visualization of this analysis is shown in Fig. 6, and a correlation matrix for all variables of interest can be found in Table 5. Separate analyses were conducted with the WM composite *z*-score calculated using a partial scoring method and an absolute scoring method to facilitate comparison with past studies and examine if scoring impacted the observed correlation magnitude. Using a partial scoring method, the correlation between WM and the retrieval practice difference score was not significant, $r(208) = -.09$, $p = .185$, CI [−.22,.04]. Using JASP, this non-significant correlation had a $BF_{01} = 4.84$, which is classified as 'moderate evidence' in favor of the null hypothesis (Wagenmakers et al., 2018). Similarly, the correlation between WM and the retrieval practice difference score was not significant when an absolute scoring method was used, $r(208) = -.05$, $p = .512$, CI [−.18,.09]. This correlation had a $BF_{01} = 9.83$, which is also classified as 'moderate evidence' in favor of the null hypothesis (Wagenmakers et al., 2018). Following this analysis, we used a Hotelling-Williams *t*-test to compare the two correlations and observed no significant difference between the two scoring methods, $t(207) = -1.79$, $p = .075$. Overall, the magnitude of the retrieval practice effect did not differ between individuals with high and low WM, and this result was obtained using both WM scoring methods.

We also computed the correlation between performance in the repeated retrieval condition and the WM composite *z*-score, partialing out performance in the restudy condition. Consistent with the other methods of analyzing the relationship between WM and retrieval practice, the partial correlation was not significant, $pr(207) = -.01$, $p = .901$, CI [−.16,.14].[2]

[2] Five subjects did not learn all items to criterion after 8 repetitions. The results of the analyses with these subjects excluded are reported in the Supplemental Materials. These results did not differ from the pattern of results for the full sample. Analyses were also conducted utilizing only operation span scores for greater ease of comparison to past studies. Results did not differ from the reported analyses using both the full sample and the reduced sample.

Analyses across experiments 1 and 2

Given the similarity between Experiments 1 and 2, we pooled the data together. That is, because Experiments 1 and 2 (a) included students from the same university subject pool, (b) used both operation and symmetry span to measure WM, and (c) used the same materials and nearly identical procedures in the repeated retrieval practice and restudy conditions, we aggregated and analyzed the data across experiments. The WM composite *z*-score for this analysis was created based on this new combined sample (for a related discussion, see Brewer, Robey, & Unsworth, 2021). In this combined sample of $N = 341$ subjects, the correlation between the retrieval practice difference score and the WM composite *z*-score was $r(339) = -.07$, $p = .195$. In the combined sample, the $BF_{01} = 6.39$, which again shows moderate evidence in favor of the null hypothesis.

Discussion

The results of Experiment 2 replicated the results of Experiment 1 with a larger sample. We observed a significant retrieval practice effect: Practicing retrieval enhanced final recall relative to restudy. However, we again failed to observe a significant relationship between WM and the retrieval practice effect. The absence of a relationship was found using both the partial and absolute scoring method of the span tasks, and the conclusions were strengthened further with an analysis that combined the data from Experiments 1 and 2. The Bayes Factors reported in both experiments and the combined analysis indicated moderate evidence in favor of the null hypothesis. Both experiments provide further evidence that retrieval practice is equally beneficial for students with differing levels of WM ability.

Experiment 3

Experiment 3 aimed to determine whether the results of Experiments 1 and 2 could be replicated using different materials. Although evidence is limited due to a small number of past studies (Table 1), the experiments that found a significant association between WM ability and the retrieval practice effect more often used paradigms or materials other than paired-associates learning. For example, past studies have used multi-list free recall and materials such as educational texts and general knowledge questions. In contrast, of the experiments that observed no significant WM association, a paired-associates task was more often used. Tasks differ in their reliance on various cognitive processes and may use materials applicable to different educational settings. For this reason, we wanted to extend our results to a different paradigm and select a study to replicate in which a relationship between individual differences in WM and the retrieval practice effect was obtained. We elected to model Experiment 3 of the current study on Agarwal et al. (2017).

Agarwal et al. (2017) investigated the relationship between individual differences in WM and retrieval practice benefit by presenting subjects with various general knowledge questions. Following the initial study phase, questions were then presented as study or test trials.

Table 5
Experiment 2 correlation matrix.

Variables	1	2	3	4	5	6	7
1. Repeated Retrieval	–						
2. Restudy	**.71**	–					
3. RP Difference Score	**.25**	**−.50**	–				
4. Operation Span Absolute	**.14**	**.18**	−.07	–			
5. Symmetry Span Absolute	.09	.09	−.01	**.44**	–		
6. Operation Span Partial	.13	**.20**	−.12	**.90**	**.48**	–	
7. Symmetry Span Partial	.07	.09	−.05	**.47**	**.90**	**.53**	–

Note: RP Difference Score: Retrieval Practice – Restudy; Bolded values indicate a significant correlation ($p < .05$).

Additional factors such as lag (intervening items between initial study and study/test trial), feedback presence, and retention interval were manipulated to investigate if optimal conditions for learning could be identified. Because Agarwal et al. (2017) manipulated feedback presence (yes/no) and retention interval (10 min/2 days) between subjects, four independent samples were used. Agarwal and colleagues observed a significant association ($r = -.42$) between WM and retrieval practice benefit *only* in the 2-day retention interval when feedback was given, such that individuals with lower WM showed a greater benefit of retrieval practice. There were no significant WM and retrieval practice relationships in the other three groups.

