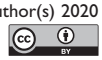
Article

PSYCHOLOGY LEARNING & TEACHING

Psychology Learning & Teaching
2021, Vol. 20(1) 21–39



DOI: 10.1177/1475725720973494
journals.sagepub.com/home/plj

Retrieval Practice: Beneficial for All Students or Moderated by Individual Differences?

Frida Bertilsson
Department of Psychology, Umeå University, Sweden

Tova Stenlund
Department of Psychology, Umeå University, Sweden

Carola Wiklund-Hörnqvist
Department of Psychology, Umeå University, Sweden

Bert Jonsson
Department of Applied Educational Science, Umeå University, Sweden

Abstract

Retrieval practice is a learning technique that is known to produce enhanced long-term memory retention when compared to several other techniques. This difference in learning outcome is commonly called "the testing effect". Yet there is little research on how individual differences in personality traits and working memory capacity moderate the size of the retrieval-practice benefits. The current study is a conceptual replication of a previous study, further investigating whether the testing effect is sensitive to individual differences in the personality traits Grit and Need for Cognition, and working memory capacity. Using a within-subjects design ($N = 151$), participants practiced 60 Swahili–Swedish word pairs (e.g., adhama–honor) through retrieval practice and re-studying. Learning was assessed at three time points: five minutes, one week, and four weeks after practice. The results revealed a significant testing effect at all three time points. Further, the results showed no association between the testing effect and the personality traits, or between the testing effect and working memory, at any time point. To conclude, retrieval practice seems to be a learning technique that is not moderated by individual differences in these specific personality traits or with working memory capacity, thus possibly beneficial for all students.

Corresponding Author:
Frida Bertilsson, Department of Psychology, Umeå University, SE-901 87 Umeå, Sweden.
Email: frida.bertilsson@umu.se

Keywords
Retrieval practice, the testing effect, individual differences, personality traits, working memory capacity

Retrieval Practice: Beneficial for All Students or Moderated by Individual Differences?

Our ability to learn is affected by individual differences in a number of different attributes related to learning, such as working memory capacity (WMC); personality traits; social, emotional, and physical factors; and teaching competence. This means that students' experiences are partly based on their prerequisites for learning. An aim for the educational community should be to give all students an equal chance to succeed in reaching the required educational goals. One way of giving equal opportunities is by investigating and applying evidence-based learning techniques that are beneficial for learning, irrespective of individual prerequisites. The present study focuses on the learning technique retrieval practice and its relation to working memory and personality characteristics.

Retrieval practice is a learning technique that repeatedly has been shown to enhance long-term retention when compared to other methods of learning, such as re-reading (Roediger & Karpicke, 2006a; Wiklund-Hörnqvist et al., 2014), group discussions (Stenlund et al., 2017), and concept mapping (Karpicke & Blunt, 2011). This retrieval-based benefit on long-term learning is commonly denoted as the *testing effect* (for reviews, see Dunlosky et al., 2013; Roediger & Karpicke, 2006a, 2006b; Rowland, 2014). There are several explanations for the testing effect. For example, the desirable difficulties framework (Bjork, 1994) states that desirable "struggle" during learning improves long-term retention (for related arguments, see the retrieval effort hypothesis, Pyc & Rawson, 2009; the elaborative retrieval hypothesis, Carpenter & Delosh, 2006; the mediator effectiveness hypothesis, Pyc & Rawson, 2010). Another account is the transfer appropriate processing hypothesis (TAP) (Morris et al., 1977) stating that if the cognitive processing during learning matches the subsequent retrieval event, long-term memory is enhanced. A recent additional explanation is the episodic context account (Karpicke et al., 2014; Lehman et al., 2014). According to this account, the superior retention of the to-be-learned material following retrieval practice is related to a reinstatement of the contextual features associated with the target. During retrieval practice, the to-be-learned material is continuously updated with contextual features, making the memory *richer*. This on-going process restricts the search set and increases the likelihood of successfully recovering a target in the future (Karpicke et al., 2014; Lehman et al., 2014)

The testing effect is a well-established empirical phenomenon, and many ways to increase its effectiveness have been identified. For example, it has been found that some test formats are more favorable than others (Kang et al., 2007; Stenlund et al., 2016), that a number of successful retrievals are required to give a lasting effect (Rawson & Dunlosky, 2011), and that correct answer feedback is beneficial for learning (Kang et al., 2007; Roediger & Butler, 2011; Wiklund-Hörnqvist et al., 2014) and retention (Pashler et al., 2005). Providing correct answer feedback keeps the learner from learning the wrong answer and also gives an opportunity to remember the correct answer (Roediger & Butler, 2011). Another important aspect

of including feedback in retrieval-practice research is to equate exposure to the learning materials between learning conditions (Kang et al., 2007).

Some aspects of the effectiveness of retrieval practice are still uncertain. For example, it is still underexplored whether individual differences in personal attributes associated with learning will moderate the effect of retrieval practice, or if the method is equally beneficial regardless of academic aptitude. Tse and Pu (2012) demonstrated that cognition and personality can interact in such a way that people with higher levels of a personality trait, such as trait test anxiety, and lower cognitive levels (measured by WMC) benefit less from retrieval practice than people with a higher cognitive level. However, in a follow-up study, Tse et al. (2019) were unable to replicate the results, which calls for further research.

In this study, the focus is on investigating the impact of personality traits and WMC on the testing effect, in order to bring more clarity into this area. We argue that it is important to evaluate whether individual characteristics that are said to influence school performance (see e.g., Alloway & Alloway, 2010; Cacioppo et al., 1996; Duckworth & Quinn, 2009) are also critical for the effects of retrieval practice.

