Journal of Memory and Language 142 (2025) 104608

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Do individual differences in working memory capacity, episodic memory ability, or fluid intelligence moderate the pretesting effect?☆

Steven C. Pan *, Liwen Yu, Marcus J. Wong, Ganeash Selvarajan, Andy Z.J. Teo

Department of Psychology, National University of Singapore, 9 Arts Link, Singapore City 117572, Singapore

ARTICLE INFO

Keywords:
Pretesting
Errorful generation
Working memory capacity
Episodic memory ability
Fluid intelligence
Individual differences

ABSTRACT

The *pretesting effect* refers to the finding that guessing the answers to test questions before learning the correct answers improves memory relative to studying (or reading) without prior guessing. Although the pretesting effect is robust and has been demonstrated across multiple studies, its magnitude varies across individuals. Two studies investigated whether individual differences in working memory capacity (WMC), episodic memory ability (EM), and/or fluid intelligence (gF) help explain that variation. In Study 1, lower gF scores were associated with a larger pretesting effect among undergraduate students, stemming from lower performance on read items. In Study 2, involving adult online participants, observed patterns were less consistent, but lower WMC scores were associated with larger pretesting effects, again due to lower performance on read items. Together, these patterns suggest that pretesting can homologize memory ability across individuals, although to an extent that may vary across learner populations and cognitive abilities. That conclusion and other findings are interpreted in the context of relevant individual differences research and theories related to pretesting and memory phenomena.

Introduction

Practice testing improves memory. Over a century of research and hundreds of studies have shown that attempting to recall information *after* it has been studied, or retrieval practice, benefits long-term memory retention of that information (Roediger & Butler, 2011; Rowland, 2014; see also Pan et al., 2024). More recently, a smaller but growing literature has shown that taking practice tests *before* information has been studied, a strategy known as *pretesting* (and sometimes errorful generation, failed testing, or unsuccessful testing), can also enhance long-term memory (e.g., Kornell et al., 2009; for reviews see Kornell & Vaughn, 2016; Metcalfe, 2017; Pan & Carpenter, 2023). Given the lack of a prior study episode, participants' guesses to pretest questions are mostly or entirely incorrect, leaving the actual answers to be learned via study of correct answer feedback (which is usually provided after each guessing attempt). Relative to non-testing conditions such as reading or studying, pretesting improves memory for stimulus materials such as paired associate words (e.g., Knight et al., 2012; Vaughn & Rawson, 2012), word triplets (e.g., Metcalfe & Huelser, 2020), trivia facts (e.g., Kornell et al., 2009), and expository texts (e.g., Pan & Sana, 2021; Richland et al., 2009). That memorial benefit is known as the *pretesting effect.*

The pretesting effect is typically demonstrated using a three-phase experimental design that features two training conditions: pretesting and reading (or studying). That design is depicted in Fig. 1. The training phase occurs first. During that phase, in the pretested condition, participants attempt to guess the answers to pretest questions (e.g., *chop –???*) and then receive immediate correct answer feedback (e.g., *chop – dice*). In the read condition, participants view intact, correct information and do not make any guesses at all (e.g., *moss – grass*). Next, participants undergo a distractor task (or retention interval) phase, wherein they engage in unrelated activities for a period of time, usually for at least several minutes. Finally, in the criterial test phase, participants take a test on all previously learned items (e.g., *chop –???*; *moss –???*). On that criterial test, it is common for performance to be better for previously pretested rather than previously read items.

☆ 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 Psychology, Faculty of Arts and Social Sciences, National University of Singapore. Mailing address: 9 Arts Link, Singapore City 117572, Singapore.
E-mail address: scp@nus.edu.sg (S.C. Pan).

https://doi.org/10.1016/j.jml.2025.104608
Received 1 April 2024; Received in revised form 22 December 2024; Accepted 2 January 2025
Available online 10 January 2025

Theoretical accounts of the pretesting effect

Several theoretical accounts have been proposed to explain the pretesting effect (for discussions, see Mera et al., 2021; Metcalfe, 2017; Pan & Carpenter, 2023; see also Kornell & Vaughn, 2016). The most prominent mechanistic accounts address the pretesting effect in the context of paired-associate learning and were adapted from theories of the retrieval practice effect or other memory phenomena. The semantic mediator or mediation hypothesis, for example, posits that the erroneous guesses that participants make during pretest trials act as mediators between originally presented cues and subsequently presented targets (Huelser & Metcalfe, 2012; Kornell et al., 2009). Those mediators support improved criterial test performance for pretested items. Search set theory, on the other hand, posits that pretest questions activate sets of potential candidate answers in memory, and when correct answer feedback is presented, encoding of the correct answer with the relevant cue is enhanced (Grimaldi & Karpicke, 2012) and incorrect retrieval routes may also be suppressed (Kornell et al., 2009). The net result, again, is enhanced retrievability of correct targets on a criterial test. The recursive reminding theory posits that the experience of generating an error and learning the answer through correct answer feedback are encoded in the same episodic event (Wahlheim & Jacoby, 2013; see also Mera et al., 2021; Jacoby & Wahlheim, 2013). On a criterial test, if participants are able to recall that event, then successful retrieval of the correct response should follow. Finally, the prediction error or error correction account suggests that awareness of incorrect responding triggers an error signal, leading to enhanced attention and better learning of the correct response, resulting in enhanced memory for pretested items (Kang et al., 2011; although see Seabrooke et al., 2021).

It should be noted that the aforementioned theoretical accounts are not exhaustive and more general accounts exist (e.g., metacognitive explanations as in Carpenter & Toftness, 2017; increased attention as in Pan et al., 2020). For instance, Kornell et al. (2015) proposed a two-stage framework wherein learners first attempt to retrieve the correct answer from memory, and failing that, learn the answer through correct answer feedback. In that framework, it is assumed that engaging in the first stage yields more learning in the second stage, but the precise mechanisms involved are not specified. Moreover, theoretical development pertaining to the pretesting effect remains a work in progress. One important early finding is that the pretesting effect for paired associates (as assessed on a cued recall test) manifests with semantically associated but not unassociated word pairs (e.g., Huelser & Metcalfe, 2012; Knight et al., 2012). Semantic associations in the learning materials may be a precondition for mediator generation or the formation of an appropriate search set. More recently, Metcalfe and Huelser (2020) found equally strong pretesting effects for word triplets that had consistent semantic associations (e.g., *wrist-palm-hand*) or inconsistent associations (e.g., *tree-palm-hand*) and only in cases where participants remembered their original guess. Mediator generation, it was theorized, was only feasible in the case of consistently associated word triplets, and difficult if not impossible for inconsistently associated triplets. That research suggests that although making guesses on a pretest involves semantic memory, as is widely assumed (e.g., Knight et al., 2012; Kornell et al., 2009; Pan et al., 2019), and semantic mediators may be formed (e.g., Leonard et al., 2023), the pretesting effect is dependent on episodic memory processes.

Individual differences and the pretesting effect

Although very little research on individual differences and the pretesting effect has been conducted to date, research on the topic has the potential to add theoretical and practical insights. Such individual differences exist: For example, Pan and Rivers (2023) observed across five experiments that up to 30 % of adult online participants did not exhibit at least a numerical pretesting effect for paired associate words on an initial learning cycle—and in fact, up to 23 % of participants exhibited better performance in the read condition. In another example, Janelli and Lipnevich (2021) found that pretesting in an online course was associated with a greater likelihood of students dropping out of the course, with a pretesting effect only observed among the students that completed the entire course. Among all of the students that began the course, however, there was no pretesting effect. In addition, Cyr and Anderson (2012) investigated errorful guessing followed by feedback in younger versus older (approximately 70 years old) adults; on a subsequent recognition test, they observed larger magnitude effects of such pretesting for older adults (although see Cyr & Anderson, 2015). They speculated that engaging in errorful guessing could have prompted older adults to engage in greater-than-typical amounts of semantic elaboration for their age group, yielding criterial test performance that was on par with that of younger adults.

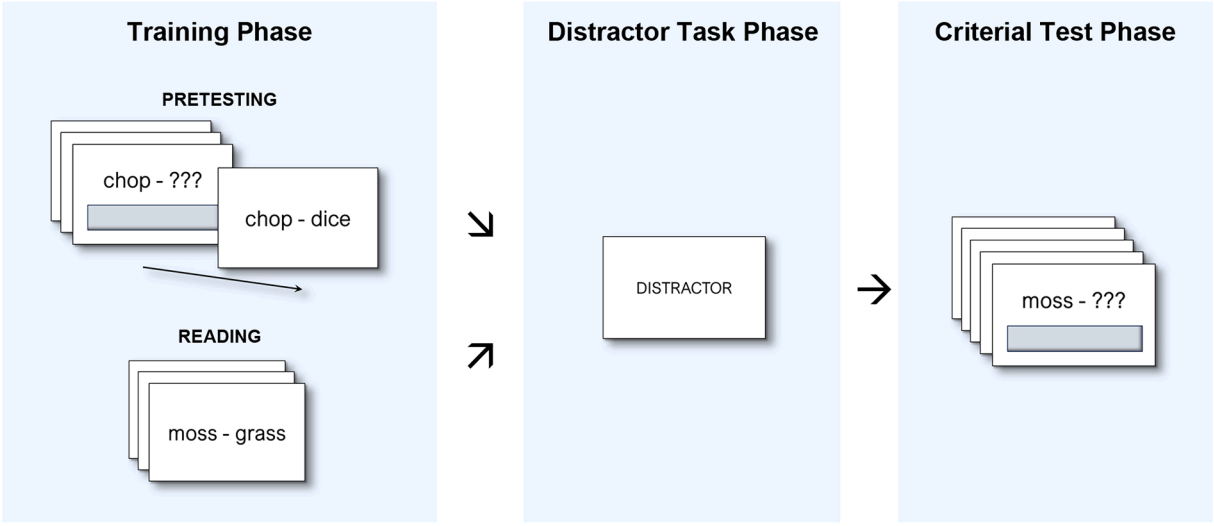

Fig. 1. *Schematic of the Pretesting Task. Note.* All participants completed three phases. First, they learned 32 word pairs, half via pretesting (5 s guessing plus 5 s correct answer feedback) and half via reading (5 s each). Pretested and read pairs were randomly intermixed. Next, they spent 5 min in an unrelated activity (in the present research, the number series task). The final phase was the criterial test in which memory for the target word of all 32 word pairs was assessed individually in a self-paced manner with no feedback, one word pair at a time. The pretesting effect entailed the performance difference for pretested versus read items on the criterial test.

When one considers the steps that are involved in generating a pretesting effect, the potential roles of well-established cognitive abilities—and consequently, effects of individual differences in those abilities—begin to emerge. The act of reading a pretest question and formulating a reasonable guess, for example, may be affected not just by semantic memory processes but also by working memory capacity (WMC), which entails the ability to focus on goal-relevant memories (Engle & Kane, 2003), as well as fluid intelligence (gF), involving the ability to solve novel or abstract problems (Kievit et al., 2016). WMC is presumably necessary to help keep pretest questions and guesses in mind, whereas gF could influence the types of strategies that participants use when answering pretest questions and during the training phase more generally. Further, given that episodic memory appears to play a role in the pretesting effect, as suggested by Wahlheim and Jacoby's (2013) recursive reminding account and Metcalfe and Huelser's (2020) results, individual differences in episodic memory (EM) ability, or the extent to which one can remember prior experiences, may also influence the pretesting effect. For instance, if memory for erroneous guesses is necessary, then EM ability could be a moderator of the pretesting effect, with stronger abilities possibly leading to a larger pretesting effect.

Given the possibilities outlined above, it is plausible that individual differences in WMC, EM ability, and/or gF might explain variations in pretesting effect magnitude among learners. The fact that WMC and gF commonly decline with age (Murman, 2015) may also be relevant for interpreting Cyr and Anderson's (2012) finding of a larger pretesting effect for older adults. At a broader level, there is an abundance of evidence showing that individual differences in WMC predict cognitive abilities such as reading ability, vocabulary learning, and performance on standardized tests (Engle & Kane, 2003; Unsworth, 2019). WMC is strongly correlated with gF (Kyllonen & Christal, 1990; Shipstead et al., 2016) and there is some evidence that individual differences in EM ability moderate the effects of such test-based learning strategies as retrieval practice (Unsworth, 2019). Whether similar relationships exist between those cognitive abilities and the pretesting effect, however, remains to be determined.

The current study

The current investigation examined the potential role of individual differences in three well-established cognitive abilities—namely working memory capacity, episodic memory ability, and fluid intelligence—on the pretesting effect. It included two studies of nearly identical design that sampled from different populations: undergraduate students at a large public research university (Study 1) and adult participants of a broader age range and different educational backgrounds sampled online (Study 2). All participants underwent a pretesting effect task involving paired associate words that has been used repeatedly in prior research. Moreover, similar to prior research (see Engle & Kane, 2003; Unsworth, 2019; Wingert & Brewer, 2018, and others), we used three tasks each to measure WMC, EM ability, and gF.