However, a few aspects of the sample used by Agarwal et al. (2017) warrant further examination. First, as noted above, because feedback presence and retention interval were manipulated between subjects, the sample size in each condition ($n = 39$) was quite small for a correlational study. Indeed, although the correlation ($r = -.42$) was statistically significant, the sample size used for the 2-day/feedback present condition resulted in a 95 % confidence interval of [−.65 to −.12]. In addition, using the scores from Redick et al. (2012) as a reference (see Supplemental Materials), the Agarwal et al. sample obtained a high mean operation span score ($M = 60.3$, $SD = 14.3$). The operation scores for the Agarwal et al. sample in the 2-day/feedback present condition were fairly homogeneous, with the majority of subjects obtaining an operation span score of 60–70 using the absolute scoring method, and very few subjects obtaining a score below 40 (cf. Fig. 2, Agarwal et al., 2017). Agarwal et al. noted this characteristic of their sample, stating that "subjects in our sample demonstrated working memory capacities toward the higher end of the scale" (p. 768). Therefore, we wanted to examine whether the results obtained by Agarwal et al. would be replicable with a larger sample size and a wider range of WM scores.

An additional benefit to using the general knowledge task was that Agarwal et al. (2017) reported high initial accuracy during the learning phase of the task, despite a learning-to-criterion procedure not being used. However, in piloting prior to the current study, participants showed low initial accuracy on this task. Therefore, we elected to make some minor adjustments to the general knowledge task procedure used by Agarwal et al. Specifically, rather than including only one repetition of the study/test trials as in Agarwal et al., we included two repetitions to improve initial retrieval success. We also used a retention interval before the final test of 16–32 h, which was shorter than the 48-hour retention interval used by Agarwal et al.

Finally, in addition to using a general knowledge task rather than paired associates, one other change from Experiments 1 and 2 was the use of operation span to measure WM, rather than a composite score of operation span and symmetry span. This change was made because Agarwal et al. (2017) only administered operation span and because the general knowledge task took longer to administer than the paired-associates task we used in Experiments 1 and 2.

Method

Subjects

This experiment was pre-registered (https://aspredicted.org/fc3qy.pdf). One hundred and eighty-two subjects were recruited from the subject pool at Purdue University and the Lafayette, Indiana community through online advertisements and flyers and participated in person.

Subjects either received course credit or monetary compensation ($10/hour) for completion of this study. From this sample, 18 subjects were excluded for failure to meet a pre-registered criterion of 25 % accuracy in the learning phase. Ten subjects were excluded for failure to complete both sessions. Finally, 6 subjects were excluded due to computer program crashes. Thus, data from 148 subjects (84 female; 20 non-native English speakers; $M_{age} = 18.91$, $SD_{age} = 1.63$) was used in the

reported analyses. With a final sample size of $N = 148$, a zero-order correlation of $r = +/-.162$ is statistically significant (alpha = .05, two-tailed).[3]

Design & materials

A within-subjects design was used for this experiment. Specifically, subjects engaged in three different learning activities (repeated retrieval, restudy, and no practice). Seventy-eight general knowledge questions were taken from the Nelson and Narens (1980) norms for use in this experiment. Selection of stimuli and division into sets of six questions that were equated on probability of recall and feeling of knowing was originally completed by Agarwal et al. (2017). For the current study, the questions were then reorganized into three sets of 26 general knowledge questions to be assigned to each learning activity. This assignment was counterbalanced across conditions and subjects.

Procedure

All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the Purdue University Institutional Review Board (Protocol Number: IRB-2021-413; Approval Date: March 17, 2021). For all sessions of the experiment, subjects were tested individually, and instructions were presented on the computer. Session 1 began with the subject signing informed consent. Following this, subjects completed a general knowledge question task. An overview of the general knowledge task procedure is provided in Fig. 7. Following the general knowledge task, subjects completed an operation span task and answered various demographics questions (age, gender, age that they learned English). Session 2 occurred 16–32 h later, but always on the subsequent day. For Session 2, subjects completed the final test phase of the general knowledge task.

The general knowledge task was based on the task used by Agarwal et al. (2017) with minor adjustments. Subjects began Session 1 by viewing 78 general knowledge questions. Questions were presented onscreen with their one-word answer directly below for 9 s. Questions were separated by a 1 s inter-stimulus-interval. The presentation order was randomized at the block's onset. Following this phase of the task, 26 of the questions received two additional study trials (restudy), 26 questions received two additional recall trials (repeated retrieval), and the remaining 26 questions were dropped and not presented again in the learning activity (no practice). For restudy trials, both the general knowledge question and its answer were presented in the center of the screen with the answer appearing directly beneath the question. The question–answer pair was presented for 11 s. For repeated retrieval trials, the general knowledge question was presented in the center of the screen with a blue cursor below. Subjects had 8 s to recall and type the answer. The correct answer was then presented for 3 s, regardless of whether the subject had generated a correct or incorrect answer on the previous screen. Thus, the total time was matched between the repeated retrieval and restudy conditions. The order of restudy and repeated retrieval trials as well as the order of the questions within each trial type was randomized. Questions were counterbalanced across all 3 learning conditions.