Research has identified several personality traits that affect people's aptitude for learning (Arbabi et al., 2015; Duckworth & Quinn, 2009; Lounsbury et al., 2003; Sadowski & Gulgoz, 1992). One specific trait is Grit, which has received much attention in recent years. Grit is defined as "perseverance and passion for long term goals" (Duckworth et al., 2007, p. 1087) and contains two subconstructs—that is, consistency of interest and perseverance of effort. Grit has been found to be predictive of both academic performance and other types of success (e.g., completion of the summer training program at West Point Military Academy and performance in Scripps National Spelling Bee; Duckworth & Quinn, 2009). However, Grit is known to have a strong positive correlation with Big Five Conscientiousness ($r = .77$; Duckworth et al., 2007; Duckworth & Quinn, 2009; Meriac et al., 2015), and some researchers question the distinctness of Grit from Conscientiousness (Credé et al., 2017). Another personality trait is Need for Cognition (NFC), defined as "the tendency for an individual to engage in and enjoy thinking" (Cacioppo & Petty, 1982, p. 116). NFC explains individual differences in motivation and effort when engaging in cognitive activities (van Seggelen-Damen, 2013). The concept of NFC has been examined for a long time and shown to be positively related to enhanced academic performance (Sadowski & Gulgoz, 1992), attending gifted classes (Meier et al., 2014), problem solving (Cacioppo et al., 1996), and the use of more advanced strategies for learning (Cazan & Indreica, 2014).

It seems likely that theoretical explanations for the testing effect emphasizing the advantage of retrieval effort should be more influenced by Grit and NFC, while explanations such as TAP would be less influenced. The present study focuses on the acquisition of a foreign language vocabulary for upper-secondary school students. The typical experimental procedure includes a learning phase and a manipulation phase (retrieval vs re-study practice) followed by retention tests assessing learning (in this case, five minutes, one week, and four weeks after the learning phase). Participants with high Grit should be able to persevere through the learning phase to a higher extent and thus learn more and perform better across all three retention tests, while individuals with high NFC should engage more in assignments involving thinking and would therefore perform better at the retention tests irrespectively of material. High motivation and effort when engaging in cognitive activities, which, as pointed out above, characterize individuals with high NFC, are aspects found to be especially important for test performance (Unsworth et al., 2013; Van Barneveld, 2007). Further, knowledge of effective strategies (for example, retrieval practice), which also characterize

individuals with high NFC, should be especially helpful for long-term memory consolidation (see e.g., Antony et al., 2017) independent of cognitive ability.

Only two previous studies have specifically examined NFC and Grit in relation to the effects of retrieval practice (Bertilsson et al., 2017; Stenlund et al., 2017). Stenlund et al. (2017) used a between-subjects design ($N = 98$) to compare the learning effects of retrieval practice with group discussions with or without feedback, and whether retention was influenced by NFC. The results showed no relationship between NFC and the testing effect. Bertilsson et al. (2017) conducted two experiments, in which the participants learned Swedish–Swahili word pairs. Experiment 1 ($N = 39$) investigated the effect of retrieval practice relative to repeated studying using a between-subjects design. In Experiment 2 ($N = 29$), a within-subjects design was employed, and all participants used both retrieval practice and re-study to learn the materials. The learning outcome was assessed by means of cued recall tests at three different time points: immediately, one week, and four weeks after the intervention. The result in both experiments showed that neither Grit nor NFC were related to the effect of retrieval practice. While these findings are interesting, the conclusions are based on only two studies with rather few participants, and only one of the studies included Grit (i.e., Bertilsson et al., 2017). To make firm conclusions and be able to generalize the findings, more research is needed targeting both NFC and Grit using a larger sample.

Beside specific personality traits, various cognitive abilities have been shown to have a significant impact on our ability to learn. Studies show that students with high WMC (Alloway & Alloway, 2010; Cowan, 2014), executive functioning (St. Clair-Thompson & Gathercole, 2006), or IQ (Deary et al., 2007) perform better in school than students who possess lower abilities in these areas. Working memory (Baddeley, 2010) has been suggested to have an essential role in a number of skills required for being successful in school, as well as for coping well with classroom activities in general (see Alloway, 2006, for a review). There are several hints that WMC may play an important role in learning word pairs (Swedish–Swahili word pairs were used in the present study) and to retrieving them across a period of four weeks. For example, studies have shown that the frontal lobe is critical during the acquisition of vocabulary (Karlsson Wirebring et al., 2015) and that the search process for long-term memory retrieval is driven by WMC (Unsworth et al., 2013). However, prior studies investigating the relationship between WMC and the testing effect have so far produced quite differing results. In a sample of college students who were instructed to learn general knowledge facts, Agarwal et al. (2017) found that on a delayed test two days after learning, retrieval practice improved performance for all students, but more so for low WMC students. In contrast, there is also a number of studies that have not found a relationship between WMC and the testing effect (Bertilsson et al., 2017; Brewer & Unsworth, 2012; Minear et al., 2018; Tse et al., 2019; Wiklund-Hörnqvist et al., 2014). One possible explanation for these mixed results is that the experiments use different methods. Many aspects of the retrieval practice intervention (i.e., type of material, number of items, amount of practice, the lag between practice and retention test) vary between experiments and are likely to impact the relationship between WMC and the testing effect.

When studying predictors of academic performance, researchers have repeatedly found that personality traits are predictive of academic performance over and above variance predicted by cognitive ability (e.g., O'Connor & Paunonen, 2007). Under some circumstances, personality traits are better predictors of academic performance than cognitive ability is (Chamorro-Premuzic & Furnham, 2008; Furnham et al., 2003). In the present study, we, therefore, controlled for WMC before entering the traits Grit and NFC in the

analyses with the aim to investigate the amount variance remaining after controlling for WMC.

In light of previous individual-difference research on the effects of retrieval practice, the purpose of this study is to further investigate whether the testing effect is sensitive to individual differences in Grit, NFC, and WMC, or whether the technique is equally beneficial for all students. One way to achieve this is to verify results from previous studies through replication. Replicability is a vital part of all research since conclusions drawn from results are not valid if the results cannot be replicated (Asendorpf et al., 2013). This is especially true when there are very few studies, as with personality and the testing effect, and when the results are inconclusive, as with WMC and the testing effect. Of the limited number of studies that have included personality traits, only one has used a within-subjects design (i.e., Bertilsson et al. 2017), which is preferable when investigating the effects of individual differences, but it included a small number of participants. Another strength with the experimental design of Experiment 2 in Bertilsson et al. (2017) is that retention was measured at three different time points, and both accumulated and uniquely tested word pairs were included at each retention test, making it possible to investigate the relationship between the independent variables and the testing effect at different intervals. The present study will, therefore, replicate the design, procedure, and, partially, the statistical analyses of Experiment 2 in Bertilsson et al. (2017), using a larger sample. In addition, while Bertilsson et al. (2017) examined individual differences in learning effects from retrieval practice and re-study practice separately, the present study makes an innovative contribution by extracting the performance difference scores between the learning conditions (retrieval practice vs re-study practice) at each retention interval. This calculation is necessary when examining individual differences in relation to the testing effect with a within-subject design, as separate analyses of the two conditions might cause problems with validity.