Prior to conducting either study, we considered three hypotheses involving the relation of WMC, EM ability, and gF to the pretesting effect. In evaluating these hypotheses, it is important to emphasize that the pretesting effect is typically calculated as a difference score involving performance in a pretested condition versus a read (or study) condition. Hence, although we were interested in the effects of cognitive abilities on learning processes that occur during the act of pretesting, the effects of those abilities on processes that occur in the read condition should also be considered. The three hypotheses were inspired by Brewer and Unsworth (2012; see also Unsworth, 2019), who conducted an analogous study in the context of the retrieval practice effect. The three hypothetical scenarios were:

1. *Pretesting enhances memory for all or nearly all learners to a similar degree.* That is, the investigated individual differences do not moderate the pretesting effect. Persons with higher or lower ability might have correspondingly higher or lower criterial test scores, but pretesting effect magnitude is unaffected (outside of, unavoidably, the most extreme ends of test performance; for discussion see Pan et al., 2015).
2. *Pretesting is more beneficial for learners with higher cognitive ability scores.* That is, pretesting is uniquely beneficial for individuals with higher WMC, EM ability, and/or gF scores. Brewer and Unsworth (2012) described this possibility as the "rich get richer" scenario (p. 409).
3. *Pretesting is more beneficial for learners with lower cognitive ability scores.* In such a scenario, pretesting helps individuals that might be expected to perform more poorly on the criterial test perform better, relatively speaking, on pretested items. Brewer and Unsworth (2012) described this possibility as that of testing "homologizing memory across the ability range" (p. 409).

To our knowledge, no published study to date has investigated these three hypotheses in the context of the pretesting effect. Given prior evidence of individual differences in the pretesting effect, however, we considered it unlikely that Hypothesis 1 would be supported, leaving Hypotheses 2 and 3 as more plausible. In evaluating those hypotheses, we considered prior evidence that two test-based learning strategies, retrieval practice and forward testing, may be more beneficial for individuals with lower ability scores (e.g., Agarwal et al., 2017; Brewer & Unsworth, 2012; Unsworth, 2019; Yang et al., 2020; see also Minear et al., 2018). If similar patterns apply to the case of pretesting, then Hypothesis 3 would be supported. Alternatively, if learning through pretesting is uniquely dependent on WMC, EM ability, or gF in a way that reading or studying is not, then Hypothesis 2 would be more likely.

We were also mindful of the recent history of individual differences research involving many of the same cognitive abilities and the retrieval practice effect (Brewer, Robey, & Unsworth, 2021; Brewer & Unsworth, 2012; Minear et al., 2018; Robey, 2019; Unsworth, 2019). In that literature, successive studies and data combined across studies has led to more refined conclusions. Hence, we expected that this study would constitute an initial foray into individual differences and pretesting, setting the stage for follow-up studies.

Study 1

The first study investigated the potential role of individual differences in cognitive abilities on the pretesting effect among an undergraduate student sample at a large public research university.

Data availability

Data and analysis code are accessible via the Open Science Framework (OSF) at https://osf.io/r6t95/. Where available to share, materials (excepting the working memory tasks, which were commercially sourced), are accessible via https://osf.io/8g3fp/.

Methods

Participants

The target sample size for both studies, 120, was comparable to that of some prior studies of individual differences and learning strategies (e.g., Brewer & Unsworth, 2012; Pan et al., 2015; Wang et al., 2020) and was feasible given available subject pool resources. A power analysis in G*Power 3.1 (Faul et al., 2007) indicated that a sample of at least 120 is needed for 80 % statistical power to detect a correlation of 0.29 or larger (one-tailed test) at $\alpha = 0.05$ (cf. Pan et al., 2015). Statistical power to detect smaller correlations, however, would be lower. One-hundred and thirty-five undergraduate students from the psychology subject pool at a large public research university in Singapore participated in exchange for course credit. All students were native or highly fluent English

speakers (English is the language of instruction at that university). Data from all participants were included in the analyses for this study, although one participant's image recognition task data were unusable due to technical problems (consequently, that participant's data were excluded from all of the mixed-effects models and supplementary analyses involving episodic memory ability but retained in other analyses).

In the final sample of 135 participants, the mean age was 20.3 years (ranging between 18 and 32 years) and 68 % were female; 75 %, 7 %, and 18 % were in the first, second, or third (or later) years of study, respectively. Just over half of the sample were Psychology majors (51 %), followed by Nursing majors (13 %), Life Sciences majors (9 %), and other types of academic majors. Approximately 74 % of the participants were of Chinese ancestry, 8 % were Malay, 7 % were Indian, and 11 % were of other ethnic groups. All data collection occurred with ethics approval obtained at the authors' university on July 11, 2023 (Protocol 2023-June-09). All participants provided informed consent and were treated in accordance with the principles set out in the Declaration of Helsinki.

Materials

Pretesting task. Materials for the pretesting task consisted of two lists of 16 weakly associated word pairs each (e.g., *chop – dice*) drawn from Huelser and Metcalfe (2012), for which statistically significant pretesting effects have previously been demonstrated. Each pair consisted of two words of at least 4 letters in length, with forward and backward associative strengths of 0.05–0.054 and 0, respectively.

Working memory tasks. The working memory tasks were versions of the operation span (Unsworth et al., 2005), symmetry span (Unsworth et al., 2009), and reading span (Unsworth et al., 2009) tasks that are commercially available via the Inquisit (Millisecond Software, 2022) online software platform. These tasks and their constituent materials (i.e., letters and math problems for operation span, visual sequences and images for symmetry span, and letters and sentences for reading span) were used without any changes.

Episodic memory tasks. Episodic memory ability was measured using delayed free recall, cued recall, and recognition tasks, all featuring materials that were previously featured in Brewer and Unsworth (2012). The delayed free recall task featured six lists of 10 unrelated words each, the cued recall task featured three lists of 10 unrelated word pairs each, and the image recognition task featured 60 drawings of various objects.

Fluid intelligence tasks. Fluid intelligence was measured using Raven's progressive matrices, number series, and letter sets tasks, all involving materials previously developed for those tasks. Raven's progressive matrices featured 18 test items (Raven & Raven, 2003), number series featured 15 test items (Ekstrom et al., 1976), and letter sets featured 20 test items (Thurstone, 1938).

Procedure

All participants completed a single session of approximately 90 min in length in a laboratory testing room. They did so via a desktop PC or a docked laptop computer equipped with the Google Chrome internet browser, Inquisit Web 5 software, and at individual laboratory testing cubicles or desks. After giving informed consent and answering several demographic questions, participants completed the initial learning phase of the pretesting task, the number series task, and the criterial test phase of the pretesting task. The pretesting task was placed at the outset of the study to avoid any potentially contaminating effects of test-potentiated learning from attenuating the pretesting effect (i.e., forward testing, wherein test experience causes participants to alter their encoding strategies and/or pay closer attention; Gupta et al., 2024). Moreover, the placement of the number series task in between the two phases of the pretesting task not only served as a measure of fluid intelligence but also acted as a distractor task during the 5-minute retention interval prior to the criterial test.

After the initial tasks, participants completed the three working memory tasks in the following order: operation span, symmetry span, and reading span. The remaining fluid intelligence tasks, namely Raven's progressive matrices and letter sets, followed. Then, participants completed the three episodic memory tasks in the following order: delayed free recall, cued recall, and image recognition. In between tasks, all participants were permitted to take brief breaks if they wished. After completing the image recognition task, they were debriefed and dismissed.

Tasks

The pretesting task was adapted from the test-enhanced learning literature and the individual differences measures were adapted from prior studies of individual differences, learning strategies, and/or human memory (e.g., Brewer & Unsworth, 2012; Chen et al., 2017; Robey, 2019). All but the working memory tasks, which involved Inquisit Web 5 (Inquisit 5, 2016), were programmed in Qualtrics (Qualtrics, Provo, UT).

For details on how participants' scores for the three tasks measuring working memory capacity, episodic memory ability, and fluid intelligence were combined to form three separate composite scores, as well as factor scores, please refer to the Score Transformations and Analysis Plan section.

Pretesting task. The pretesting task, which is depicted in Fig. 1, was patterned after that used in many pretesting effect studies such as Huelser and Metcalfe (2012; see also Knight et al., 2012; Kornell et al., 2009; Pan & Rivers, 2023) and had three phases: training phase, distractor task, and criterial test. The training phase involved the presentation of 32 word pairs across two lists, one pair at a time in random order. One list of pairs was assigned to the *read* condition, wherein each pair was presented intact for 5 s each, whereas the other list was assigned to the *pretested* condition, wherein participants were given 5 s to guess (and type) the target word in response to presentation of the cue word, after which the intact pair was shown for an additional 5 s. This arrangement meant that the time available to view the correct answer was equated in the read and pretested conditions (for discussion see Kornell et al., 2009), as is customary in the pretesting literature. The assignment of list to condition was counterbalanced over participants and randomization involved presenting items from the two lists, intermixed together.

Once the training phase had ended, participants engaged in an unrelated distractor task activity (i.e., the number series task) for 5 min. Next, they completed the criterial test, in which all 32 pairs were again presented one at a time in a new random order with items from the two lists intermixed. On the criterial test, only the cue word for each pair was presented, with participants having to recall and type the missing target word. The criterial test was self-paced and no feedback was provided. The criterial test served as the measure of the pretesting effect (for further details please refer to the Score Transformations and Analysis Plan section).

Working memory tasks. For the working memory tasks, participants were redirected within their browser window to Inquisit Web 5, which presented the relevant tasks in full screen. After completing all three tasks, participants were redirected within the browser window to complete the remainder of the study. Each task began with a series of practice trials which were discarded for data analysis purposes. Details of the working memory tasks are as follows.

Operation span. This task measures participants' ability to hold a sequence of letters in working memory. Participants were presented with randomly ordered sequences of between three to seven letters. Each

letter was presented for approximately 1 s and preceded by a simple math problem. After a sequence of letters was shown, participants recalled the letters by selecting them in the proper order from a provided letter matrix. The actual task was comprised of 15 trials, each involving a different sequence of letters. The task can be scored in different ways, including absolute scoring (i.e., the sum of all perfectly recalled sequences of letters) or partial credit scoring (i.e., where even partly correct sequences are considered). Following Conway et al. (2005; see also Friedman & Miyake, 2005; Redick et al., 2012; Unsworth & Engle, 2007), which emphasized internal consistency and other advantages of the latter approach, we used partial credit load scoring—that is, the sum of all correctly recalled letters from all sequences, irrespective of whether an entire sequence was perfectly recalled—as each participant's operation span score.

Symmetry span. This task measures participants' ability to hold a sequence of spatial locations in working memory. Participants were presented with sequences of between two to five 4 x 4 matrices in random order for 850 ms each, all containing a red square in any part of the matrix. Each matrix was preceded by the presentation of a black-and-white shape, with participants having to decide whether the shape was symmetrical about its vertical axis or not. After a sequence was shown, participants had to recall the entire sequence in terms of location and serial order and input their answers on a 4 x 4 matrix. The task was comprised of 12 trials. Partial credit load scoring (i.e., the sum of all perfectly recalled square locations, irrespective of whether an entire sequence was perfectly recalled or not), analogous to that used for operation span, was used to derive each participant's symmetry span score.

Reading span. Similar to operation span, this task also measures participants' ability to hold a sequence of letters in working memory. Participants were presented with sequences of between three to seven letters in random order. Each letter was presented for approximately 1 s and preceded by a sentence problem wherein participants had to decide whether a presented sentence was sensical or not. After a sequence of letters was shown, participants recalled the letters by selecting them in the proper order from a provided letter matrix. The actual task was comprised of 15 trials, each involving a different sequence of letters. Partial credit load scoring in the same manner as for operation span and symmetry span was used to derive each participant's reading span score.

Episodic memory tasks. The three episodic memory tasks were as follows.

Delayed free recall. Participants learned six 10-word lists of unrelated common nouns (e.g., *farm, toast, oak*…), with each list undergoing the following three phases: list presentation, distractor task, and free recall test. During list presentation, nouns were presented individually for 1 s each and in the same order for each participant. Next, a 15-second distractor task involved attempting to solve three arithmetic problems. Finally, the free recall test involved typing as many of the words from the most recently presented list as could be recalled, in any order, within 45 s. Each participant's score was the proportion of words that they recalled correctly (out of 60, across all six lists).

