Journal of Memory and Language 66 (2012) 407–415

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

Individual differences in the effects of retrieval from long-term memory

Gene A. Brewer [a,*], Nash Unsworth [b]

[a] Department of Psychology, Arizona State University, Tempe, AZ 85287-1104, United States
[b] Department of Psychology, University of Oregon, United States

ARTICLE INFO

Article history:
Received 5 January 2011
revision received 14 December 2011
Available online 17 January 2012

Keywords:
Individual differences
Testing effect
Episodic memory
Intelligence
Attention control
Working memory

ABSTRACT

The current study examined individual differences in the effects of retrieval from long-term memory (i.e., the testing effect). The effects of retrieving from memory make tested information more accessible for future retrieval attempts. Despite the broad applied ramifications of such a potent memorization technique there is a paucity of research tailored toward scrutinizing variability in the effect. Multiple measures of working memory capacity, attention control, episodic memory, and general-fluid intelligence were collected in addition to performance in a standard paired-associate testing task. A testing effect was observed and there was a great deal of individual variability in the magnitude of the effect. This variability was best accounted for by memory and intelligence constructs. Furthermore, the pattern of results is consistent with the notion that students with poor memory abilities and lower general-fluid intelligence benefit more so from testing memory than high ability students.

Published by Elsevier Inc.

> "The relationship between test scores and school performance seems to be ubiquitous. Wherever it has been studied, children with high scores on tests of intelligence tend to learn more of what is taught in school than their lower scoring peers. There may be styles of teaching and methods of instruction that will decrease this correlation, but none that consistently eliminates it has yet been found..." – Neisser et al. (1996, p. 82)

Introduction

The act of retrieving information from memory reinforces that information thereby rendering it more accessible during later retrieval attempts (Abbott, 1909; Bjork, 1975; Izawa, 1967). The effects of retrieval on subsequent memory performance are collectively referred to as the testing effect. Recently, there has been a surge of empirical research applying cognitive principles toward understanding the testing effect (for a review see Roediger & Karpicke, 2006a). The principal reason for this recent interest is the potential benefits of implementing repeated testing in applied settings such as the classroom (McDaniel, Roediger, & McDermott, 2007). However, there is limited individual differences research investigating the testing effect. That is, the effects of retrieving information from memory may or may not extend to all people in a similar manner. Finding individual difference constructs that are related to the benefits of testing is an important endeavor for both theoretical and applied psychologists. For example, it is an outstanding question whether testing should be uniformly applied in the classroom, or whether testing helps certain subpopulations more than others. The primary goal of the current report was to provide one of the first individual differences analysis of the testing effect.

The testing effect

In standard testing-effect paradigms, a set of to-be-remembered material is encoded and subsequently retrieved. After the initial test, participants engage in some other activity or delay before having their memory for that same information probed again later in the future. Memory for initially tested information is more immune to forgetting and is also more accessible for future retrieval attempts (Roediger & Karpicke, 2006b). The benefits of testing extend

[*] Corresponding author.
E-mail address: gene.brewer@asu.edu (G.A. Brewer).

0749-596X/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jml.2011.12.009

beyond re-presentation and are generally apparent after some significant delay (Carrier & Pashler, 1992). Thus, retrieval makes memories for tested material more durable than restudied material even though more test-relevant information is processed through re-presentation than retrieval. Generally, the results from a multitude of studies implementing different testing procedures have supported the hypothesis that one causal mechanism underlying the testing effect is controlled and effortful retrieval from long-term memory (e.g., Carpenter, 2009). Given the reliability of the testing effect across a variety of paradigms (Roediger, Agarwal, Kang, & Marsh, 2010), the current study will focus specifically on cued-recall paradigms.

With respect to the testing effect, the cued-recall paradigm has been fruitful for exploring the cognitive processes underlying retrieval benefits (Carpenter, Pashler, & Vul, 2006; Izawa, 1967; Pyc & Rawson, 2009). Carpenter et al. (2006) hypothesized that the association between cue-target pairs would be strengthened by testing as compared with restudy. In the initial testing phase, participants were always given the cue (A) and they tried to retrieve the target (B). On the final test, participants were either given the original cue (A), or the original target (B), and they were asked to produce the other member of the pair. Across conditions, participants consistently produced the pair member with higher probability when it had been previously tested roughly 18–48 h earlier. This result extended prior theorizing in the testing literature by suggesting that retrieval strengthens associative information in a bidirectional manner (see also Rizzuto & Kahana, 2001).

Based on these types of results, the retrieval-effort hypothesis has been proposed by Carpenter and Delosh (2006). Essentially, they argued that the difficulty of retrieval processes and the degree of cue elaboration operating during an initial test are partly responsible for improved subsequent memory (see also Carpenter, 2009; Pyc & Rawson, 2009). As the current research will demonstrate, the notion of difficult, or elaborate, retrieval processes could be interpreted in any number of ways including the strategic use of working memory processes, controlled attention, or long-term memory search mechanisms. Thus, it is not clear whether the beneficial effect of difficult retrieval exerts itself through one or several cognitive control processes. Also, it is not clear how these processes interact to establish a memorial benefit to previously tested information. Individual differences methodology can help to clarify these issues as well as other important aspects of the testing effect.