In Session 2, subjects were tested on all 78 general knowledge questions. In each trial, the general knowledge question was presented in the center of the screen with a blue cursor below. Subjects had 14 s to recall and type the answer. The order of questions was randomized. Spelling errors were evaluated during response scoring, and answers were considered correct if they were within 2 letters of the target spelling.

[3] Although a power analysis was reported in the pre-registration, there was an error in calculating it that made it invalid. Thus, we do not report the power analysis here. However, consistent with Experiments 1 and 2, the sample size used in Experiment 3 shows that a relatively small correlation would be statistically significant.

Operation Span. The same E-Prime version of the task used in Experiment 1 was used for the current experiment.

Analyses

To examine the differences in recall performance among the learning conditions, we conducted a repeated-measures ANOVA with learning condition as a within-subjects variable with three levels: repeated retrieval, restudy, and no practice. Next, we examined associations with WM. To examine if learning phase performance differed as a function of WM, we conducted an ANCOVA with repetition as a within-subjects variable and operation span score (partial scoring) as a continuous covariate. To investigate the effect of retrieval practice on final performance in Session 2 controlling for the effect of WM, we conducted an ANCOVA with learning condition (repeated retrieval, restudy, and no practice) as a within-subjects variable and operation span score as a continuous covariate. Following this analysis, we examined the correlation between WM and retrieval practice. To quantify the retrieval practice effect, we calculated a difference score by subtracting the proportion correct in the restudy condition from the proportion correct in the repeated retrieval condition. We also conducted a partial correlation analysis, where the correlation between WM and final test performance was calculated after first partialing out restudy condition performance. For all analyses, operation span scores were calculated utilizing both partial and absolute scoring methods, to examine the potential effect of scoring method on the results. Bayes Factors were provided where appropriate to show degree of evidence in favor of the null hypothesis for correlation results. Partial-eta squared (η_p^2) is provided as an index of effect size. All analyses were two-tailed and conducted using $\alpha = 0.05$, and 95 % confidence intervals are reported.

Results

Learning phase performance

In Session 1, the proportion correct increased with each repetition of the items during the learning phase. A paired-samples *t*-test confirmed that recall in repetition 2 of the learning phase was higher than in repetition 1, $t(147) = 27.78$, $p < .001$, $d = 2.29$ [1.98, 2.59]. However, the proportion correct during the learning phase (Repetition 1: $M = .39$, $SD = .19$; Repetition 2: $M = .62$, $SD = .19$) was lower than it was in the previous experiments and lower than the level recommended by Rowland (2014).

Final test performance

Descriptive statistics are reported in Table 6, and reliabilities are reported in the Supplemental Materials. Performance was highest in the repeated retrieval condition and lowest in the no practice condition, with restudy performance in between these conditions. Consistent with Experiments 1 and 2, a reliable retrieval practice effect was observed.

A repeated-measures ANOVA with condition as a within-subjects variable confirmed these results and indicated a significant main effect of learning condition, $F(2, 294) = 550.51$, $MSE = .01$, $p < .001$, $\eta_p^2 = .79$. Subsequent pairwise comparisons showed that performance was better in the retrieval practice condition relative to the restudy condition, $t(147) = 11.82$, $p < .001$, $d = .95$ [.75, 1.14], and the no practice condition, $t(147) = 33.83$, $p < .001$, $d = 2.68$ [2.34, 3.03]. The restudy condition outperformed the no practice condition, $t(147) = 19.64$, $p < .001$, $d = 1.66$ [1.41, 1.91].

Associations with working memory

The mean operation span scores using both the partial and absolute scoring methods were consistent with the normative data of Redick et al. (2012) and Experiments 1 and 2.

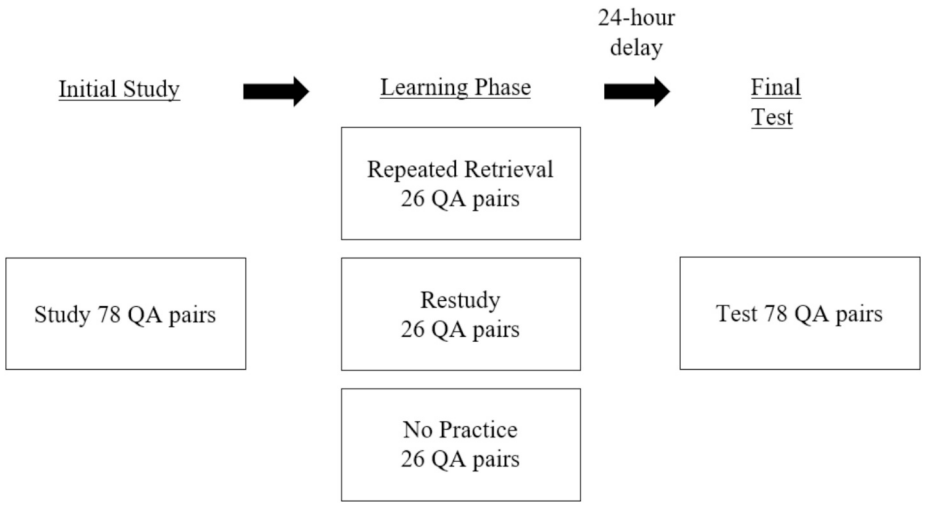

Fig. 7. Procedure for the general knowledge task. QA: Question-Answer.

Table 6
Experiment 3 descriptive statistics.