Methods

Participants

In total, 196 students (49.5% female; $M_{age} = 17.2$, $SD = .65$) from two types of study programs (natural sciences and social sciences) were recruited from an upper-secondary school in northern Sweden. Thirty-eight students did not complete all parts of the study and were therefore excluded. Seven outliers were identified in the measure for WMC using an interquartile range of 1.5, and the corresponding cases were excluded from all analyses. No outliers were identified in the measures of Grit and NFC. This resulted in a final sample of 151 participants with ages ranging from 16 to 20 years (45.7% female; $M_{age} = 17.1$, $SD = .62$). The participants received two movie tickets as a reimbursement for their participation, and written informed consent was obtained in accordance with the Declaration of Helsinki. The study was approved by the Regional Ethical Review Board, Sweden (2017/517-31).

Based on Bertilsson et al. (2017; Experiment 2, $N = 29$) two a priori power analyses were conducted using G*Power 3.1.9.7 (Faul et al., 2009). The first a priori power analysis (repeated-measures ANOVA) targeted the testing effects at each retention interval (five minutes, one week, four weeks) with the lowest effect size ($\eta_p^2 = 0.80$) from Bertilsson et al. (2017, Experiment 2) as input. The analysis indicated that with an alpha of 0.05 and a statistical power of 0.95, a minimum of six participants is required, showing that the study by Bertilsson et al. (2017) had a sufficient sample size. The second a priori power

analysis (multiple linear regression) targeted the effects of the Mental Effort Tolerance Questionnaire (METQ; Dornic et al., 1991) and Grit at each retention interval (five minutes, one week, four weeks). The correlations between METQ and retrieval practice word pairs, and Grit and retrieval practice word pairs at each retention interval was entered as input in the power analysis. The analysis indicated that with an alpha of 0.05 and a statistical power of 0.95, a sample size of 89 participants is required, which extends that of Bertilsson et al.'s (2017) sample size in Experiment 2.

Materials

The material used in the learning intervention consisted of 60 Swahili–Swedish word pairs (Karlsson Wirebring et al., 2015; Nelson & Dunlosky, 1994). Further, three instruments were used to measure WMC and personality; these are described below.

WMC. An automated version of the Operation Span task (Ospan; Unsworth et al., 2005), a complex working memory task, was used to measure WMC. The automated Ospan is administered on a computer and can be completed by the participants independently from the experiment leader. The task is comprised of two subtasks — a letter span, and a concurrent math task — that are alternated so that a letter is presented between each math operation. The participant is required to solve the math task while at the same time maintaining the presented letters in memory. This continues for 3–7 trials before the participant is shown a matrix of 12 letters and is asked to recall the letters by clicking a box next to the letters in the order they were shown. To ensure that participants do not ignore the math tasks in favor of rehearsing the letters (thus measuring short-term memory rather than working memory) there is an 85% accuracy criterion on the math tasks. The current percentage of correctly solved math tasks is displayed to the participant during letter recall. Unsworth et al. (2005) reported a good test-retest reliability, $r = .83$, and internal consistency, $\alpha = .78$. They found the automated Ospan to be a valid measure of WMC as it was significantly related to the original Ospan (Turner & Engle, 1989), $r = .45$, and the two measures correlated similarly with Ravens Progressive Matrices, a measure of fluid abilities (automated Ospan, $r = .38$; original Ospan, $r = .42$) (Unsworth et al., 2005). The statistical analyses were conducted using the number of letters recalled in the correct position (i.e., partial credit load scoring, cf. Conway et al., 2005).

NFC. NFC was measured using the METQ (Dornic & Ekehammar, 1991), a Swedish adaption of the original NFC scale. The original scale is a well-validated measure of NFC (for meta-analysis, see Cacioppo et al., 1996) that was developed by Cacioppo and Petty (1982). The METQ contains 30 items with responses given using a 5-point Likert-type scale ranging from 1 (*Do not agree at all*) to 5 (*Agree completely*). The NFC score is the total sum of all items; however, 18 of the items are phrased negatively and therefore require reverse scoring. The internal consistency in the present study was $\alpha = .88$, which is in line with previous studies that have evaluated the psychometric properties of the questionnaire (Dornic et al., 1991; Stenlund & Jonsson, 2017). Psychometric evaluations have also found evidence of validity in the Swedish adaption of the NFC scale (Stenlund & Jonsson, 2017).

Grit. The Short Grit Scale (Grit-S; Duckworth & Quinn, 2009), an eight-item adaption of the original Grit Scale, was used to measure Grit. The questionnaire was translated to Swedish and independently back-translated to English by a professional translator to ensure good quality.

Grit-S contains two subconstructs—that is, consistency of interest and perseverance of effort, each measured by four items in the questionnaire. Responses are given using a 5-point Likert-type scale ranging from 1 (*Not like me at all*) to 5 (*Very much like me*). Half of the items are phrased negatively and require reverse scoring. To generate the Grit score, the sum of the items is divided by the number of questions. Grit-S has been reported to have good validity and reliability, with an internal consistency ranging between $\alpha = .73$ and .84 (Duckworth & Quinn, 2009). Similar levels of internal consistency have also been reported in previous studies using the Swedish translation (Bertilsson et al., 2017). In the present study, the internal consistency was $\alpha = .65$, which suggests that the scale has a questionable reliability in this case. This lower level might be explained by differences between the two subconstructs that the scale measures. Hence, a single construct scale would potentially have had a higher internal consistency than a scale that is built on two subscales (for an overview of the Grit scale, see Credé et al., 2017).