Cued recall. Participants learned three lists of 10 unrelated word pairs each using the following procedure. First, each pair (e.g., *soup – ski*) was presented individually for 2 s each and in the same order for each participant. Second, a cued recall test occurred wherein only the cue word for each pair was presented, with participants having to recall and type the missing target word. The cued recall test was self-paced and no feedback was provided. Each participant's score was the proportion of word pairs that they recalled correctly (out of 30, across all three lists).

Image recognition. Participants viewed 30 drawings of various common objects (e.g., food, animals), with each image presented individually for 3 s and in random order for each participant. Next, a recognition test occurred wherein they viewed 60 drawings individually for 5 s each and in random order (including the 30 drawings that they had previously seen and 30 new ones) and were tasked with identifying each image as new or old. Each participant's score was the proportion of images (out of 60) that was correctly identified (Note: this task, which is commonly used in memory studies (e.g., Chen et al., 2017), was simpler over some prior versions in that it did not require memorization of spatial locations, and consequently took less time than such versions).

Fluid intelligence tasks. Details of the three fluid intelligence tasks are as follows.

Raven's progressive matrices. Participants were given 10 min to solve up to 18 logic problems of generally increasing difficulty, presented in the same order for each participant. Each problem was comprised of a 3 x 3 matrix of geometric patterns, with the bottom right pattern missing; from eight presented choices, participants had to select the pattern that would correctly complete the matrix. Each participant's score was the proportion of correctly solved problems (out of 18).

Number series. Participants were given 5 min to solve up to 15 problems. Each problem consisted of a sequence of numbers that followed an unidentified rule, along with five potential answer options. Problems were presented in the same order for each participant. Participants were instructed to choose the answer option that represented the next number in the series. A practice example was provided beforehand. Each participant's score was the proportion of correctly solved problems (out of 15).

Letter sets. Participants were given up to 5 min to solve up to 20 problems. Each problem consisted of five sets of four letters each. Four of the sets followed the same unidentified rule; participants were instructed to identify the one set that did not follow that rule. Problems were presented in the same order for each participant. Each participant's score was the proportion of correctly solved problems (out of 20).

Score transformations and analysis plan

Data processing and all analyses involved use of R version 4.1.0 (R Core Team, 2021).

Pretesting effect. Following Kornell et al. (2009) and other pretesting studies, all pretested items that were guessed correctly during the training phase by a given participant were removed prior to calculation of the pretesting effect for that participant. This procedure affected a very small amount of data (in the case of pretesting with paired associates, less than 5 % of pretested items are usually guessed correctly) and served to remove typically easier items that participants may have already learned prior to participating in the study (although, as Kornell et al. noted, doing so might create a small selection bias that favored the reading condition, yielding smaller pretesting effects).

As is typical in the test-enhanced learning literature, the pretesting effect was calculated for each participant by computing the difference score between the (a) average performance for all 16 pretested items on the criterial test (minus any items that were guessed correctly during the training phase) and the (b) average performance for all 16 read items on the criterial test. Difference scores, however, tend to have low reliability (for discussions see Robey, 2019; Unsworth, 2019). Indeed, the split-half reliability of the pretesting effect in Study 1 (0.56) was lower than the reliability of performance in the pretested condition (0.70) and the read condition (0.70) in the same study, and to foreshadow, pretesting effect reliability in Study 2 was even lower (0.24) whilst the pretested and read conditions reliability remained about the same (approximately 0.70). In response, following Robey's (2019) strategy for addressing the low reliability of difference scores, we eschewed analyses relying on pretesting effect values. Rather, as described in the Analysis Plan section, we conducted linear mixed-effect model analyses using separate data for the pretested and the read conditions. While still addressing the pretesting effect, such analyses also provided insight into the two conditions that contribute to that effect. Pretesting effect difference scores remained in use to describe the pretesting effect across the study sample and in some supplementary analyses.

Composite z-scores. Drawing on approaches taken in prior individual differences research (e.g., Brewer & Unsworth, 2012; Robey, 2019; see also Wingert & Brewer, 2018), we first *z*-score transformed the results for each measure—that is, rescaling the data from each measure so that the data has a mean of 0 and a standard deviation of 1. Then, separately for WMC, EM ability, and gF, we calculated the mean of the respective *z*-scores for each participant to produce composite *z*-scores. For example, the composite *z*-score for WMC for a given participant included the average of the *z*-scores for the operation span, symmetry span, and reading span tasks for that participant.

Factor scores. Besides computing composite *z*-scores, we also used confirmatory factor analysis (CFA) to output factor scores separately for WMC, EM ability, and gF. An advantage of factor scores is that they avoid the measurement error associated with composite scores (Robey, 2019; see also Wingert & Brewer, 2018). CFA models were fitted using lavaan version 0.6–9 (Rosseel, 2012) in R. Each model was specified with three different task scores representing the indicators of each construct (WMC, EM ability, and gF).[1] The factor loadings were estimated, with the variance of latent factors set to 1. All the factor loadings exceeded the threshold of 0.35, which is considered acceptable (Jung & Lee, 2011). The factor scores were then derived from these estimated factor loadings. Factor scores were used in a separate set of analyses to those conducted using composite z-scores.

Analysis plan. First, we computed descriptive statistics for all measures. Second, we assessed the magnitude and variation of the pretesting effect across the entire sample, reporting the percentages of participants exhibiting positive, negative, and null pretesting effects. This analysis provided insights into the dataset, particularly regarding the variability of pretesting effect magnitude among participants. Third, we explored the influence of individual differences in cognitive abilities by fitting two sets of three linear mixed-effects models, each focusing on a specific ability, using criterial test data for both pretested and read items. To provide converging evidence, one set involved composite *z*-scores and the other set involved factor scores. For each analysis set, we fitted models separately for WMC, EM, and gF to examine the interaction effects between training condition (pretested vs. read) and the individual differences of interest. This approach allowed us to investigate whether individual differences in WMC, EM ability, or gF predicted performance in the pretested condition, the read condition, or both. Fourth, we conducted two linear mixed-effects models, each considering all three cognitive abilities simultaneously (one using composite z-scores and the other using factor scores) to examine the interaction effects between training condition (pretested vs. read) and the three cognitive abilities. Finally, for additional insights, we also performed supplementary analyses that involved dividing up study data into quartiles (analogous to the approaches used in Brewer & Unsworth, 2012; Minear et al., 2018).[2]

[1] An initial model with three latent variables (WMC, EM, and gF) showed poor fit, $\chi^2(24) = 51.92$, $p < .001$, CFI = 0.91, TLI = 0.87, SRMR = .074, RMSEA = 0.094, 90% CI [0.059, 0.129], likely due to insufficient sample size, as adding latent variables increases sample size requirements (Wolf et al., 2013). In response, we used single-factor models to generate the factor scores. Following Robey (2019), model fit indices are not reported given that these models involved only one latent variable and three tasks each.

[2] Examination of the WMC scatterplots for Study 1 indicated the presence of a possible outlier—that is, a participant with WMC composite *z*-scores and factor scores < −3. As an exploratory measure, we repeated the entire analysis sequence reported here but with that participant's data removed. The results, which are detailed in the Supplementary Materials, did not appreciably differ from the analyses reported here.

Results

Descriptive statistics and split-half reliabilities are presented in Table 1 and a correlation matrix of all measures is presented in Table 2. Within each of the individual differences categories, the three measures were all significantly correlated with one another, as expected, and with correlations that are generally similar in magnitude to that observed by Brewer and Unsworth (2012), Minear et al. (2018), Pan et al. (2015), and Robey (2019). We next report learning phase performance, the extent of the pretesting effect among the study sample, results of linear mixed-effects models, and supplementary analyses.

Pretesting task

Learning phase. In line with typical patterns in the pretesting literature, participants rarely guessed pretested items correctly during the learning phase (only 4 % of all items were guessed correctly). As previously noted, prior to analyses of the criterial test data, test trials involving those correctly guessed pretest items were removed.

Pretesting effect. Across the entire sample, participants tended to recall more pretested than read items on the criterial test, $t(134) = 10.84$, $p < .0001$, $d = 0.93$. The mean pretesting effect was 0.21 proportion correct, which is comparable to pretesting effects that have been observed in the literature. As in prior work, there was variation in that effect: 80 % of participants exhibited a numerically positive pretesting effect, 16 % of participants exhibited a numerically negative pretesting effect (i.e., they remembered more word pairs in the read condition than in the pretested condition), and 4 % of participants exhibited zero pretesting effect. Across the entire sample, pretesting effect magnitude ranged from −0.44 to 0.75.

Linear mixed-effects models

Individual Cognitive Abilities and Criterial Test Performance. We submitted criterial test scores to three linear mixed-effects models, each conducted separately with the composite scores of WMC, EM ability, and gF. Additionally, we fitted three analogous linear mixed-effects models using factor scores for WMC, EM ability, and gF. Each

Table 1
Descriptive statistics for study 1.

Measure	Mean	*SD*	Skewness	Kurtosis	Reliability
Pretested condition	0.74	0.18	−1.00	1.20	0.70
Read condition	0.53	0.21	−0.0032	−0.65	0.70
Pretesting effect (all items)	0.21	0.22	−0.061	−0.20	0.59
Pretesting effect	0.21	0.22	−0.029	−0.18	0.56
Operation span	65.30	9.59	−1.86	4.53	0.77
Reading span	62.21	12.54	−1.80	4.41	0.83
Symmetry span	34.19	6.40	−1.21	1.25	0.67
Cued recall	0.49	0.16	0.57	0.33	0.92
Delayed free recall	0.55	0.23	−0.12	−0.93	0.82
Image recognition	0.93	0.066	−1.65	4.33	0.55
Raven's matrices	0.81	0.19	−1.68	2.67	0.82
Letter sets	0.53	0.12	−0.23	0.12	0.57
Number series	0.67	0.19	−0.33	−0.61	0.75

Note. all items = including word pairs that were guessed correctly during pretesting. Per the conventions of the pretesting effect literature, those items are removed prior to calculating the pretesting effect. Such removal was performed prior to all subsequent analyses, but the pretesting effect in the case of no items being removed is included here for completeness. All reliabilities are split-half reliability calculated using trial-level data. *SD* = standard deviation.

Table 2
Correlation Matrix for Study 1.

	Pretest	Read	PTeffect	Ospan	Rspan	Symspan	CR	DFR	Recog	Raven	Lsets	Nseries
Pretest	1.00											
Read	0.360***	1.00										
PTeffect	0.480***	−0.646***	1.00									
Ospan	0.029	0.073	−0.045	1.00								
Rspan	−0.009	0.047	−0.051	0.659***	1.00							
Symspan	−0.034	0.038	−0.063	0.518***	0.458***	1.00						
CR	0.326***	0.460***	−0.166	0.153	0.155	0.243**	1.00					
DFR	0.247**	0.222**	−0.006	0.253**	0.281***	0.289***	0.610***	1.00				
Recog	0.211*	0.397***	−0.201*	0.138	0.184*	0.265**	0.469***	0.248**	1.00			
Raven	0.142	0.346***	−0.208*	0.351***	0.457***	0.391***	0.247**	0.321***	0.400***	1.00		
Lsets	−0.011	0.027	−0.034	0.379***	0.262**	0.273**	0.098	0.123	0.207*	0.184*	1.00	
Nseries	−0.023	0.239**	−0.244**	0.335***	0.306***	0.258**	0.113	0.173*	0.079	0.355***	0.264**	1.00

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; Pretest = performance in pretest condition; Read = performance in read condition; PTeffect = pretesting effect; Ospan = operation span; Rspan = reading span; Symspan = symmetry span; CR = cued recall; DFR = delayed free recall; Recog. = image recognition; Raven = Raven's progressive matrices; Lsets = Letter sets; Nseries = Number series.

model included training condition (read = 0 vs. pretested = 1), the respective scores (composite or factor), and their interaction as predictors, with crossed random intercepts for participants. All models were fitted using the lme4 package version 1.1.34 (Bates et al., 2015) in R. Scatterplots corresponding to the composite score-based models are presented in the left-side panels of Fig. 2, whereas scatterplots for the factor score-based models are shown in the right-side panels. Interaction effect results are detailed in Table 3.

In the mixed-effects models using composite scores, EM ability was a significant predictor of criterial test scores, $b = 0.095$, $SE = 0.016$, $p < .001$, $d = 0.77$, and gF was also a significant predictor of criterial test scores, $b = 0.059$, $SE = 0.016$, $p < .001$, $d = 0.47$. WMC scores, however, were not a significant predictor ($p = .439$). In essence, higher EM ability and higher gF scores were associated with higher overall criterial performance, whereas higher WMC scores were not. Crucially, there was a significant interaction between gF composite scores and training condition, b = -0.050, $SE = 0.019$, $p = .008$, d = -0.47. Specifically, in the read condition, gF scores significantly and positively predicted criterial test scores (95 % CI = [0.03, 0.09]), whereas the relationship was not significant in the pretested condition (95 % CI = [-0.02, 0.04]). That result suggests that gF is associated with performance for read items, which in turn influences the magnitude of the pretesting effect. No significant interactions were found between training condition and WMC ($p = .462$) or EM ability ($p = .067$).