Variables	Mean	SD	Min	Max	Skew	Kurtosis
Repeated Retrieval	.66	.20	.27	1.00	−0.11	−1.00
Restudy	.53	.22	.12	1.00	0.22	−1.04
No Practice	.26	.16	.00	.73	0.73	0.24
RP Difference Score	.13	.14	−.19	.46	0.07	−0.44
Operation Span Absolute	45.44	17.82	0.00	75.00	−0.66	−0.17
Operation Span Partial	59.26	13.89	9.00	75.00	−1.72	2.87

Note: RP Difference Score: Repeated Retrieval – Restudy; Min: Minimum; Max: Maximum.

Learning Phase Performance. To examine whether learning phase performance differed as a function of WM ability, we conducted an ANCOVA with repetition as a within-subjects variable and operation span score (partial scoring) as a continuous covariate. There was a significant main effect of repetition, $F(1, 146) = 772.05$, $MSE = .01$, $p < .001$, $\eta_p^2 = .84$, reflecting that performance was better in the second repetition compared to the first. There was also a significant main effect of WM, $F(1, 146) = 4.65$, $MSE = .07$, $p = .033$, $\eta_p^2 = .03$, as individuals with higher operation span scores performed better during the learning phase. The interaction between repetition and WM was not significant, $F(1, 146) = 1.11$, $MSE = .01$, $p = .294$, $\eta_p^2 = .01$.

Final Test Performance. We conducted an ANCOVA with learning condition (repeated retrieval, restudy, and no practice) as a within-subjects variable and operation span score as a continuous covariate. Reported first is the analysis utilizing the partial scoring method. Mirroring the ANOVA results above, there was a significant main effect of learning condition, $F(2, 292) = 16.46$, $MSE = .01$, $p < .001$, $\eta_p^2 = .10$. The main effect of WM ability was significant, $F(1, 146) = 8.86$, $MSE = .09$, $p = .003$, $\eta_p^2 = .06$, indicating that individuals with high WM performed better than individuals with low WM. Additionally, the interaction between learning condition and WM was significant, $F(2, 292) = 3.05$, $MSE = .01$, $p = .049$, $\eta_p^2 = .02$. However, the final test performance difference between individuals with high and low WM was numerically largest in the no practice condition, not the repeated retrieval or restudy conditions. Therefore, to further understand this significant interaction, we conducted the same analysis excluding the no practice condition and including only the repeated retrieval and restudy conditions. In this analysis, the learning condition and WM interaction was not significant, $F(1, 146) = 0.48$, $MSE = .01$, $p = .490$, $\eta_p^2 < .01$, indicating that performance in the no practice condition was likely driving the significant interaction in the prior analysis.

Next, we conducted another ANCOVA with learning condition (repeated retrieval, restudy, and no practice) as a within-subjects variable and operation span score as a continuous covariate calculated using the absolute scoring method. The main effect of learning condition was again significant, $F(2, 292) = 57.43$, $MSE = .01$, $p < .001$, $\eta_p^2 = .28$, as well as the main effect of WM, $F(1, 146) = 12.50$, $MSE = .09$, $p < .001$, $\eta_p^2 = .08$. However, the interaction between learning condition and WM was not significant, $F(2, 292) = 2.07$, $MSE = .01$, $p = .128$, $\eta_p^2 = .01$.[4]

To further examine the relationship between the retrieval practice effect and WM, we computed the correlation between the retrieval practice difference score (performance in the repeated retrieval condition minus performance in the restudy condition) and the operation span score. Visualization of this analysis is shown in Fig. 8, and a correlation matrix for all variables of interest can be found in Table 7. Reported first is the analysis utilizing the partial scoring method. The magnitude of the retrieval practice effect did not differ as a function of WM ability. The correlation was not significant, $r(146) = -.06$, $p = .490$, CI [−.22,.11]. Using JASP, this non-significant correlation had a $BF_{01} = 7.68$, which is classified as 'moderate evidence' in favor of the null hypothesis (Wagenmakers et al., 2018). Similarly, the correlation was not significant utilizing the absolute scoring method, $r(146) = -.07$, $p = .430$, CI [−.22,.10]. This correlation had a $BF_{01} = 7.23$, which is classified as 'moderate evidence' in favor of the null hypothesis (Wagenmakers et al., 2018). Following this analysis, we used a Hotelling-Williams *t*-test to compare the two correlations and observed no significant difference between the two scoring methods, $t(145) = -0.24$, $p = .814$.

We also computed the correlation between performance in the repeated retrieval condition and operation span scores, partialing out performance in the restudy condition. Consistent with the other methods of analyzing the relationship between WM and retrieval practice, the partial correlation was not significant, $pr(145) = .07$, $p = .413$, CI [−.11,.25].

Discussion

The results of Experiment 3 were consistent with the results of Experiments 1 and 2. With the procedure and general knowledge materials

[4] Given the finding that learning phase proportion correct was lower than expected, we conducted an additional ANCOVA analysis with learning phase proportion correct added as a covariate. Results are reported in the supplemental materials section.