Design

A schematic overview over the study design and experimental procedure can be seen in Figure 1 (a)–(c). A 2×3 factorial within-subjects design was used, meaning that the participants learned half of the word pairs using retrieval practice—repeated testing with immediate feedback—and the other half using repeated studying. Three retention tests were given at different delays in order to assess the amount of learning: five minutes after learning, after one week, and after four weeks. In addition, the word pairs were randomly divided into three groups to be measured at different lags.

Procedure

The automated Ospan (WMC), METQ, and Grit were completed one week before the intervention as part of a data collection within a larger research project with the title

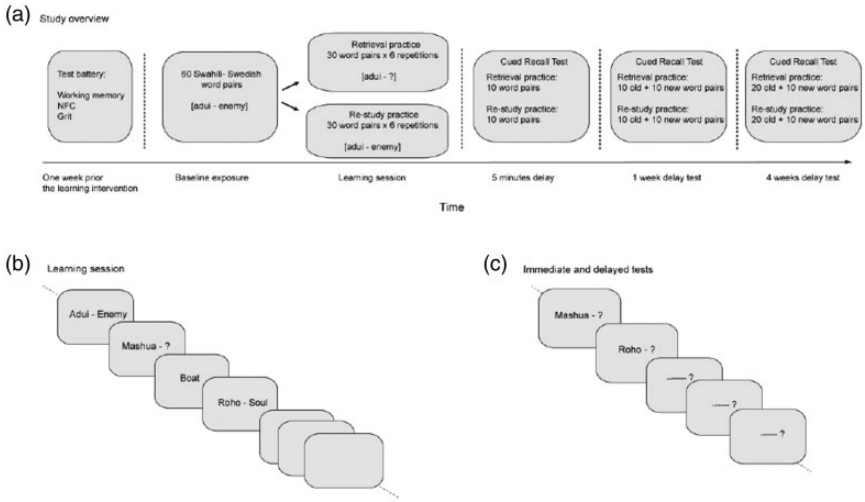


Figure 1. Study Design and Experimental Setup
Note. Schematic illustration depicting an overview of the experimental setup (a), the learning session (b), and the retention tests (c).

'Learning to engage the brain'. The intervention consisted of a learning session, and an assessment session including three retention tests.

Learning session. The learning session took place in the participants' classrooms during a class period and was conducted on individual computers using a web-based program that was designed to present the 60 Swahili–Swedish word pairs. Immediately prior to the learning session, all 60 word pairs were presented, one at a time, on the participants' computer screens in order to familiarize the participants with the material. Each pair was shown for eight seconds. Next, the learning session started and consisted of six practice rounds in which all participants learned half of the material via retrieval practice and the other half through repeated study (see Figure 1(b)). Re-study word pairs and retrieval practice word pairs were interleaved and randomly assigned to one of the two conditions on an individual level, meaning that retrieval practice and re-study word pairs differed between participants. The instructions explained that when both words in a pair were presented (Adhama–Honor, re-study condition) the participants should read and learn the words, and when only the Swahili word was presented (Mashua–?, retrieval practice condition) the participants were instructed to type in the Swedish equivalent. Word pairs in the retrieval practice condition were shown for eight seconds while the participants wrote the Swedish translation, followed by one second of correct answer feedback. To equate exposure to the material in the two conditions, word pairs in the re-study condition were shown for nine seconds each.

Immediate and Delayed Tests. Retention was assessed by means of a cued recall test at three separate time points (see Figure 1(c)). A five-minute break separated the learning phase from the retention test. At each retention interval (5 minutes, 1 week, and 4 weeks) participants were tested on 20 unique Swahili–Swedish word pairs, 10 re-study, and 10 retrieval practice pairs. In addition, the word pairs tested in prior retention rests were included in the subsequent tests as well. As a result of this procedure, each retention test contained an increasing number of word pairs. For example, after 4 weeks, participants were tested on all 60 word pairs. Of those 60 pairs, 20 had previously been tested after 5 minutes and 1 week, 20 had previously been tested after 1 week, and the final 20 (10 re-study and 10 retrieval practice pairs) were unique to the 4-week test. See Table 1 for a correlation matrix of all variables used in this study.

Results

The alpha level was set to .05, and as measures of effect size, partial eta square (η_p^2) and coefficient of determination (r^2) were used, where applicable. Greenhouse-Geisser corrected degrees of freedom were reported when Mauchly's test indicated that the assumption of sphericity had been violated.

First, a factorial 2×3 repeated-measures ANOVA was used to investigate changes in retention for retrieval practice and re-study word pairs, with the variables retention interval (five minutes, one week, and four weeks) and practice condition (retrieval practice and re-study practice) as within-subjects factors. The analysis was conducted using the word pairs that were unique to each retention test—that is, 20 word pairs at each interval. The results revealed main effects of retention interval, $F(1.8,\ 268) = 396.687$, $p < .001$, $\eta_p^2 = .73$, and practice condition, $F(1,\ 150) = 248.982$, $p < .001$, $\eta_p^2 = .62$, as well as a practice condition $\times$ retention interval interaction, $F(2,\ 300) = 6.247$, $p = .002$, $\eta_p^2 = .04$.

Table 1. Mean, Standard Deviation and Pearson Correlation Matrix for the Predictor Variables WMC, Grit, and NFC, as well as Recall Performance After Various Delays for Re-study and Retrieval Practice Word Pairs.