In the mixed-effects models using factor scores, EM ability was a significant predictor of criterial test scores, $b = 0.096$, $SE = 0.016$, $p < .001$, $d = 0.79$, and gF was also a significant predictor, $b = 0.062$, $SE = 0.016$, $p < .001$, $d = 0.49$. WMC scores were not a significant predictor ($p = .40$). These patterns mirrored the analyses with composite scores. Further, there was a significant interaction between gF factor scores and the training condition, b = -0.057, $SE = 0.018$, $p = .002$, d = -0.54. Specifically, in the read condition, gF scores significantly and positively predicted criterial test scores (95 % CI = [0.03, 0.09]), whereas the relationship was not significant in the pretested condition (95 % CI = [-0.03, 0.04]). That result again suggests that gF is associated with performance for read items, which in turn influences the magnitude of the pretesting effect. No significant interactions were found between training condition and factor scores for WMC ($p = .51$) and EM ability ($p = .050$).

Multiple Cognitive Abilities and Criterial Test Performance. We submitted criterial test scores to two linear mixed-effects models in which all three cognitive abilities were considered simultaneously, conducted separately using composite scores and factor scores. Both models included training condition (read = 0 vs. pretested = 1), scores for WMC, EM ability, and gF, and interaction terms between each cognitive ability and training condition, with crossed random intercepts for participants. Interaction effect results are detailed in Table 4.

In the model using composite scores, WMC significantly predicted criterial test scores, b = -0.047, $SE = 0.019$, $p = .013$, d = -0.32, as did EM ability, $b = 0.091$, $SE = 0.017$, $p < .001$, $d = 0.71$, and gF, $b = 0.054$, $SE = 0.019$, $p = .004$, $d = 0.37$. Notably, while WMC was negatively related to criterial test performance, the relationships with EM ability and gF were positive. Moreover, there was a significant interaction between gF composite scores and training condition, b = -0.056, $SE = 0.023$, $p = .016$, d = -0.43. Specifically, in the read condition, gF composite scores significantly and positively predicted criterial test scores (95 % CI = [0.02, 0.09]), whereas the relationship was not significant in the pretested condition (95 % CI = [-0.04, 0.04]). This interaction aligns with the patterns observed in the linear mixed-effects models for gF alone. No significant interactions were found between composite scores and training condition for WMC ($p = .243$) or EM ability ($p = .295$).

In the model using factor scores, EM ability scores significantly and positively predicted criterial test scores, $b = 0.091$, $SE = 0.016$, $p < .001$, $d = 0.75$, as did gF scores, $b = 0.062$, $SE = 0.018$, $p < .001$, $d = 0.45$. WMC factor scores, however, were not a significant predictor ($p = .060$). Further, there was a significant interaction between gF factor scores and training condition, b = -0.064, $SE = 0.021$, $p = .004$, d = -0.52. Specifically, in the read condition, gF factor scores significantly and positively predicted criterial test scores (95 % CI = [0.03, 0.10]), whereas the relationship was not significant in the pretested condition (95 % CI = [-0.04, 0.03]). This pattern aligns with that observed in the aforementioned model using composite scores. No significant interactions were found between factor scores and training condition for WMC ($p = .263$) or EM ability ($p = .116$).

Quartile analyses

Similar to approaches taken by Brewer and Unsworth (2012), Minear et al. (2018), and others, in supplementary analyses we investigated whether the pretesting effect differed among the lowest and highest quartiles of each composite *z*-score and each factor score. These analyses involved (a) computing the pretesting effect for the highest and lowest quartiles of the cognitive ability being examined and (b) performing a mixed-factors Analysis of Variance (ANOVA) on criterial test scores in the pretested and read conditions among the lowest and highest ability quartiles with factors of training condition (pretested vs. read) and quartile (lowest vs. highest). A statistically significant interaction constituted evidence that the pretesting effect differed according to the relevant cognitive ability. Performance in the pretested and read conditions for the lowest and highest composite score quartiles of each cognitive ability are presented in the left-side panels of Fig. 3, while those for quartiles drawn from factor scores are shown in the right-side panels.

We first consider the quartile-based analyses involving composite scores. There were statistically significant ($p < .001$) pretesting effects in the lowest and highest quartiles, respectively, for WMC ($ds = 0.95$, 0.94), EM ability ($ds = 1.15$, 0.78), and gF ($ds = 1.44$, 0.80). Pretesting

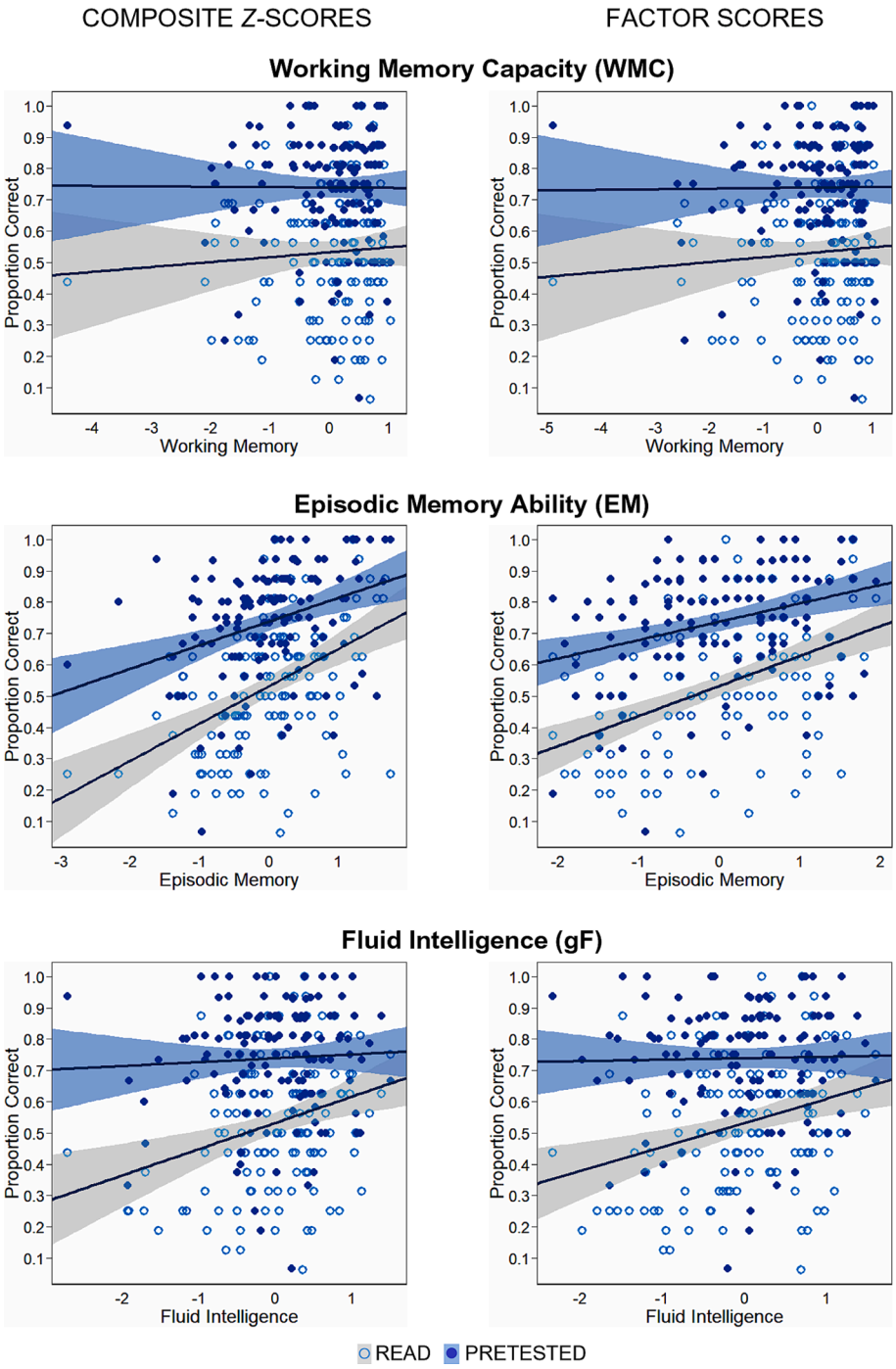

Fig. 2. Criterial test performance as a function of working memory capacity, episodic memory ability, and fluid intelligence scores in study 1. Note: Line = best fitting regression line and shading = 95 % CI.

Table 3
Interaction effects of cognitive ability and training condition in separate linear mixed-effects models for each ability in study 1.

Model type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Composite scores	Working memory	−0.014	0.019	.462	−0.13
	Episodic memory	−0.035	0.019	.067	−0.32
	Fluid intelligence	−0.050	0.019	.008**	−0.47
Factor scores	Working memory	−0.013	0.019	.510	−0.11
	Episodic memory	−0.038	0.019	.050	−0.34
	Fluid intelligence	−0.057	0.018	.002**	−0.54

Note. ** = $p < .01$.

Table 4
Interaction effects of cognitive ability and training condition in simultaneous linear mixed-effects models in study 1.

Model type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Composite scores	Working memory	0.024	0.023	.295	0.18
	Episodic memory	−0.024	0.020	.243	−0.21
	Fluid intelligence	−0.056	0.023	.016*	−0.43
Factor scores	Working memory	0.024	0.022	.263	0.20
	Episodic memory	−0.030	0.019	.116	−0.28
	Fluid intelligence	−0.064	0.021	.004**	−0.52

Note. * = $p < .05$, ** = $p < .01$.

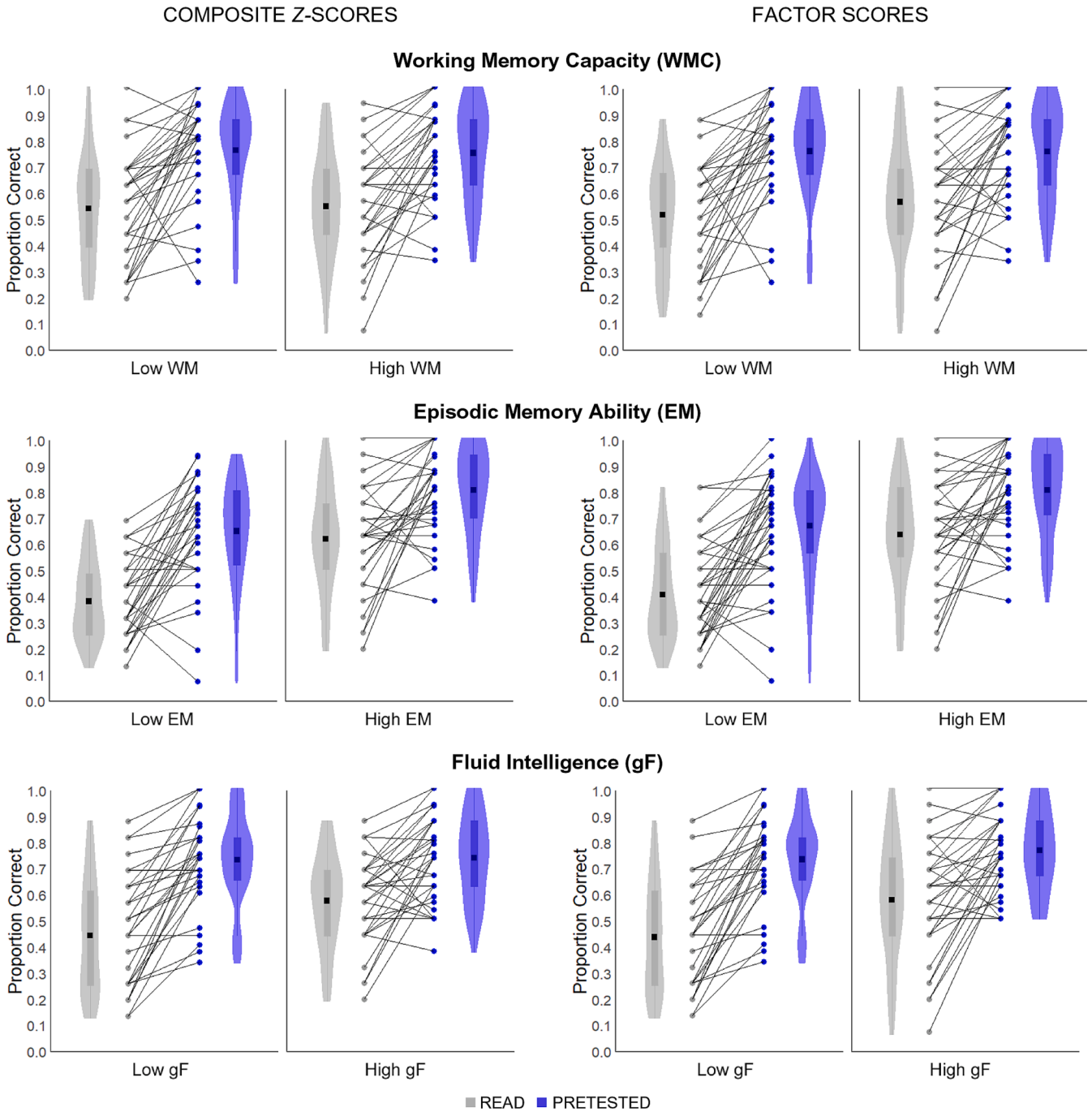

Fig. 3. Criterial test performance among the highest and lowest ability quartiles in study 1.

effect magnitude appeared to be larger, at least numerically, for the lowest quartiles in the case of EM ability and gF. Mixed-factors ANOVAs, however, did not find a significant interaction for the cases of WMC or EM ability (p-values $\geq .16$), but did reveal a significant interaction for gF, $F\,(66) = 6.48$, $p = .013$, $\eta_p^2 = 0.089$. Hence, the pretesting effect was significantly larger for the lowest ability quartile only for the case of gF. That result is consistent with patterns evident in the corresponding panel of Fig. 3, wherein performance among both gF quartiles was comparable for pretested items ($M_{\text{lowestgF}} = 0.73$, $M_{\text{highestgF}} = 0.73$) but differed more substantially for read items ($M_{\text{lowestgF}} = 0.44$, $M_{\text{highestgF}} = 0.57$). In other words, low-gF individuals differed from high-gF individuals in performing worse on read items whilst performing similarly on pretested items.