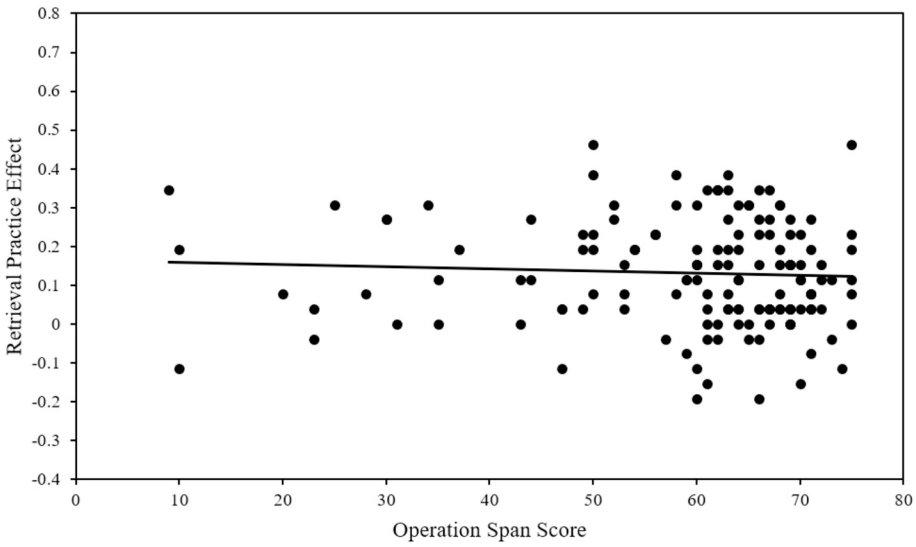

Fig. 8. Magnitude of the retrieval practice effect (*repeated retrieval* − *restudy*) as a function of operation span score (partial scoring) in Experiment 3.

Table 7
Experiment 3 correlation matrix.

Variables	1	2	3	4	5	6
1. Repeated Retrieval	–					
2. Restudy	**.79**	–				
3. No Practice	**.67**	**.67**	–			
4. RP Difference Score	.16	**−.47**	−.13	–		
5. Operation Span Absolute	**.26**	**.27**	**.22**	−.07	–	
6. Operation Span Partial	**.23**	**.24**	.15	−.06	**.92**	–

Note: RP Difference Score: Repeated Retrieval – Restudy. Bolded values indicate a significant correlation at the $p < .05$.

used by Agarwal et al. (2017), practicing retrieval enhanced performance on a delayed final test relative to repeated studying. Most importantly, the magnitude of this retrieval practice effect did not differ between individuals with high and low WM ability. This finding is consistent with the bulk of prior studies reviewed earlier and with the results of Experiments 1 and 2, though it is inconsistent with the results of Agarwal et al. (2017), despite the use of the same materials and procedure. The likely explanation for why the results of Experiment 3 differed from the results obtained by Agarwal et al. is due to differences in scoring methods and sample compositions. In Agarwal et al. (2017), using the absolute scoring method, the overall sample scored 15 points higher on operation span than our Experiment 3 sample, and was more restricted in range. In the current study, WM scores, on average, aligned with normative values (Redick et al., 2012) and showed variation across the possible range of scores (Fig. 8).

General discussion

The current study replicated the retrieval practice effect using a paired-associates task and a general knowledge task. In all three experiments, retrieval practice produced better final test performance than did repeated studying (Fig. 9). In all three experiments, although there were substantial individual differences in the magnitude of the retrieval practice effect (Figs. 3, 6, and 8), there was no association with performance on measures of WM ability, using multiple analytic approaches. The lack of a relationship between WM ability and the retrieval practice effect is consistent with several previous studies on this topic (Table 1).

One source of variation that had been previously overlooked in past studies investigating the association between individual differences in WM and retrieval practice is initial recall success. Novel to the current study, we implemented a learning-to-criterion procedure in Experiments 1 and 2 to address this issue. Initial retrieval success is a critical factor for obtaining retrieval practice effects (Karpicke, 2017; Rowland, 2014).

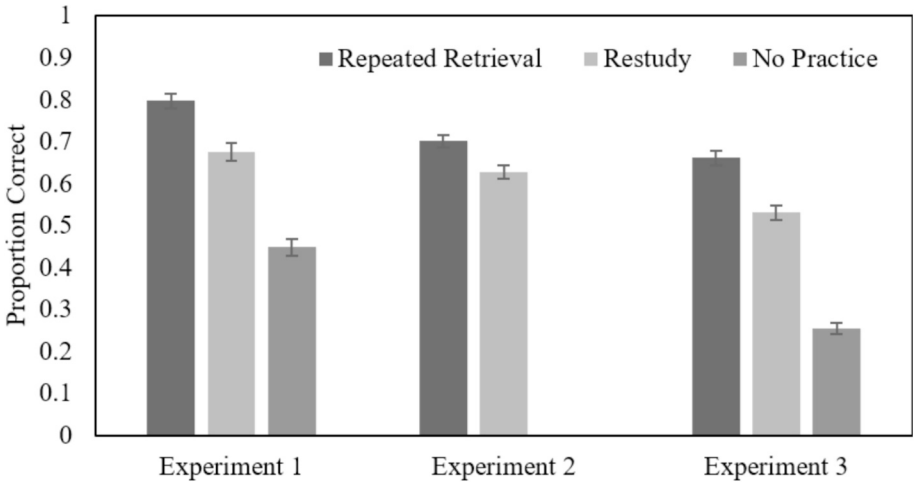

Fig. 9. Final test performance by condition for Experiments 1–3.