Variables	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.
1. WMC	53.17	12.44	–	−.04	.31**	.15	.21*	.24**	.09	.10	.22**	.30**	.09	.20*	.14	.07	.17*	−.10	.12	.05
2. Grit	3.37	.53		–	.27**	.12	.05	.02	.14	.10	.10	.08	.09	.00	.04	−.02	.06	.08	−.14	−.05
3. NFC	102.05	15.58			–	.29**	.20*	.20*	.22*	.19*	.27**	.31**	.12	.21*	.20*	−.03	.20*	−.03	.00	.02
4. Re-study 5 min.	.52	.27				–	.63**	.35**	.74**	.68**	.67**	.65*	.33**	.56**	.51**	−.44**	.20*	.07	−.13	−.13
5. Re-study 1 week uni.	.25	.20					–	.37**	.62**	.56**	.54**	.65**	.36**	.62**	.59**	−.14	−.18*	.08	.04	.07
6. Re-study 4 weeks uni.	.11	.16						–	.35**	.36**	.36**	.45**	.41**	.35**	.33**	.00	.20*	−.35**	.03	.00
7. Re-study 1 week acc.	.69	.26							–	.87**	.51**	.58**	.27**	.53**	.48*	−.30**	.12	.01	−.43**	−.33**
8. Re-study 4 weeks acc.	.45	.25								–	.46**	.51**	.29**	.47**	.46**	−.30**	.09	.02	−.36**	−.47**
9. RP 5 min.	.23	.21									–	.63**	.44**	.70**	.64**	.37**	.26**	.18*	.25**	.21**
10. RP 1 week uni.	.33	.25										–	.41**	.63**	.53**	−.05	.63**	.08	.09	.06
11. RP 4 weeks uni.	.29	.25											–	.31**	.37**	.12	.17*	.71**	.06	.10
12. RP 1 week acc.	.49	.27												–	.83**	.15	.18*	.05	.53**	.40**
13. RP 4 weeks acc.	.43	.27													–	.14	.08	.13	.42**	.57**
14. TE 5 min	.17	.22														–	.07	.12	.47**	.42**
15. TE 1 week uni.	.20	.19															–	.02	.07	.00
16. TE 4 weeks uni.	.12	.21																	.04	.11
17. TE 1 week acc.	.16	.26																		.74**
18. TE 4 weeks acc.	.23	.27																		–

Note. *$p < .05$; **$p < .01$; RP = retrieval practice; TE = testing effect (retrieval practice – re-study); uni. = unique word pairs; acc. = accumulated word pairs; WMC = working memory capacity; NFC = Need for Cognition.

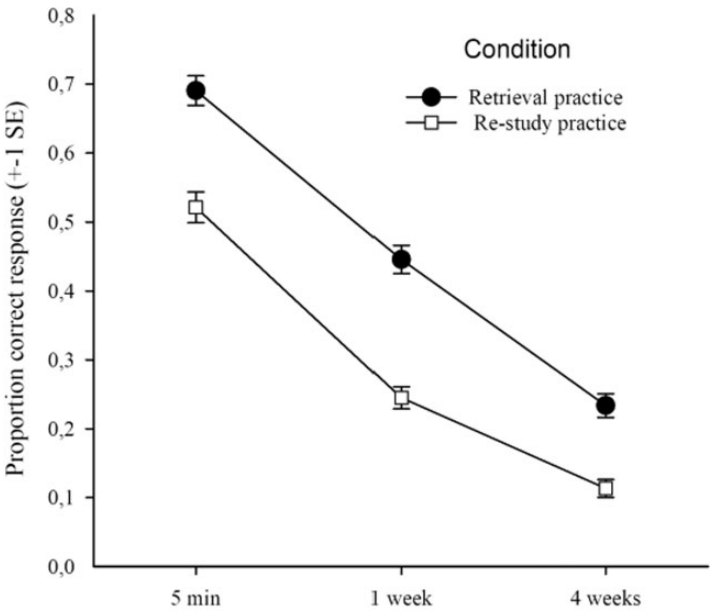


Figure 2. Recall Performance for Unique Word Pairs
Note. Proportion of correctly recalled unique word pairs as a function of practice (retrieval vs re-study) and retention interval (5 min, 1 week, vs 4 weeks).

The interaction was driven by a decreased difference between retrieval practice and re-study word pairs from the one-week to the four-week test, potentially caused by recall of re-study word pairs approaching the floor (see Figure 2).

To determine whether there was a significant testing effect—that is, that retrieval practice led to better retention than re-studying at each retention interval—the pairwise comparisons (Bonferroni corrected) for the main effects were inspected. They revealed that the significant main effect of retention interval reflects significant differences between all levels of the variable (all ps $< .001$) and that the significant main effect of practice reflects a significant difference in favor of the retrieval-practice condition ($p < .001$, see Figure 2).

Next, the effect of WMC, Grit, and NFC on the testing effect was investigated using hierarchical regression analyses where WMC was entered in the first step, and Grit and NFC were added in the second step. The three regression analyses were Bonferroni corrected for multiple analyses ($p < .017$). As mentioned in Bertilsson et al. (2017), the reason for setting up the analyses this way was to examine the effects of WMC separately, and then to control for WMC when analyzing the effects of Grit and NFC. The difference in retention between retrieval practice and re-study practice on the three retention tests were the dependent variables (i.e., the testing effect). This setup is in contrast to Bertilsson et al. (2017) where separate regression analyses were conducted using performance in the retrieval-practice and re-study-practice conditions at each retention interval. The regression analyses for unique word pairs in the present study revealed no significant relations between WMC, Grit, or NFC and the testing effect (Table 2).

To investigate whether the results would be confounded by retrieval-practice effects that arise from taking the tests after 5 minutes and 1 week (i.e., accumulated word pairs), a

Table 2. Hierarchical Regression Analyses using WMC, Grit, and NFC as Predictors of the Testing Effect (i.e., the Difference in Retention Between Retrieval Practice and Re-study Word Pairs) for Word Pairs Unique to Each Retention Test.