We next turn to the quartile-based analyses involving factor scores. There were statistically significant ($p < .001$) pretesting effects in the lowest and highest quartiles, respectively, for WMC ($ds = 1.06, 0.82$), EM ability ($ds = 1.12, 0.73$), and gF ($ds = 1.48, 0.86$). There appeared to be larger pretesting effect sizes for the lowest ability quartiles, at least numerically. Mixed-factors ANOVAs, however, did not find a significant interaction for the cases of WMC or EM ability (p-values $\geq .089$), but did find a significant interaction for gF, $F\,(66) = 4.44$, $p = .039$, $\eta_p^2 = 0.063$. The latter result is consistent with patterns evident in the corresponding panel of Fig. 3, wherein performance among both gF quartiles was comparable for pretested items ($M_{\text{lowestgF}} = 0.73$, $M_{\text{highestgF}} = 0.76$) but differed more substantially for read items ($M_{\text{lowestgF}} = 0.43$, $M_{\text{highestgF}} = 0.58$). These results mirror the analyses performed using composite scores.

Study 2

The first study revealed that among a sample of university students, individuals with lower gF scores tended to have a larger magnitude pretesting effect, driven by reduced performance on read items, whereas individual differences in WMC and EM ability did not explain variations in pretesting effect magnitude. We next expanded the scope of our investigation to include a multi-national, adult, English-speaking sample that was recruited online. This sample was more heterogeneous than that of the first study, including with respect to age, educational attainment, and national origin.

Data availability

Data, analysis code, and materials, are available via OSF and accessible via the same links as provided for Study 1.

Methods

This study design and originally intended analysis plan was preregistered at AsPredicted.org (URL: https://aspredicted.org/JS4_8KF). Deviations from the analysis plan are discussed below.

Participants

The target sample size was identical to that of the first study. One-hundred and thirty-four participants were recruited online via Prolific Academic (Prolific, London, UK) in exchange for a payment of USD $14.05 per participant. Prolific is a commonly used crowdsourcing platform in academic research that has a good reputation for data quality (Palan & Schitter, 2017). All participants had to be in an English-speaking country (e.g., Australia, Canada, New Zealand, the United Kingdom, or the United States), be fluent in English, be aged between 21 and 45 years (the lower limit dictated by ethics board requirements for online studies) and have an approval rate of 95 % or higher on prior Prolific studies. These sample characteristics reflected our intent to sample broadly from English-speaking adult participants that are common to the Prolific platform. Of that original sample, 11 participants were excluded for multiple submissions and 1 participant was excluded for experiencing technical difficulties during the experiment, leaving a final sample of 122 participants.

In the final sample, the mean age was 31.1 years (ranging between 21 and 45 years) and 43 % were female. Fifty-two percent of the participants were of Caucasian/White ancestry, 22 % were Asian, 19 % were Black, 4 % had a multi-ethnic background, 2 % were of other ethnic groups, and < 1 % declined to provide ethnicity information. Most participants (55 %) were from the United Kingdom, followed by Australia (19 %), Canada (16 %), the United States (10 %), and New Zealand (<1%). The highest level of education among 40 % participants was an undergraduate degree, whereas for 27 %, 17 %, and 16 % of the participants, it was a graduate degree, high school diploma, or other levels of education, respectively.

Materials, procedure, and tasks

All aspects of the materials, procedure, and tasks were largely identical to that of Study 1, except that participation occurred online in an unsupervised environment (participants were urged to complete the study in an undisturbed location). We also implemented a technical check to rule out potential software or hardware incompatibility issues. Participation was only permitted using a desktop or laptop computer that was equipped with the Google Chrome browser and the Inquisit Web application. The demographic questions were further modified to reflect the online setting and different sample characteristics. Finally, to verify that participants were not engaged in off-task activity, their browser activity was tracked using TaskMaster (Permut al., 2019). No participants were removed from the study due to evidence of substantial off-task activity, which suggests that participant compliance with instructions was good.

Score transformations and analysis plan

Computing of composite *z*-scores, factor scores, and the analysis plan that was implemented were all identical to that for Study 1. In the interests of complete transparency (Willroth & Atherton, 2024), the preregistered analysis plan, which was filed in advance of any analysis of either Study 1 or 2 and prior to data collection for Study 2, specified a *t*-test to determine the pretesting effect, the calculation of composite *z*-scores for WMC, EM ability, and gF (both of which we conducted), followed by correlational and regression analyses involving those composite scores and pretesting effect scores (which were largely not conducted due to low reliability). Although such an approach has been used in prior studies (e.g., Brewer & Unsworth, 2012; Pan et al., 2015, etc.), to avoid using pretesting effect difference scores, we adopted the same analysis approach that was used in Study 1 (which was based in part on analysis methods also used in the individual differences and learning strategies literature, e.g., Robey, 2019), which included the calculation of both composite *z*-scores and factor scores for all three cognitive abilities.[3]

Results

Descriptive statistics and split-half reliabilities are presented in Table 5 and a correlation matrix of all measures is presented in Table 6. As in Study 1, within each of the individual differences categories, the three measures were all significantly correlated with one another. As previously noted, however, that reliability of the pretesting effect dif-

Table 5
Descriptive statistics for study 2.

Measure	Mean	*SD*	Skewness	Kurtosis	Reliability
Pretested condition	0.69	0.20	−0.54	−0.53	0.72
Read condition	0.48	0.22	0.17	−0.76	0.69
Pretesting effect (all items)	0.22	0.18	0.43	0.11	0.27
Pretesting effect	0.21	0.19	0.50	0.31	0.24
Operation span	59.34	15.30	−1.80	3.33	0.80
Reading span	57.98	16.71	−1.34	1.37	0.89
Symmetry span	30.39	8.23	−0.72	0.082	0.79
Cued recall	0.49	0.27	0.38	−0.90	0.91
Delayed free recall	0.51	0.21	0.46	−0.64	0.93
Image recognition	0.88	0.12	−2.15	5.96	0.68
Raven's matrices	0.69	0.23	−0.58	−0.89	0.87
Letter sets	0.49	0.12	−0.032	−0.051	0.64
Number series	0.61	0.21	−0.42	−0.52	0.79

Note. all items = including word pairs that were guessed correctly during pretesting. Per the conventions of the pretesting effect literature, those items are removed prior to calculating the pretesting effect. Such removal was performed prior to all subsequent analyses, but the pretesting effect in the case of no items being removed is included here for completeness. All reliabilities are split-half reliability calculated using trial-level data. *SD* = standard deviation.

[3] A CFA model with a three-factor structure had a poor fit, $\chi^2(28) = 412.01$, $p < .001$, CFI = 0.96, TLI = 0.93, SRMR =.069, RMSEA = 0.086, 90% CI [0.041, 0.128] and single-factor models were used instead to generate the factor scores. All the factor loadings exceeded the threshold of 0.35. This approach was identical to that used in Study 1.

Table 6
Correlation matrix for study 2.

	Pretest	Read	PTeffect	Ospan	Rspan	Symspan	CR	DFR	Recog	Raven	Lsets	Nseries
Pretest	1.00											
Read	0.606***	1.00										
PTeffect	0.331***	−0.549***	1.00									
Ospan	0.055	0.236	−0.223	1.00								
Rspan	0.077	0.243	−0.203	0.772***	1.00							
Symspa	0.213	0.175	0.019	0.465***	0.552***	1.00						
CR	0.338***	0.437***	−0.170	0.416	0.404	0.215**	1.00					
DFR	0.240**	0.347**	−0.153	0.477**	0.509***	0.364***	0.777***	1.00				
Recog	0.263*	0.226***	0.007*	0.444	0.399*	0.342**	0.332***	0.302**	1.00			
Raven	0.212	0.086***	0.116*	0.179***	0.096***	0.346***	0.091**	0.110***	0.252***	1.00		
Lsets	0.3217	0.105	0.104	0.058***	0.118**	0.192**	0.052	0.061	0.028*	0.248*	1.00	
Nseries	0.142	0.004**	0.148**	0.113***	0.071***	0.217**	0.114	0.096*	0.113	0.590***	0.329**	1.00

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; Pretest = performance in pretest condition; Read = performance in read condition; PTeffect = pretesting effect; Ospan = operation span; Rspan = reading span; Symspan = symmetry span; CR = cued recall; DFR = delayed free recall; Recog. = image recognition; Raven = Raven's progressive matrices; Lsets = Letter sets; Nseries = Number series.

ference score was low (0.24, which is noticeably lower than in Study 1), whereas reliability in the pretested and read conditions of Study 2 remained substantially higher and comparable to that observed in the same conditions in Study 1 (~0.70). Pretesting effect magnitude and other distributional characteristics were similar to the prior study. Given that the task was entirely identical to that used in Study 1, including randomized trial order, we speculate tentatively that the reliability differences may have been due to sample characteristics and/or the online testing environment. We next report analyses of Study 2 in the same order as in Study 1. Analyses described as exploratory were not preregistered.

Pretesting task

Learning phase. Similar to prior results, participants rarely guessed pretested items correctly during the learning phase (less than 4 % of all items were guessed correctly). Prior to analyses of the criterial test data, test trials involving those correctly guessed pretest items were removed.

Pretesting effect. Across the entire sample, participants tended to recall more pretested than read items on the criterial test, $t\,(121) = 12.46$, $p < .0001$, $d = 1.13$. The mean pretesting effect was 0.21 proportion correct, which is identical to that of Study 1. Variation in that effect was evident given that 85 % of participants exhibited a numerically positive pretesting effect, 12 % of participants exhibited a numerically negative pretesting effect, and 3 % of participants exhibited zero pretesting effect. Across the entire sample, pretesting effect magnitude ranged from −0.18 to 0.81.

Linear mixed-effects models

Individual Cognitive Abilities and Criterial Test Performance. In an approach akin to that used for Study 1, we submitted criterial test scores to three linear mixed-effects models, each conducted separately with the composite scores and factor scores of WMC, EM ability, or gF. Scatterplots corresponding to the composite score-based models are presented in the left panels of Fig. 4, while those for the factor score-based models are shown in the right panels. Interaction effect results are detailed in Table 7.

In the mixed-effects models using composite scores, WMC significantly and positively predicted criterial test scores, $b = 0.056$, $SE = 0.019$, $p = .003$, $d = 0.45$, and so did EM ability, $b = 0.093$, $SE = 0.018$, $p < .001$, $d = 0.77$. On the other hand, gF composite scores were not a significant predictor ($p = .318$). Further, no significant interactions were found between training condition and composite scores for WMC ($p = .081$), EM ability ($p = .150$), or gF ($p = .079$).