When taking WM ability into account, utilizing measures to equate initial retrieval success may be especially important. Past research has provided evidence that individuals with high WM ability perform better on episodic memory tasks like those used in typical retrieval practice studies (Rosen & Engle, 1997; Martin et al., 2020; Unsworth, 2007; Unsworth, 2009a). It is reasonable to expect that high WM ability individuals would have greater levels of initial retrieval success relative to lower WM ability individuals, and this difference would impact any retrieval practice effect.

This difference in retrieval success was indeed demonstrated in Experiments 1 and 3 in this report. During the learning-to-criterion phase in Experiment 1, individuals with high WM recalled more items than individuals with low WM on early trials. Similarly, in the learning phase in Experiment 3, initial success was better for individuals with high WM. Findings by Nordstrand et al. (2016), Agarwal et al. (2017), and Conway and Engle (1994) also support this idea. Nordstrand et al. found that individuals with high WM improved more during the initial learning phase of a paired-associates task relative to individuals with low WM. Similarly, Agarwal et al. reported that WM was positively correlated with initial retrieval success ($r = .31, p < .001$). Conway and Engle found that low WM individuals needed more cycles to reach criterion than did high WM individuals. Critically, the learning-to-criterion method used in Experiments 1 and 2 was an effective way to equate initial retrieval success prior to the retrieval practice/restudy manipulation. In both experiments, by the final trial of the learning-to-criterion phase, there was no difference in recall success as a function of WM. Therefore, we can be more confident that retrieval practice effects observed in these experiments reflect the effects of retrieval, not differences in learning rate or initial retrieval success across subjects with varying WM ability. This reiterates the advantage of using a learning-to-criterion method when investigating individual differences in retrieval practice effects.

Previous individual differences in WM studies have also used different types of learning tasks to examine retrieval practice. The importance of considering the impact of alternative tasks is highlighted by Jenkins (1978), who describes a tetrahedral model of memory in which results of a given experiment may depend on the interaction between different variables. These variables include the subjects, the criterial task, the materials, and the orienting task of an experiment. As an example of a relevant subject-criterial task interaction, Unsworth (2009b) provided evidence that the relationship between recall performance and individual differences in WM varied depending on the amount of self-initiated processing (Craik, 1983) required by the task. Tasks that involved more self-initiated processing, such as free recall, were more strongly correlated with measures of cognitive ability (WM, fluid intelligence). Our use of a paired-associates task (Experiments 1 and 2) and general knowledge task (Experiment 3) were directly influenced by their use in the retrieval practice literature (Table 1) and their applicability to learning activities students often engage in, but they are variations of a cued-recall task. However, one speculation is that individual differences in WM may be more likely to modulate the retrieval practice effect when the learning activity demands more self-initiated processing, as in free recall. Other characteristics of a task may also be influential when considering differences in cognitive ability. For example, individuals with low WM have been shown to be more susceptible to buildup of proactive interference in a task relative to individuals with high WM (Kane & Engle, 2000). These ideas would be consistent with the findings of Yang et al. (2020), who used a free recall paradigm designed to increase proactive interference across multiple lists and observed a significant relationship between WM and the retrieval practice effect.

Other factors we have highlighted in the current study that may have influenced the mixed results reported in past research involve differences in WM scoring and the ability range of the sample. The present experiments addressed these issues by conducting analyses utilizing both partial and absolute scoring. We observed no difference between the two scoring methods. However, in contrast to almost all previous studies (see Table 1), the ability range of our sample was consistent with the normative data of Redick et al. (2012). Scoring methods for the complex span tasks may have greater influence on samples clustered at high or low ends of the WM distribution (for discussion and an example, see Unsworth & Engle, 2007a). These differences in the sample distribution across studies complicate generalizability of results and may underlie the variability in correlation strength observed in past research on this topic.

Finally, the choice of how to assess the association between WM and the retrieval practice effect may also be influential for the observed relationship. Past research has cautioned against the use of difference scores in individual differences research due to low reliability (for recent review, see Draheim et al., 2019). Indeed, reliabilities for the retrieval practice difference score were low across all three experiments of the current study, consistent with prior research (e.g., Brewer & Unsworth, 2012; Minear et al., 2018). Additionally, as discussed previously, there are known problems interpreting difference scores that are only exacerbated when reliability is low (Cohen et al., 2002). Our results using ANCOVAs and partial correlation analyses showed the same results across all three studies: there was no relationship between WM and the retrieval practice effect, indicating the benefits of retrieval practice were similar across levels of WM ability.

Several theories have been proposed to explain the benefit to memory following the use of retrieval practice. For example, the *episodic context account* (Karpicke, Lehman, & Aue, 2014) proposes that retrieval reinstates the previous context associated with an item. The item is then updated with additional features of the current context. Subsequently, this enhanced context representation enables a more efficient search of memory by decreasing the number of possible retrieval paths. In contrast, the *elaborative retrieval hypothesis* proposes that the benefit of retrieval practice is due to the creation of additional retrieval paths through the process of elaboration (Carpenter, 2009). When recalling information, additional cues semantically related to the target are activated and can be used to aid in retrieval. Another theory with implications for the current research stems from the idea that retrieval requires more effort than other study strategies, such as rereading. Bjork and Bjork (2011) refers to this as a *desirable difficulty*, where conditions that induce increased effort lead to better long-term retention. None of these theories are specific about the impact of WM on the proposed mechanisms. However, many theories of WM describe long-term memory as a component of WM (for brief review see Cowan, 2017). Additionally, it has been proposed that high WM individuals use more effective retrieval cues to guide memory search processes (Unsworth & Engle, 2007b). This includes both temporal-contextual retrieval cues (Spillers & Unsworth, 2011; Unsworth & Engle, 2006) and semantic retrieval cues (Rosen & Engle, 1997; Unsworth, Brewer, & Spillers, 2013). As mentioned previously, prior research has also suggested WM and retrieval practice influence cognitive processes such as narrowing of a search set during recall and reducing proactive interference (e.g., Kane & Engle, 2000; Lehman et al., 2014; Szpunar et al., 2008; Unsworth, 2019). While the current study was not designed to examine these particular mechanisms, and is therefore limited in our ability to provide evidence for or against any particular account of retrieval practice, future research investigating this research question may wish to examine these cognitive processes in more depth by using tasks that may vary the level of cue support or interference and/or by conducting analyses of retrieval dynamics in addition to retrieval accuracy.