Delay	Predictors	β	t	p	r^2	F(total model)
5 min	Step 1				.00	$F(1,149) = 0.72$, $p = .40$
	WMC	.07	0.85	.40		
	Step 2				.01	$F(3,147) = 0.37$, $p = .77$
	WMC	.09	0.99	.33		
	Grit	−.00	−0.03	.98		
	NFC	−.05	−0.60	.55		
1 week	Step 1				.03	$F(1,149) = 4.54$, $p = .04$
	WMC	.17	2.13	.04		
	Step 2				.05	$F(3,147) = 2.75$, $p = .05$
	WMC	.13	1.46	.14		
	Grit	.02	0.23	.82		
	NFC	.15	1.74	.08		
4 weeks	Step 1				.01	$F(1,149) = 1.39$, $p = .24$
	WMC	−.10	−1.18	.24		
	Step 2				.02	$F(3,147) = 0.76$, $p = .52$
	WMC	−.09	−1.00	.32		
	Grit	.08	0.94	.35		
	NFC	−.02	−0.25	.81		

Note. WMC = working memory capacity; NFC = Need for Cognition.

second factorial 2×3 repeated-measures ANOVA was conducted using the accumulated word pairs tested at each retention interval (5 minutes = 20 items; 1 week = 40 items; 4 weeks = 60 items). The results showed main effects of retention interval, $F(1.6, 242) = 232.906$, $p < .001$, $\eta_p^2 = .61$, as well as practice condition, $F(1, 150) = 83.264$, $p < .001$, $\eta_p^2 = .36$ (see Figure 3), but no interaction effect. The lack of interaction between retention interval and practice condition suggests that retention of both types of word pairs declined between each of the three consecutive retention tests (see Figure 3).

Identical hierarchical regression analyses as for the unique word pairs were conducted, but now including word pairs that had been tested in previous retention tests (i.e., accumulated; Bonferroni corrected for multiple analyses, $p < .017$). The results revealed again that none of the independent variables were related to the difference in performance on any of the retention tests (Table 3).

Because of the non-significant effects of NFC, Grit, and WMC, a post-hoc power analysis was conducted (Onwuegbuzie & Leech, 2004) using G*Power 3.1.9.7 and the statistical test of multiple linear regression (Faul et al., 2009). The correlations between NFC, Grit, and WMC was used as input and provided an effect size of $f^2 = 0.21$. With an alpha level set to .05 and the sample size of 151 the post-hoc power analysis provided a power of .99, indicating a small risk of a Type-II error.

Discussion

This study partially replicated the design and procedure previously used by Bertilsson et al. (2017) using a larger sample size with the purpose of adding additional insight to how the

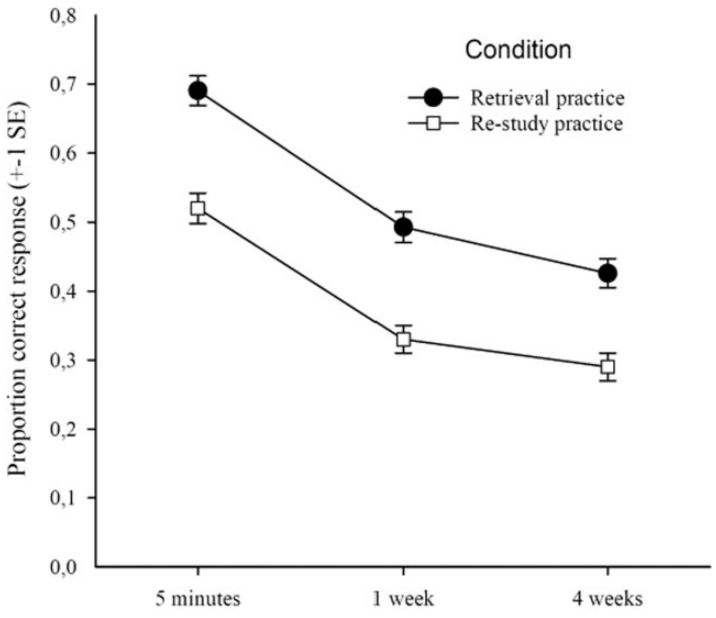


Figure 3. Recall Performance for Accumulated Word Pairs
Note. Proportion of correctly recalled accumulated word pairs as a function of practice (retrieval vs re-study) and retention interval (5 min, 1 week, vs 4 weeks).

Table 3. Hierarchical Regression Analyses Using WMC, Grit, and NFC as Predictors of the Testing Effect (i.e., the Difference in Retention Between Retrieval Practice and Re-study Word Pairs) for Accumulated Word Pairs.

Delay	Predictors	β	t	p	r^2	F (total model)
5 min	Step 1				.00	F (1, 149) = 0.72, p = .40
	WMC	.07	0.85	.40		
	Step 2				.01	F (3, 147) = 0.37, p = .77
	WMC	.09	0.99	.33		
	Grit	−.00	−0.03	.98		
	NFC	−.05	−0.60	.55		
1 week	Step 1				.01	F (1, 149) = 2.02, p = .16
	WMC	.12	1.42	.16		
	Step 2				.03	F (3, 147) = 1.57, p = .20
	WMC	.11	1.25	.21		
	Grit	−.13	−1.58	.12		
	NFC	.00	0.05	.96		
4 weeks	Step 1				.00	F (1, 149) = 0.33, p = .57
	WMC	.05	0.58	.57		
	Step 2				.01	F (3,147) = 0.28, p = .84
	WMC	.04	0.42	.67		
	Grit	−.06	−0.71	.48		
	NFC	.03	0.29	.77		

Note. WMC = working memory capacity; NFC = Need for Cognition.

benefits of retrieval practice are associated with interindividual differences in WMC and personality traits. Furthermore, this study contributes new insights compared to Bertilsson et al. (2017) by examining the testing effect using difference scores (retrieval practice – re-study practice).

As in Bertilsson et al. (2017), the ANOVA conducted using unique word pairs showed that retention of retrieval practice word pairs was significantly better than the retention of re-study word pairs on all three retention tests—that is, testing effects. However, in contrast to the results in Bertilsson et al. (2017), an interaction effect between retention interval and practice condition was found, meaning that the decline in performance between the retention tests differed between the learning techniques. When it comes to the ANOVA conducted using accumulated word pairs, within-subjects testing effects were obtained, and, in line with the previous results, no interaction effect was found between retention interval and practice condition. This indicates that for unique word pairs there is a difference in the decline in performance between the two learning techniques, while for accumulated word pairs the decline is similar irrespective of learning technique. However, as can be seen in Figure 2, a floor effect can potentially explain the interaction rather than a real decreased difference between retrieval practice and re-study word pairs at the four-week test (but for related findings, see Carpenter et al., 2008). The better retention of accumulated word pairs, compared to the unique word pairs, across time illustrates that in a pedagogical setting it is essential as a learner to have the possibility to retrieve the to-be-learned material several times in order to better consolidate the information (e.g., Rawson & Dunlosky, 2011; Wiklund-Hörnqvist et al., 2020).