In the mixed-effects models using factor scores, WMC significantly and positively predicted criterial test scores, $b = 0.055$, $SE = 0.019$, $p = .004$, $d = 0.44$, as did EM ability, $b = 0.097$, $SE = 0.018$, $p < .001$, $d = 0.81$. As in the aforementioned analysis with composite scores, gF factor scores were not a significant predictor ($p = .714$). There was, however, a significant interaction between WMC factor scores and training condition, b = -0.038, $SE = 0.017$, $p = .024$, d = -0.42, in contrast with the results in the prior analysis. Specifically, in the read condition, WMC significantly and positively predicted criterial test scores (95 % CI = [0.02, 0.09]), whereas the relationship was not significant in the pretested condition (95 % CI = [-0.02, 0.05]). In other words, individuals with higher WMC scores exhibited a smaller pretesting effect, and that difference was due to relatively higher scores in the read condition. No significant interactions were found between training condition and factor scores for EM ability ($p = .061$) and gF ($p = .086$).

Multiple Cognitive Abilities and Criterial Test Performance. As in Study 2, we also submitted criterial test scores to two linear mixed-effects models, conducted separately using composite scores and factor scores, in which WMC, EM ability, and gF were considered simultaneously. Interaction effect results are detailed in Table 8. In the model using composite scores, EM ability was a significant positive predictor of criterial test scores, $b = 0.091$, $SE = 0.022$, $p < .001$, $d = 0.63$. WMC scores ($p = .887$) and gF composite scores ($p = .867$), however, were not significant in predicting criterial test scores. There was also a significant interaction between gF scores and training condition, $b = 0.039$, $SE = 0.017$, $p = .023$, $d = 0.42$. Specifically, in the pretested condition, gF scores significantly and positively predicted criterial test scores (95 % CI = [0.01, 0.08]), whereas the relationship was not significant in the read condition (95 % CI = [-0.03, 0.04]). This pattern differs from that observed in the linear mixed-effects models with gF alone and contrasts with the findings from Study 1. No significant interactions were found between composite scores and training condition for WMC ($p = .131$) or EM ability ($p = .525$).

In the model using factor scores, EM ability significantly and positively predicted criterial test scores, $b = 0.092$, $SE = 0.020$, $p < .001$, $d = 0.68$. WMC ($p = .578$) and gF ($p = .736$), however, were not significant predictors. As in the analysis with composite scores, there was a significant interaction between gF factor scores and training condition, $b = 0.036$, $SE = 0.017$, $p = .034$, $d = 0.39$. Moreover, although there was no significant effect of gF factor scores on criterial test scores in the read condition (95 % CI = [-0.04, 0.03]) or the pretested condition (95 % CI = [-0.01, 0.06]), the relation between gF factor scores and criterial test scores was positive in the pretested condition ($b = 0.03$) and negative in the read condition (b = -0.01). No significant interactions were found between training condition and factor scores for WMC ($p = .089$) and EM ($p = .278$).

Quartile analyses

As in Study 1, we performed supplementary analyses involving the lowest and highest quartiles of each factor score and each composite *z*-score. We first consider the analyses involving composite scores, which

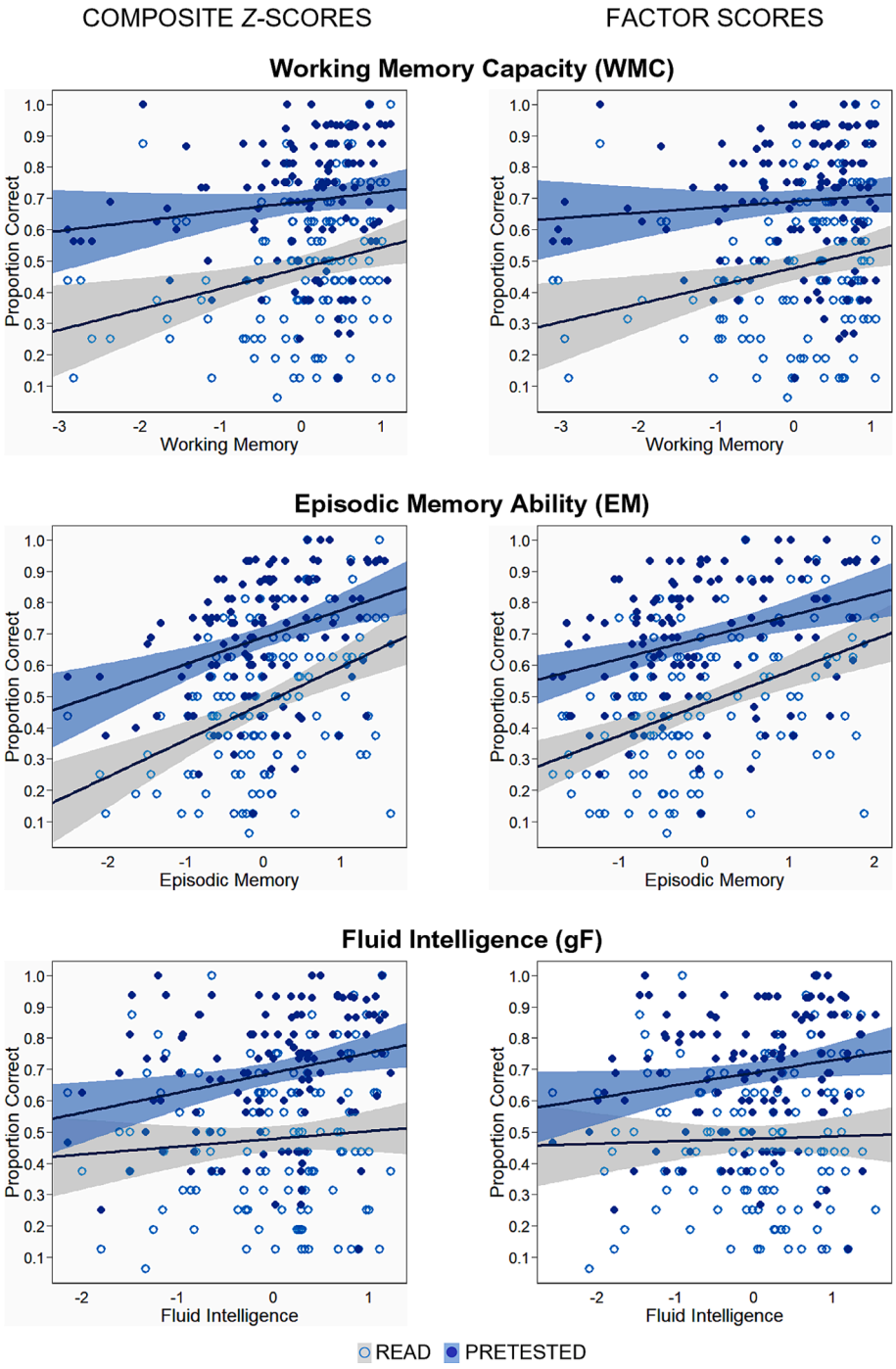

Fig. 4. Criterial test performance as a function of working memory capacity, episodic memory ability, and fluid intelligence scores in study 2. Note: Line = best fitting regression line and shading = 95 % CI.

Table 7
Interaction effects of cognitive ability and training condition in separate linear mixed-effects models for each ability in study 2.

Model type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Composite scores	Working memory	−0.030	0.017	.081	−0.32
	Episodic memory	−0.025	0.017	.150	−0.26
	Fluid intelligence	0.030	0.017	.079	0.33
Factor scores	Working memory	−0.038	0.017	.024*	−0.42
	Episodic memory	−0.032	0.017	.061	−0.35
	Fluid intelligence	0.029	0.017	.086	0.31

Note. * = $p < .05$.

Table 8
Interaction effects of cognitive ability and training condition in simultaneous linear mixed-effects models in study 2.

Model type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Composite scores	Working memory	−0.031	0.021	.131	−0.28
	Episodic memory	−0.013	0.020	.525	−0.12
	Fluid intelligence	0.039	0.017	.023*	0.42
Factor scores	Working memory	−0.032	0.019	.089	−0.32
	Episodic memory	−0.021	0.019	.278	−0.20
	Fluid intelligence	0.036	0.017	.034*	0.39

Note. * = $p < .05$, ** = $p < .01$.

are depicted in the left-side panels of Fig. 5. There were statistically significant ($p < .001$) pretesting effects in the lowest and highest quartiles, respectively, for WMC ($ds = 1.78, 0.88$), EM ability ($ds = 1.54, 0.87$), and gF ($ds = 0.94, 1.00$). The effect sizes for the lowest WMC and EM ability quartiles were numerically larger. Mixed-factors ANOVAs did not find a significant interaction for the cases of EM ability or gF (p-values $\geq .113$), but revealed a significant interaction for WMC, $F(60) = 8.092$, $p = .0061$, $\eta_p^2 = 0.12$. That interaction is consistent with patterns evident in the corresponding panel of Fig. 5, namely that performance among WMC quartiles was comparable for pretested items ($M_{\text{lowestWMC}} = 0.66$, $M_{\text{highestWMC}} = 0.73$) but differed substantially for read items ($M_{\text{lowestWMC}} = 0.38$, $M_{\text{highestWMC}} = 0.57$). In other words, low-WMC individuals differed from high-WMC individuals in performing worse on read items whilst performing more similarly on pretested items.

We next turn to the quartile-based analyses involving factor scores. There were statistically significant ($p < .0001$) pretesting effects in the lowest and highest quartiles, respectively, for WMC ($ds = 1.52, 0.80$), EM ability ($ds = 1.58, 0.86$), and gF ($ds = 0.94, 1.10$). The effect sizes for the lowest WMC and EM ability quartiles were numerically larger. Mixed-factors ANOVAs did not find a significant interaction for the cases of EM ability or gF (p-values $\geq .129$), but revealed a significant interaction for WMC, $F(60) = 8.95$, $p = .0040$, $\eta_p^2 = 0.13$. That interaction is consistent with patterns evident in the corresponding panel of Fig. 5, namely that performance among WMC quartiles was comparable for pretested items ($M_{\text{lowestWMC}} = 0.65$, $M_{\text{highestWMC}} = 0.69$) but differed substantially more for read items ($M_{\text{lowestWMC}} = 0.36$, $M_{\text{highestWMC}} =$

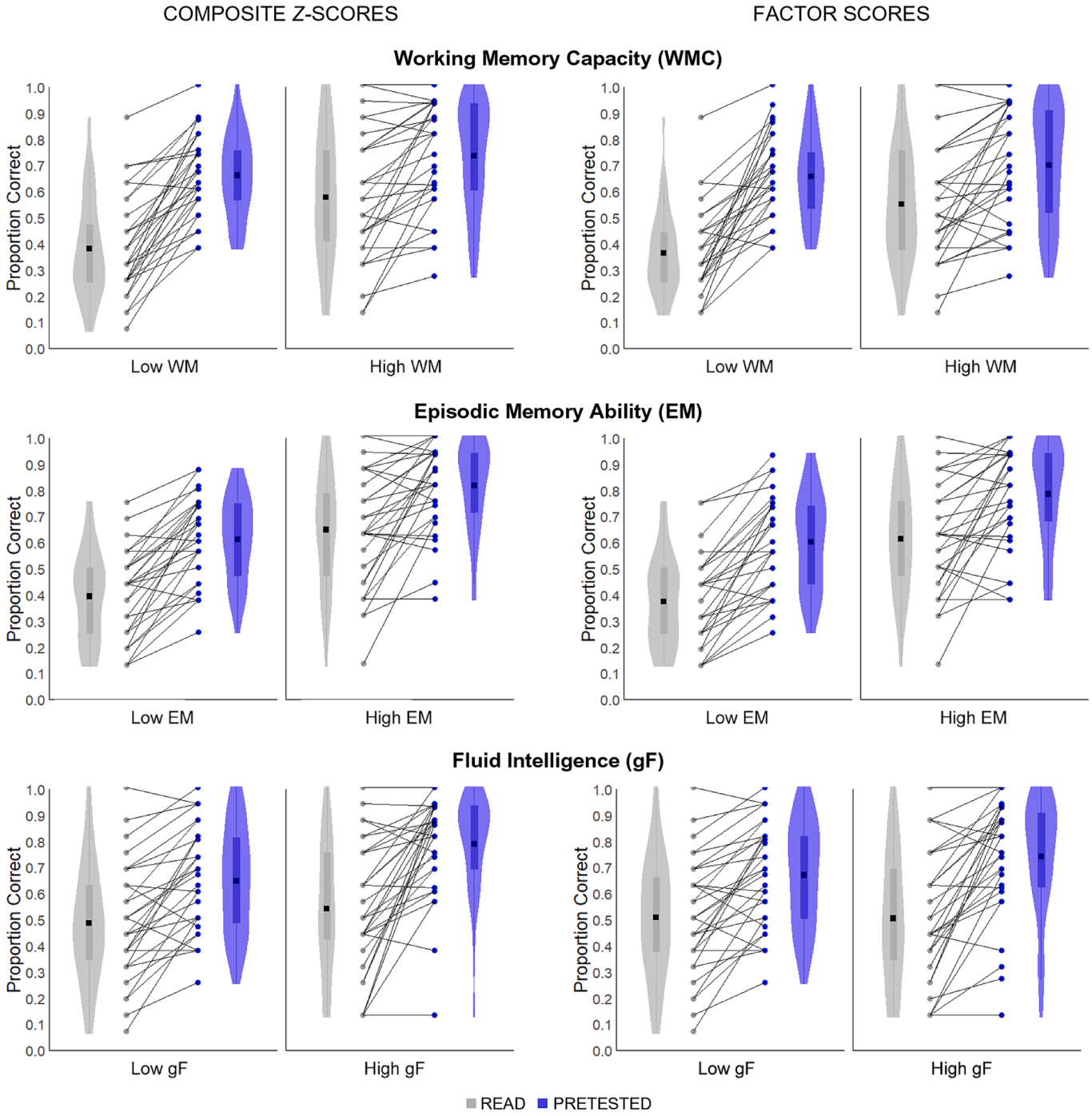

Fig. 5. Criterial test performance among the highest and lowest ability quartiles in study 2.