Finally, another limitation of the current study was that we only examined WM. Other measures of cognitive ability such as fluid intelligence, episodic memory, and attentional control may play an important role in retrieval. For example, Brewer and Unsworth (2012) observed a larger benefit of retrieval practice for individuals with low episodic memory ability and low fluid intelligence, but not WM. Minear et al. (2018) also reported a differential benefit of retrieval practice for individuals with high and low fluid intelligence. However, Minear and colleagues observed a significant interaction between fluid intelligence

and item difficulty. Individuals with high fluid intelligence showed a larger retrieval practice effect for difficult items while individuals with low fluid intelligence showed a larger retrieval practice effect for easy items. Future research may benefit from examining the role of various cognitive abilities in retrieval as well as examining the effect of different factors such as item difficulty, retention interval, or feedback presence. Additionally, this line of research may be particularly relevant for atypical populations which have shown deficits in some of these cognitive processes (see Knouse et al., 2016; Leonard et al., 2020; and Levlin et al., 2022 for examples on the effects of retrieval in clinical populations).

Conclusion

The current study provides further evidence to the literature that the benefit gained from retrieval practice is independent of WM ability. This result was observed after controlling for differences in initial retrieval success through the use of a learning-to-criterion procedure in Experiments 1 and 2. Additionally, this result was replicated utilizing a paired-associates task and a general knowledge task. From an applied standpoint, this result adds to the extensive literature implicating retrieval practice as an effective study strategy to be used in the classroom. It is important for instructors to know whether study strategies such as retrieval practice produce benefits broadly across learners, or whether some benefit more than others, and whether such differences are related to cognitive abilities. The results of the current study suggest that benefits of retrieval practice are not related to WM, consistent with the overall picture of the literature. Independent of the WM ability of the student, retrieval practice is a beneficial strategy for increasing retention of learned material.

CRediT authorship contribution statement

Andy L. Fordyce: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thomas S. Redick:** Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Joseph P. Bedwell:** Writing – review & editing, Supervision, Software, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Jeffrey D. Karpicke:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

We thank Joshua Whiffen for his contributions to Experiment 1 and Pooja Agarwal for providing the materials for the general knowledge task used in the current study. Additionally, we thank Chunliang Yang and Bert Jonsson for providing information on the data for their previous studies. Finally, we thank the Karpicke Cognition and Learning Lab members for assistance with Experiments 1 & 2 data collection as well as Ashley Egler, Larissa Olivas, Lindsey Fishman, Bradley Tonjes, and Emily Sanders for assistance with Experiment 3 data collection. While working on this manuscript, TSR was supported by Office of Naval Research grant N000142312768.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104664.

Data availability

Materials, data, and analysis files for this study are available on OSF (https://osf.io/7jf9m/?view_only = f1317737f1a9430f96e4bc057066780b).

References

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

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., III (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory, 25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval practice: Beneficial for all students or moderated by individual differences? *Psychology, Learning, & Teaching, 20*(1), 21–39. https://doi.org/10.1177/14757257209734 94

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society, 2*(59–68).

Brewer, G. A., Robey, A., & Unsworth, N. (2021). Discrepant findings on the relation between episodic memory and retrieval practice: The impact of analysis decisions. *Journal of Memory and Language, 116*, Article 104185. https://doi.org/10.1016/j.jml.2020.104185

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*, 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*, 1563–1569. https://doi.org/10.1037/a0017021

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Conway, A. R., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology. General, 123*(4), 354–373.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769–786. https://doi.org/10.3758/BF03196772

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*, 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society, London, Series B, 302*, 341–359.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*(5), 508–535. https://doi.org/10.1037/bul0000192

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14*(1), 4–58.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). New York: Elsevier.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods, 37*, 581–590. https://doi.org/10.3758/BF03192728

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Education & Psychology, 106*(1), 58–68. https://doi.org/10.1037/a0033208

Harrison, T. L., Whiffen, J. W., Ware, T. M., & Engle, R. W. (2012, November). Is working memory capacity related to the magnitude of the test effect?. Poster given at the annual meeting of the Psychonomic Society, Minneapolis, MN.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*(4), 597–606. https://doi.org/10.1002/acp.3032

Jenkins, J. J. (1978). Four points to remember: A tetrahedral model of memory experiments. *Levels of processing in human memory*, 429–446.

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability.