Bertilsson et al. (2017) did not find any relationships between NFC, Grit, or WMC and performance on word pairs, in either retrieval-practice or re-study-practice conditions. However, the analyses were conducted separately for retrieval practice and re-study word pairs, using performance at each retention interval as the dependent variables, which could be interpreted as reflecting episodic memory performance rather than the testing effect. In the present study, the regression analyses were performed using the performance difference score between the two conditions at each retention interval. The analyses were again conducted on both unique and accumulated word pairs, and, in line with the findings from Bertilsson et al. (2017), no relationships were found between any of the predictors and performance on any of the retention tests (albeit corrected for multiple comparisons).

The non-significant relationships between Grit, NFC, and retrieval practice performance found in both the current study and Bertilsson et al. (2017), with respect to both unique and accumulated word pairs, suggest that individual differences in personality traits emphasizing effort seem to be unrelated to the effects of retrieval practice. It therefore seems plausible that the testing effects in the present study and in Bertilsson et al. (2017), to a large extent, were driven by TAP—an argument that is in line with Agarwal's (2019) findings that TAP is the critical process in retrieval practice. However, an important caveat is, of course, that we did not directly manipulate practice and retention test format. Further, since it was found that the measure of Grit had a lower internal consistency in the present study than it has previously been shown to have, we have to be cautious in drawing conclusions about the relationship between Grit and the testing effect.

With that in mind, one aspect to consider is whether students who enjoy thinking (NFC) or show Grit found the experimental setup and the vocabulary language materials to be cognitively stimulating or motivational. Perhaps grittiness and/or NFC would have a significant impact on performance with more complex material, if the tasks had been subject

for grading, or if the participants were allowed to choose whether they wanted to use retrieval practice or not. Moreover, it is also possible that the use of a more comprehensive measure of personality (e.g., the 48-item Conscientiousness-scale of the Revised NEO Personality Inventory [Costa & McCrae 2008]) would have yielded a different result. Such aspects should be investigated in future studies. Thus, it cannot be ruled out that for Grit, the small number of items included in the instrument and possible differences in the two subconstructs is a validity problem.

The non-significant relation of WMC and the testing effect is in line with previous studies (Brewer & Unsworth, 2012; Minear et al., 2018; Wiklund-Hörnqvist et al., 2014) and further underscore the conclusion that cognitive abilities (at least WMC) are of less importance for the use of retrieval practice. However, it should be noted that both Brewer and Unsworth (2012) and Minear et al. (2018) included measurements of intelligence (gf) beyond examining the non-significant effects of WMC. While Brewer and Unsworth (2012) found that retrieval practice was most beneficial for those with lower gf (relative higher gf), Minear et al. (2018) found that students with lower (compared to higher) gf showed a larger testing effect for easy items. The opposite pattern was found for difficult items, such that students with higher (compared to lower) gf showed a larger testing effect. However, no significant relationship between gf and the overall testing effect was evident (Minear et al., 2018). In line with the current study, both studies used foreign language vocabulary as the to-be-learned material, but in contrast to the current study the testing effect was examined after one day (Brewer & Unsworth, 2012) or after two days (Minear et al., 2018), while the current study spanned across weeks (see also Wiklund-Hörnqvist et al., 2014). In addition, the current study included both accumulated and unique word pairs, and, as evident from the results, the level of performance differed between accumulated and unique word pairs such that accumulated word pairs were retained at a higher degree relative to unique word pairs, possibly also due to testing. Importantly, independent of accumulated or unique word pairs, non-significant effects of WMC were evident across all three retention intervals, suggesting that cognitive load associated with retrieving unique word pairs relative to accumulated word pairs was comparable despite differences in performance level.

In sum, the results from Bertilsson et al. (2017) and the present study indicate that retrieval practice is a useful learning strategy in the context of acquiring a foreign language vocabulary. The present study also highlights retrieval practice as an effective learning strategy useful for students irrespective of the cognitive prerequisites and personality characteristics targeted. Such scientific evidence further emphasizes the significance of merging psychological and didactical knowledge for the purpose to optimize learning outcomes in the classroom. Together with previous studies, we argue that retrieval practice should be explained and taught to students and teachers both as a pedagogical tool and as an individual learning strategy.

The present study contains some limitations regarding the conclusions that can be drawn from the results. While a testing effect was identified, it is not possible to discern whether it is the result of a direct effect of testing, an indirect effect of testing (e.g., test-potentiated learning or forward testing effect), or whether both direct and indirect effects contribute to the advantage of retrieval practice. Arnold and McDermott (2013) suggest that the observed benefit of feedback on the testing effect may actually be the result of the testing effect and test-potentiated learning in conjunction. While the majority of the literature regarding the testing effect largely ignores this ambiguity, some studies have made attempts at separating the effects (e.g., Arnold & McDermott, 2013; Kubik et al., 2016).

Future research aiming to investigate the effectiveness of retrieval practice should differentiate between direct and indirect effects of testing to enable interpreting individual differences in regard to different types of the testing effect. Although we know from a wealth of studies that retrieval practice produces superior long-term retention even in the absence of feedback (e.g., Roediger & Butler, 2011), and independent of being a teacher or student, acquiring long-lasting learning is one of the challenges within the educational system. From an educational perspective, the results in the current study indicate that retrieval practice accompanied by feedback can be one way for educators to respond to individual variability in terms of personality traits and cognitive abilities well associated with learning.

Acknowledgement

We thank Tony Quillbard for conducting the computer programming required for this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding was received from Umeå School of Education to Bert Jonsson and Frida Bertilsson (project memory and learning) and the Swedish Research Council (grant number 721-2014-2099) to Bert Jonsson.