0.55).

Overall, although the analyses performed for Study 2 do not paint as consistent a picture as that obtained for Study 1, both the quartile-based analyses with composite and factor scores, as well as the linear mixed-effects model with WMC factor scores, suggest that WMC was associated with performance for read items. That performance, in turn, influenced the magnitude of the pretesting effect. In these analyses, lower-WMC individuals exhibited a larger pretesting effect than higher-WMC individuals.

Analyses involving educational attainment and age

Given the variation in educational attainment among the Study 2 sample, we were interested in whether the pretesting effect differed according to education status. To address this issue, we performed exploratory analyses similar to the quartile analyses described previously. There were statistically significant ($p < .0001$) pretesting effects among participants that at least had an undergraduate degree ($d = 1.16$) and among participants that did not ($d = 1.04$). Note that the effect size magnitude was quite similar between the two groups. Moreover, in a mixed-factors ANOVA on criterial test scores with factors of training condition (pretested vs. read) and education status (undergraduate degree or higher vs. no undergraduate degree), the interaction was not significant ($p = .58$). That result is consistent with the fact that performance levels were similar between groups for both training conditions (with pretested scores of $M_{\text{UGdegree}} = 0.68$ and $M_{\text{NoUGdegree}} = 0.72$; read scores of $M_{\text{UGdegree}} = 0.47$ and $M_{\text{NoUGdegree}} = 0.49$). Hence, educational attainment did not appear to influence the magnitude of the pretesting effect.

Given the diversity in age among the Study 2 sample, we were also interested in whether older versus younger participants performed differently. In another exploratory analysis, we submitted criterial test scores to a linear mixed-effects model with training condition (read = 0 vs. pretested = 1), age, and their interaction as predictors, including crossed random intercepts for participants. The interaction between training condition and age was not significant ($p = .956$), indicating that the pretesting effect did not vary across the sampled age range (21–45 years).

Comparison of cognitive abilities in studies 1 and 2

To explore potential differences in cognitive abilities between the samples from Study 1 and Study 2, which may help explain the divergent results across the studies, we conducted a series of exploratory analyses. These analyses involved combining data from both studies, calculating composite *z*-scores and factor scores for the merged dataset, and performing independent-samples *t*-tests for each cognitive ability. The analyses using composite scores found that participants in Study 1 had significantly higher scores for WMC, $t(255) = 4.02$, $p < .001$, EM ability, $t(254) = 2.25$, $p = .026$, and gF, $t(255) = 4.32$, $p < .001$, compared to participants in Study 2. The analyses using factor scores found that participants in Study 1 had significantly higher WMC scores versus those in Study 2, $t(255) = 3.49$, $p < .001$, as well as significantly higher gF scores, $t(255) = 3.70$, $p < .001$, whereas the EM ability factor scores were not significantly different ($p = .050$). Overall, these results indicate that the distributions of cognitive abilities differed between the two studies, with a higher ability sample present in Study 1 compared to Study 2.

Discussion

In the present research, a substantial pretesting effect was observed across two studies, replicating findings from prior studies using the same pretesting task and materials (e.g., Huelser & Metcalfe, 2012; Pan & Rivers, 2023). As anticipated, there was variability in the magnitude of that effect: While most participants exhibited a positive pretesting effect, approximately 20 % of participants in Study 1 and 15 % in Study 2 showed no effect or even a negative effect. We explored whether differences in working memory capacity, episodic memory ability, and/or fluid intelligence could help explain that variability. Results indicated that at least two of these cognitive abilities had an influence on the pretesting effect: In Study 1, fluid intelligence appeared to play a role, whereas in Study 2, apparent influences of working memory capacity and fluid intelligence were observed. In both studies, participants with lower scores on specific ability measures (gF in Study 1 and WMC in Study 2) showed larger pretesting effects compared to those with higher scores, driven largely by poorer performance for read items. Conversely, in Study 2, there were some indications that participants with higher gF scores also exhibited larger pretesting effects, driven by better performance for pretested items. Overall, although the results from both studies do not point to the exact same cognitive abilities in an identical manner, they clearly falsify the hypothesis that pretesting is equally beneficial for all learners. Rather, the advantages of pretesting appear to vary based on individual cognitive abilities.

Greater benefits of pretesting for lower-ability learners

A pattern was evident across both studies: With respect to gF in Study 1 and WMC in Study 2, lower-ability participants scored much lower than higher-ability participants in the read condition of the pretesting task, but their performance in the pretested condition was comparable. Scores for the relevant ability measures were only predictive of criterial test performance in the read condition and not performance in the pretested condition. That pattern was consistent across all analyses in Study 1 and for several analyses in Study 2 (linear-mixed effects model with WMC factor scores, quartile analyses with WMC composite scores, and quartile analyses with factor scores). A plausible interpretation of these results is that pretesting enabled lower-ability participants to learn pretested items at a level similar to that of higher-ability participants. That conclusion is consistent with the third hypothesis articulated at the outset of this manuscript, namely that lower-ability learners benefit more from pretesting. In other words, pretesting homologized memory ability across individuals.

As related evidence, consider the results of Pan and Rivers (2023). Re-analysis of their Experiments 1–4 indicates that participants with a positive pretesting effect scored lower in the read condition (*M*s = 0.38–0.43) than participants with null or negative pretesting effects (*M*s = 0.68–0.72), but both groups were comparable in the pretested condition (*M*s = 0.60–0.70). Thus, just as in the present research, variations in pretesting effect size were due to performance on read items, not on pretested items (for similar patterns see Cyr & Anderson, 2012).

Prior research from beyond the pretesting literature may help explain these patterns. First, with respect to gF, Minear et al. (2018) found that students scoring higher in gF also reported significantly higher use of "deep" processing strategies (e.g., imagery) and greater use of the keyword method (Fritz et al., 2007; Waldeyer & Roelle, 2021). If higher-gF participants already use superior memory strategies, then it is likely that they would exhibit better learning in a reading condition compared to lower-gF participants; moreover, when an effective memory strategy such as pretesting is used, then they should be expected to derive less learning benefits from it given that they already use other effective memory strategies (for related discussions see Brewer & Unsworth, 2012; Unsworth, 2019). Relatedly, Robey (2019) asked undergraduate participants to report the memory strategies that they used during a retrieval practice task and found that more effective strategy use was associated with smaller retrieval practice effects. It is conceivable in our view that similar patterns can occur for the case of pretesting, which would explain the results of Study 1.

With respect to WMC, studies indicate that higher-WMC individuals also tend to employ more effective memory strategies. For instance, Unsworth (2016) found that higher-WMC individuals were more likely to use strategies such as imagery and sentence generation during a free recall task. Similar findings have been reported by Bailey et al. (2008;

see also Unsworth & Spillers, 2010, and for the case of episodic memory, Kirchhoff, 2009). These findings support the possibility that, similar to gF, individual differences in strategy use among high-WMC versus low-WMC participants explain the results of Study 2.

It should be noted, however, that a larger benefit of pretesting for lower-ability individuals was not observed in the case of EM ability in either study. One possibility is that EM ability is a weaker predictor of the pretesting effect than other abilities. Alternatively, perhaps larger sample sizes may be needed to detect such patterns. Further research could address those possibilities.

Pretesting benefits for higher-ability learners and accounting for discrepant results

Discrepant results across studies also present challenges for interpretation. Notably, in Study 2, although gF did not significantly interact with training condition (pretested vs. read) when examined independently, it exhibited a significant interaction when analyzed alongside WMC and EM ability. In those analyses, a significant positive relationship emerged between gF scores and performance in the pretested condition, whereas no significant relationship was observed in the read condition. Those patterns align with the second hypothesis posited earlier (i.e., higher-ability learners can benefit more from pretesting) yet contrast with the results from Study 1, suggesting an alternative interpretation. Rather than the preexisting use of superior memory strategies among higher-gF individuals eliminating the need to engage in pretesting to enhance memory, these individuals may be better equipped to adapt to and learn from pretesting than their lower-gF counterparts. Compared to the proposal that lower-ability learners may experience larger pretesting effects, however (which, as previously noted, aligns with prior research), we consider this explanation more tentative.

Why might such a pattern have emerged in Study 2 only when gF was analyzed alongside other cognitive abilities, and moreover, why was an interaction involving WMC (that was observed repeatedly when WMC was analyzed separately) absent in the simultaneous analyses? To begin, it is important to consider the strong correlations between WMC and gF (Kyllonen & Christal, 1990; Shipstead et al., 2016). If the relationship between gF and performance in the pretested and/or read conditions relies on the presence of other cognitive abilities, such as WMC, then this relationship may only become evident when gF is considered in conjunction with these abilities. Moreover, the lack of a significant interaction involving WMC in the simultaneous models could be attributed to the overlap in variance explained by gF and WMC. Further research is needed, however, to further investigate these possibilities.

It is also important to consider the differences in sample characteristics between studies 2, with Study 1 comprising a higher ability sample. The differing ranges of cognitive abilities in the two samples may lead to corresponding differences in individual differences patterns (for related discussions, see Brewer et al., 2021; Pan et al., 2015). Further, these studies were conducted in different settings, and responses to pretesting procedures may vary between a laboratory setting with introductory psychology students versus an online platform with participants that have extensive experience with a variety of psychology studies. Relatedly, the authors of this manuscript have occasionally observed differences in the efficacy of various learning strategies (e.g., interleaved practice) when conducted in the laboratory versus online (Pan et al., 2025).

Theoretical implications for pretesting and other learning strategies

The current findings can be interpreted from the lens of major theoretical accounts of pretesting effects and related memory phenomena. For instance, if high-gF or high-WMC individuals generate effective mediators more readily when learning verbal materials, then there would be less need for pretesting to facilitate such generation (if the semantic mediator account is correct), in turn leading to smaller pretesting effects. Another possibility is that in such individuals, the more extensive use of effective memory strategies may obviate or lessen the need for the activation of a search set, encoding of an episodic memory of the pretesting event, or learning in response to an error signal in order to enhance memory. In low-gF or low-WMC individuals, on the other hand, any of those processes might be especially beneficial and enhance memory (although for evidence that low-WMC individuals are less effective at generating and using search sets, see Unsworth, 2009). Alternatively, higher-ability learners may experience greater benefits from pretesting in certain contexts, particularly if the task demands exceed their usual strategies or if pretesting effectively activates deeper cognitive processing.

Other factors unrelated to strategy use may also contribute to the observed patterns. For example, high-WMC individuals may be less prone to interference in paired-associate list learning paradigms (e.g., Rosen & Engle, 1998; although see Oberauer et al., 2004). Recent findings by Kliegl et al. (2023) show that pretesting can protect learners from retroactive interference caused by studying additional word pair lists. Although that research and the present studies were differently designed, participants may have experienced interference during the training phase of the pretesting task in Studies 1 and 2, and low-WMC individuals, whom may be more susceptible to such interference, benefitted especially from pretesting as a result.

The finding that pretesting benefits learners of differing abilities to varying degrees also raises the prospect that other learning strategies may benefit learners differently. For instance, strategies such as retrieval practice, distributed practice, elaboration, or the keyword method may ultimately be more advantageous for learners that typically rely on less efficient encoding methods (for related discussions, see Bjork et al., 2013; Pan & Bjork, 2022; Unsworth, 2019). If so, then tailoring a variety of learning strategies towards learners of varying abilities may be especially impactful.