Journal of Education & Psychology, 113(5), 972–985. https://doi.org/10.1037/edu0000627

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*, 336–358. https://doi.org/10.1037/0278-7393.26.2.336

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology. General, 138*(4), 469–486. https://doi.org/10.1037/a0017341

Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. in J. H. Byrne & J. T. Wixted (Eds.), Cognitive psychology of memory. Learning and memory: a comprehensive reference (Vol. 2, pp. 487-514). Oxford: Academic Press. Doi: 10.1016/B978-0-12-809324-5.21055-9.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego, CA: Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1.

Karpicke, J. D., & Roediger, H. L. (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

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(2012), 17–29. https://doi.org/10.1016/j.jml.2012.02.004

Knouse, L. E., Rawson, K. A., Vaughn, K. E., & Dunlosky, J. (2016). Does testing improve learning for college students with attention-deficit/hyperactivity disorder? *Clinical Psychological Science: A Journal of the Association for Psychological Science, 4*(1), 136–143.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*(6), 1787. https://doi.org/10.1037/xlm0000012

Leonard, L. B., Deevy, P., Karpicke, J. D., Christ, S. L., & Kueser, J. B. (2020). After initial retrieval practice, more retrieval produces better retention than more study in the word learning of children with developmental language disorder. *Journal of Speech, Language, and Hearing Research, 63*, 2763–2776.

Levlin, M., Wiklund-Hörnqvist, C., Sandgren, O., Karlsson, S., & Jonsson, B. (2022). Evaluating the effect of rich vocabulary instruction and retrieval practice on the classroom vocabulary skills of children with (developmental) language disorder. *Language, Speech, and Hearing Services in Schools, 53*(2), 542–560.

Martin, J. D., Shipstead, Z., Harrison, T. L., Redick, T. S., Bunting, M., & Engle, R. W. (2020). The role of maintenance and disengagement in predicting reading comprehension and vocabulary learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 46*(1), 140–154. https://doi.org/10.1037/xlm0000705

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annu Rev in Psych Sci, 72*, 609–633.

Minear, M., Coane, J. H., Boland, S. C., & Cooney, L. H. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*(3), 338–368. https://doi.org/10.1016/S0022-5371(80)90266-2

Nordstrand, D., Hansson, P., & Wiklund-Hörnqvist, C. (2016). *Test enhanced learning, working memory, and fluid intelligence*. Umea University. Bachelor's Thesis.

Poole, B. J., & Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention. *The Quarterly Journal of Experimental Psychology, 62*(7), 1430–1454.

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

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review, 20*(6), 1102–1113. https://doi.org/10.3758/s13423-013-0453-9

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology. General, 145*(11), 1473–1492.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335–343.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology. General, 126*(3), 211–227. https://doi.org/10.1037/0096-3445.126.3.211

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language, 39*(3), 418–436. https://doi.org/10.1006/jmla.1998.2590

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

Salthouse, T. A., Siedlecki, K. L., & Krueger, L. E. (2006). An individual differences analysis of memory control. *Journal of Memory and Language, 55*(1), 102–125. https://doi.org/10.1016/j.jml.2006.03.006

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*(4), 628–654.

Spillers, G. J., & Unsworth, N. (2011). Variation in working memory capacity and temporal–contextual retrieval from episodic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37*(6), 1532–1539.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(6), 1392. https://doi.org/10.1037/a0013082

Tse, C., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology. Applied, 18*(3), 253–264. https://doi.org/10.1037/a0029190

Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*, 1020–1034. https://doi.org/10.1037/0278-7393.33.6.1020

Unsworth, N. (2009a). Examining variation in working memory capacity and retrieval in cued recall. *Memory, 17*, 386–396. https://doi.org/10.1080/09658210902802959

Unsworth, N. (2009b). Individual differences in self-initiated processing at encoding and retrieval: A latent variable analysis. *The Quarterly Journal of Experimental Psychology, 62*(2), 257–266. https://doi.org/10.1080/17470210802373092

Unsworth, N. (2010). Interference control, working memory capacity, and cognitive abilities: A latent variable analysis. *Intelligence, 38*(2), 255–267. https://doi.org/10.1016/j.intell.2009.12.003

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin, 145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition, 41*, 242–254.

Unsworth, N., & Engle, R. W. (2006). A temporal–contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language, 54*(3), 346–362.

Unsworth, N., & Engle, R. W. (2007a). On the division of short-term and working memory: An examination of simple and complex spans and their relation to higher-order abilities. *Psychological Bulletin, 133*, 1038–1066. https://doi.org/10.1037/0033-2909.133.6.1038

Unsworth, N., & Engle, R. W. (2007b). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505. https://doi.org/10.3758/BF03192720

Unsworth, N., & Redick, T. S. (2017). Working memory and intelligence. In J. Wixted (Ed.), Cognitive psychology of memory, Vol. 2 of learning and memory: a comprehensive reference, 2nd edition, Byrne, J. H. (ed.). pp. 163-180. Oxford: Academic Press. Doi: 10.1016/B978-0-12-809324-5.21041-9.

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

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wiemers, E. A., & Redick, T. S. (2019). Task manipulation effects on the relationship between working memory and go/no-go task performance. *Consciousness and Cognition, 71*, 39–58. https://doi.org/10.1016/j.concog.2019.03.006

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology, 55*, 10–16. https://doi.org/10.1111/sjop.12093

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of Experimental Psychology. Applied, 26*(4), 724–738. https://doi.org/10.1037/xap0000278