ORCID iD

Frida Bertilsson https://orcid.org/0000-0002-0709-3647

References

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

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (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

Alloway, T. P. (2006). How does working memory work in the classroom? *Educational Research and Reviews*, *1*(4), 134–139.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Arbabi, T., Vollmer, C., Dörfler, T., & Randler, C. (2015). The influence of chronotype and intelligence on academic achievement in primary school is mediated by conscientiousness, midpoint of sleep and motivation. *Chronobiology International*, *32*(3), 349–357. https://doi.org/10.3109/07420528.2014.980508

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), 136–140. https://doi.org/10.1016/j.cub.2009.12.014

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *Journal of Cognitive Education and Psychology*, *16*(3), 241–259. https://doi.org/10.1891/1945-8959.16.3.241

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing*. MIT Press. https://doi.org/10.1016/S0013-4694(97)84006-4

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

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition*, *36*(2), 438–448. https://doi.org/10.3758/MC.36.2.438

Cazan, A.-M., & Indreica, S. E. (2014). Need for cognition and approaches to learning among university students. *Procedia—Social and Behavioral Sciences*, *127*, 134–138. https://doi.org/10.1016/j.sbspro.2014.03.227

Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences*, *44*(7), 1596–1603. https://doi.org/10.1016/j.paid.2008.01.003

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

Costa, P. T., Jr., & McCrae, R. R. (2008). *The Revised NEO Personality Inventory (NEO-PI-R)*. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (pp. 179–198). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *In Educational Psychology Review*, *26(*2): 197–223. https://doi.org/10.1007/s10648-013-9246-y

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492–511. https://doi.org/10.1037/pspp0000102

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Dornic, S., Ekehammar, B., & Laaksonen, T. (1991). Tolerance for mental effort: Self-ratings related to perception, performance and personality. *Personality and Individual Differences*, *12*(3), 313–319. https://doi.org/10.1016/0191-8869(91)90118-U

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

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

Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences*, *14*(1), 47–64. https://doi.org/10.1016/j.lindif.2003.08.002

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karlsson Wirebring, L., Wiklund-Hornqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015). Lesser neural pattern similarity across repeated tests is associated with better long-term memory retention. *Journal of Neuroscience*, *35*(26), 9595–9602. https://doi.org/10.1523/JNEUROSCI.3550-14.2015

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-Based Learning: An Episodic Context Account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Kubik, V., Olofsson, J. K., Nilsson, L. G., & Jönsson, F. U. (2016). Putting action memory to the test: Testing affects subsequent restudy but not long-Term forgetting of action events. *Journal of Cognitive Psychology*, *28*(2), 209–219. https://doi.org/10.1080/20445911.2015.1111378

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lounsbury, J. W., Sundstrom, E., Loveland, J. L., & Gibson, L. W. (2003). Broad versus narrow personality traits in predicting academic performance of adolescents. *Learning and Individual Differences*, *14*(1), 65–75. https://doi.org/10.1016/j.lindif.2003.08.001

Meier, E., Vogl, K., & Preckel, F. (2014). Motivational characteristics of students in gifted classes: The pivotal role of need for cognition. *Learning and Individual Differences*, *33*, 39–46. https://doi.org/10.1016/j.lindif.2014.04.006

Meriac, J. P., Slifka, J. S., & LaBat, L. R. (2015). Work ethic and grit: An examination of empirical redundancy. *Personality and Individual Differences*, *86*, 401–405. https://doi.org/10.1016/j.paid.2015.07.009

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335. https://doi.org/10.1080/09658219408258951

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971–990. https://doi.org/10.1016/j.paid.2007.03.017

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*(4), 201–230. https://doi.org/10.1207/s15328031us0304_1

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

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

Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science*, *330*(6002), 335–335. https://doi.org/10.1126/science.1191465

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

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

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

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1467-8721.2008.00612.x

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

Sadowski, C. J., & Gulgoz, S. (1992). Association of need for cognition and course performance. *Perceptual and Motor Skills*, *74*(2), 498–498. https://doi.org/10.2466/pms.1992.74.2.498

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Stenlund, T., & Jonsson, B. (2017). Assessing the willingness to elaborate among young students: Psychometric evaluation of a Swedish need for cognition scale. *Frontiers in Education*, *2*(2). https://doi.org/10.3389/feduc.2017.00002

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710–1727. https://doi.org/10.1080/01443410.2014.953037

Tse, C.-S., Chan, M. H.-M., Tse, W.-S., & Wong, S. W.-H. (2019). Can the testing effect for general knowledge facts be influenced by distraction due to divided attention or experimentally induced anxious mood? *Frontiers in Psychology*, *10*, 969. https://doi.org/10.3389/fpsyg.2019.00969

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

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory and Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

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

Van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*(1), 31–46. https://doi.org/10.1177/0146621606286206

Van Seggelen-Damen, I. C. M. (2013). Reflective personality: Identifying cognitive style and cognitive complexity. *Current Psychology*, *32*(1), 82–99. https://doi.org/10.1007/s12144-013-9166-5

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10–16. https://doi.org/10.1111/sjop.12093

Wiklund-Hörnqvist, C., Stillesjö, S., Andersson, M., Jonsson, B. & Nyberg, L. (2020). Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus. *Brain and Behavior* 00, e01909. https://doi.org/10.1002/brb3.1909

Author Biographies

Frida Bertilsson is a PhD student with a masters degree in cognitive science. She is part of a research project denoted as 'The learning brain' which focuses on studying learning strategies in relation to cognition.

Tova Stenlund is an associate professor in the Department of Psychology, Umeå University, Sweden. She has a PhD degree in educational measurement, and her research and teaching interests lie in the field of educational psychology and cognitive psychology, with a particular focus on validity of test and assessment results, test-taking behavior, and effects of repeated testing on memory and learning.

Carola Wiklund-Hörnqvist is an assistant professor in the Department of Psychology, Umeå University, Sweden. She has a PhD in psychology, and her research and teaching interests lie in the field of educational neuroscience, focusing on the relationship between memory and learning, specifically; on the effects of repeated testing on memory and learning related to both behavioral and neuroimaging data.

Bert Jonsson is a professor in the Department of Applied Educational Science. He is the principal investigator of the project 'The learning brain' financed by the Swedish Research Council. The project investigates fundamental questions arising in educational science and pertaining to the cognitive neuroscience of children's learning. The project focuses mainly on learning strategies in relation to cognition.