Interpretative considerations, study limitations, and future research

Mindful of sample size limitations (i.e., our studies would have benefited from stronger statistical power to detect smaller effects), we considered pooling data from both studies for combined analyses. Robey (2019) and Unsworth (2019) demonstrated the value of pooled data in revealing significant individual differences in retrieval practice effects. Relatedly, Schönbrodt and Perugini (2013) showed through simulations that a sample size approaching 250 is necessary for stable correlational estimates in many scenarios (see also Kretzschmar & Gignac, 2019). Given differences in age, academic background, and study experience, however—which could complicate the interpretation of a combined analysis—we opted for separate analyses for Studies 1 and 2. As an additional concern, Krefeld-Schwalb et al. (2024) recently showed differences in gF, crystallized intelligence, attentiveness to study instructions, and prior research experience among participants from different crowdsourcing platforms, including Prolific Academic, and argued that those differences explain variability observed across those platforms even for well-established psychological phenomena.

To address these issues, future research could feature larger and more diverse samples. From a cognitive ability perspective, the samples used in Study 1 and Study 2 may not fully represent the range of cognitive abilities present in the general population. Samples with wider ability ranges may help better determine whether the relationship between the pretesting effect and WMC, EM ability, or gF is linear, curvilinear, or follows another pattern that may not yet be understood. Larger sample sizes would also better support the use of multiple-factor CFA to generate factor scores.

Another limitation of the present research is the lack of an extended retention interval in the pretesting task. Although the pretesting effects observed in our studies were substantial ($ds = 0.93, 1.13$), they may be even more pronounced with longer retention intervals, such as several days (Kliegl et al., 2022). Investigating the pretesting effect with

different learning materials could provide valuable insights (for related discussion, see Cronbach & Snow, 1977). Examining the strategies individuals employ during pretesting may also be useful.

Finally, future studies could explore alternative data analysis approaches. Our multipronged approach featured a series of analyses, each with its unique advantages and disadvantages. Different scoring approaches (for instance, using absolute scoring for the WMC tasks rather than partial credit scoring) could be considered. Additionally, some researchers caution against the use of quartile-based analyses due to the data-driven nature of cut-off points and issues of statistical power (Bennette & Vickers, 2012). Although that approach has precedent in individual differences research, we recognize that no single analysis is necessarily definitive and encourage future researchers to emphasize the most robust analytical approaches wherever feasible.

Pedagogical implications

In the present research, positive pretesting effects were observed in approximately 80 % of all participants, and the magnitude of those effects did not depend on whether learners had attained an undergraduate degree or not. The pretesting effect also remained largely consistent across the sampled age range. It is therefore likely that pretesting enhances memory for many or most learners, at least to some extent. Nevertheless, pretesting did not benefit all learners equally. In particular, there were indications that lower-ability learners benefited more, although the patterns were not entirely consistent across studies, and an opposite pattern may exist in some cases. Ultimately, our findings suggest that pretesting may be especially effective as a targeted intervention, with a strong possibility being that it may boost the memory performance of lower-ability learners to levels similar to their peers.

CRediT authorship contribution statement

Steven C. Pan: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Liwen Yu:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis. **Marcus J. Wong:** Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Ganeash Selvarajan:** Software, Methodology, Data curation. **Andy Z.J. Teo:** Validation, Software, Methodology, Formal analysis.

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.

Acknowledgement

The authors thank Jolynn Pek for analysis suggestions and comments on an earlier draft of this manuscript, Ahmad Naufal, Wen Kai Ling, Xinyi Lim, Yilin Hong, and Zihan Cui for help with data collection, Sze Lin Tung for technical assistance, Nathaneal Teo and Yilin Hong for work on data processing and/or data reproducibility checks, Alison Robey for sharing analysis code, and Michelle Kaku for administrative support. Thanks also to Alexander Burgoyne and Gene Brewer for sharing stimulus materials, as well as Faria Sana for assistance with software resources. This research was supported by a National University of Singapore Faculty of Arts & Social Sciences (FASS) grant awarded to S. C. Pan.

Data and analysis code are available at the Open Science Framework (OSF) and can be accessed at https://osf.io/r6t95/. Materials are on OSF at https://osf.io/8g3fp/. The authors declare no conflict of interest.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104608.

Data availability

Link to data/code have been shared at the submission step.

References

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

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition, 36*(8), 1383–1390. https://doi.org/10.3758/MC.36.8.1383

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bennette, C., & Vickers, A. (2012). Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology, 12*(1), 21. https://doi.org/10.1186/1471-2288-12-21

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417–444.

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*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition, 6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Chen, H. Y., Gilmore, A. W., Nelson, S. M., & McDermott, K. B. (2017). Are there multiple kinds of episodic memory? An fMRI investigation comparing autobiographical and recognition memory tasks. *Journal of Neuroscience, 37*(10), 2764–2775.

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

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*(5), 769–786. https://doi.org/10.3758/BF03196772

Cyr, A.-A., & Anderson, N. D. (2012). Trial-and-error learning improves source memory among young and older adults. *Psychology and Aging, 27*(2), 429–439. https://doi.org/10.1037/a0025115

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 841–850. https://doi.org/10.1037/xlm0000073

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In *Psychology of Learning and Motivation* (Vol. 44, pp. 145–199). Elsevier. Doi: 10.1016/S0079-7421(03)44005-X.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods, 37*(4), 581–590. https://doi.org/10.3758/BF03192728

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology, 21*(4), 499–526. https://doi.org/10.1002/acp.1287

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Gupta, M. W., Pan, S. C., & Rickard, T. C. (2024). Interaction between the testing and forward testing effects in the case of Cued-Recall: Implications for Theory, individual difference Studies, and application. *Journal of Memory and Language, 134*, Article 104476. https://doi.org/10.1016/j.jml.2023.104476

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40*(4), 514–527. https://doi.org/10.3758/s13421-011-0167-z

Inquisit 5 (2016). Retrieved from https://www.millisecond.com.

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition, 41*(5), 625–637. https://doi.org/10.3758/s13421-013-0298-5

Janelli, M., & Lipnevich, A. A. (2021). Effects of pre-tests and feedback on performance outcomes and persistence in Massive Open Online Courses. *Computers & Education, 161*, Article 104076. https://doi.org/10.1016/j.compedu.2020.104076

Jung, S., & Lee, S. (2011). Exploratory factor analysis for small samples. *Behavior Research Methods, 43*, 701–709.
Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology, 103*(1), 48–59. https://doi.org/10.1037/a0021977
Kievit, R. A., Davis, S. W., Griffiths, J., Correia, M. M., Cam-CAN, & Henson, R. N. (2016). A watershed model of individual differences in fluid intelligence. *Neuropsychologia, 91*, 186–198. doi: 10.1016/j.neuropsychologia.2016.08.008.
Kirchhoff, B. A. (2009). Individual differences in episodic memory: The Role of self-initiated encoding strategies. *The Neuroscientist, 15*(2), 166–179. https://doi.org/10.1177/1073858408329507
Kliegl, O., Bartl, J., & Bäuml, K.-H.-T. (2022). The pretesting effect comes to full fruition after prolonged retention interval. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1037/mac0000085
Kliegl, O., Bartl, J., & Bäuml, K.-H.-T. (2023). The pretesting effect thrives in the presence of competing information. *Memory, 31*(5), 705–714. https://doi.org/10.1080/09658211.2023.2190568
Knight, J. B., Hunter Ball, B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*(4), 731–746. https://doi.org/10.1016/j.jml.2011.12.008
Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998. https://doi.org/10.1037/a0015729
Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283–294. https://doi.org/10.1037/a0037850
Kornell, N., & Vaughn, K. E. (2016). *How retrieval attempts affect learning. In Psychology of Learning and Motivation* (Vol. 65,, 183–215.
Krefeld-Schwalb, A., Sugerman, E. R., & Johnson, E. J. (2024). Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proceedings of the National Academy of Sciences, 121*(12), Article e2306281121.
Kretzschmar, A., & Gignac, G. E. (2019). At what sample size do latent variable correlations stabilize? *Journal of Research in Personality, 80*, 17–22. https://doi.org/10.1016/j.jrp.2019.03.007
Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1
Leonard, B., Hake, H., & Stocco, A. (2023). Faulty memories, favored outcomes: How errors impact learning processes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/6175j6zs.
Mera, Y., Rodriguez, G., & Marin-Garcia, E. (2021). Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic Bulletin & Review, 13*.
Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology, 68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022
Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia, 138*, Article 107296. https://doi.org/10.1016/j.neuropsychologia.2019.107296
Millisecond Software (2022). *User manual: Inquisit automated operation span.* Retrieved from https://www.millisecond.com/download/library/v6/ospan/automatedospan/automatedospan.manual.
Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (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
Murman, D. L. (2015). The impact of age on cognition. *Seminars in Hearing, 36*(3), 111–121. https://doi.org/10.1055/s-0035-1555115
Oberauer, K., Lange, E., & Engle, R. W. (2004). Working memory capacity and resistance to interference. *Journal of Memory and Language, 51*(1), 80–96. https://doi.org/10.1016/j.jml.2004.03.003
Palan, S., & Schitter, C. (2017). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*. https://doi.org/10.1016/j.jbef.2017.12.004
Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educational Psychology Review, 35*(4), 97. https://doi.org/10.1007/s10648-023-09814-5
Pan, S. C., Dunlosky, J., Xu, K. M., & Ouwehand, K. (2024). Emerging and future directions in test-enhanced learning research. *Educational Psychology Review, 36*(1), 20. https://doi.org/10.1007/s10648-024-09857-2
Pan, S. C., Lovelett, J., Stoeckenius, D., & Rickard, T. C. (2019). Conditions of highly specific learning through cued recall. *Psychonomic Bulletin & Review, 26*(2), 634–640. https://doi.org/10.3758/s13423-019-01593-x
Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language, 83*, 53–61. https://doi.org/10.1016/j.jml.2015.04.001
Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*. https://doi.org/10.3758/s13421-022-01392-1
Pan, S. C., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology: Applied, 27*(2).
Pan, S. C., Sana, F., Schmitt, A. G., & Bjork, E. L. (2020). Pretesting reduces mind wandering and enhances learning during online lectures. *Journal of Applied Research in Memory and Cognition, 9*(4), 542–554. https://doi.org/10.1016/j.jarmac.2020.07.004
Pan, S. C., Yu, L., Hong, Y., Wong, M. J., Selvarajan, G., & Kaku, M. E. (2025). Individual differences in fluid intelligence moderate the interleaving effect for perceptual category learning. *Learning and Individual Differences, 117*, 102603. https://doi.org/10.1016/j.lindif.2024.102603
Permut, S., Fisher, M., & Oppenheimer, D. M. (2019). TaskMaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science, 2*(2), 188–196. https://doi.org/10.1177/2515245919838479
R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.
Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In *Handbook of nonverbal assessment* (pp. 223–237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11.
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
Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243–257.
Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language, 108*, Article 104029. https://doi.org/10.1016/j.jml.2019.104029
Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003
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
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.
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
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
Seabrooke, T., Mitchell, C. J., Wills, A. J., Inkster, A. B., & Hollins, T. J. (2021). The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect. *Memory & Cognition*. https://doi.org/10.3758/s13421-021-01218-6
Shipstead, Z., Harrison, T., & Engle, R. (2016). Working memory capacity and fluid intelligence: maintenance and disengagement. *Perspectives on Psychological Science, 11*, 771–799. https://doi.org/10.1177/1745691616650647
Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*. ix + 121–ix + 121.
Unsworth, N. (2009). Variation in working memory capacity, fluid intelligence, and episodic recall: A latent variable examination of differences in the dynamics of free recall. *Memory & Cognition, 37*(6), 837–849. https://doi.org/10.3758/MC.37.6.837
Unsworth, N. (2016). Working memory capacity and recall from long-term memory: Examining the influences of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 50–61. https://doi.org/10.1037/xlm0000148
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. (2009). There's more to the working memory capacity—Fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review, 16*(5), 931–937. https://doi.org/10.3758/PBR.16.5.931
Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038–1066. https://doi.org/10.1037/0033-2909.133.6.1038
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., & Spillers, G. J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review, 17*(2), 200–205. https://doi.org/10.3758/PBR.17.2.200
Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19*, 899–905.
Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition, 41*(1), 1–15. https://doi.org/10.3758/s13421-012-0246-9
Waldeyer, J., & Roelle, J. (2021). The keyword effect: A conceptual replication, effects on bias, and an optimization. *Metacognition and Learning, 16*(1), 37–56. https://doi.org/10.1007/s11409-020-09235-7
Wang, J., Liu, Z., Xing, Q., & Seger, C. A. (2020). The benefit of interleaved presentation in category learning is independent of working memory. *Memory, 28*(2), 285–292. https://doi.org/10.1080/09658211.2019.1705490
Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science, 7*(1), Article 25152459231213802.
Wingert, K. M., & Brewer, G. A. (2018). Methods of studying individual differences in memory. In H. Otani & B. L. Schwartz (Eds.), *Handbook of Research Methods in*

Human Memory (1st ed., pp. 443–458). Routledge. Doi: 10.4324/9780429439957-24.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913–934.

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