Individual differences and the effects of testing

An individual differences approach can be a useful tool for elucidating the component processes underlying memory ability (Underwood, 1975). However, very little research has examined individual differences in the magnitude of the testing effect. Chan (2009) collected data from a single working memory task (operation span) but failed to find any relationships with the testing effect. This lack of a relationship could be due to task unreliability or due to idiosyncratic effects from using a single measure. Tse, Balota, and Roediger (2010) investigated the effect of aging on the testing effect. When compared with middle age adults (60–80 years old), they found that older adults (greater than 80 years old) had larger testing effects but only when feedback was provided during initial testing. Theoretically, Tse and colleagues argued that deficits in cognitive control processes such as associative binding, controlled attention, and memory monitoring might underlie age differences in the testing effect. To date, no published study has focused explicitly on examining individual differences in cognitive abilities and the direct effects of testing memory in a college aged sample. To the extent that individuals do differ in the effects of testing, it is necessary to determine what cognitive ability constructs may be accounting for these individual differences.

To find potential relations between testing and higher-order cognition, we selected four ability constructs to examine in relation to the testing effect. Prior research has shown that performance on working memory capacity (WMC) tasks is related to various higher-order cognitive abilities related to success in school environments including reading comprehension, standardized-achievement test scores, reasoning ability, and intelligence (see Engle and Kane (2004) for a review). As mentioned earlier, research has suggested that effortful retrieval leads to the largest testing effects. To the degree that attention control (AC) is necessary for effortful retrieval (Kane & Engle, 2000), performance on attention tasks may be related to the magnitude of the testing effect. Therefore, the strategic regulation of attention may be an important factor for understanding individual differences in the testing effect (Dudukovic, DuBrow, & Wagner, 2009). Given that memory retrieval is a basic component of testing effect paradigms, it seems most likely that tasks tapping episodic memory (EM) abilities will be closely related to the size of the testing effect at an individual differences level. Both AC and EM processes underlie the relation between WMC and general-fluid intelligence (gF; Unsworth & Spillers, 2010). As such, gF may share important variation with the magnitude of the testing effect to the degree that gF broadly represents fluid reasoning and domain-general cognitive control abilities. Importantly, there is reason to believe that some (or all) of these constructs may be related to the testing effect but it is also critical to examine their shared influences. While references have been made to these underlying cognitive mechanisms responsible for the testing effect, virtually no study to date has explicitly examined them contemporaneously in an individual differences analysis.

Research has suggested that retrieval is an important determinant of both remembering and forgetting but one critical question remains unanswered: How would performance on basic attention, memory, and intelligence tasks relate to the testing effect? Primarily, there are three hypotheses that can be derived from positing this question.

1. *Testing provides general benefits across the ability range.* That is, all students are benefited equally from testing and the only differences are due to baseline differences in cognitive abilities. This would serve to shift the entire distribution of scores upwards, but would not change the rank ordering of individuals. From an applied standpoint this result would suggest that testing can be applied equally to all students.

2. *Testing allows the rich to get richer.* That is, testing allows high ability students to better utilize their inherent abilities and boost their scores more so than low ability students. From an applied standpoint this result would suggest that testing preferentially helps good students, but does little to help low ability students. As such, other intervention techniques would be needed to better help low ability students.
3. *Testing homologizes memory across the ability range.* That is, high ability students are already maximally utilizing their cognitive abilities and thus, testing does not help them much (or at all). Low ability students, however, are benefited by testing because they normally do not utilize their cognitive abilities as well and testing encourages the usage of those processes. From an applied standpoint this result would suggest that testing could be applied uniformly in the classroom given that all individuals would likely be benefited, but low ability students would be benefited the most. Moreover, high ability students may benefit from additional strategies and interventions.

In these hypotheses, "ability" simply refers to performance on various working memory, attention, episodic memory, and intelligence tasks. Notably, there are logical reasons to suspect that each of these hypotheses are plausible, and thus an individual differences examination is desired to tease apart any ability × intervention interactions. Another way to consider these three hypotheses is within the aptitude × treatment framework described by Cronbach and Snow (1977). Hypotheses 2 and 3 reflect important aptitude × treatment interactions that are rare and extraordinarily informative for both theoretical and educational purposes. A broader discussion of the current results in relation to this framework will be provided in after the results have been presented.

The current study

The current study employed a large-scale individual differences approach with multiple measures of WMC, AC, EM, and gF. Composite scores were drawn out of these measures and were used to predict performance on a paired-associate testing task. Having multiple measures of each construct allowed us to compute composite scores with better psychometric properties. By exploring variation in the testing effect, the current study has nontrivial implications for memory researchers and applied psychologists, as well as teachers and professors working within any classroom setting. Understanding whom likely benefits from testing should allow us to better understand the nature of the testing effect itself as well as devise better practices that can be utilized in the classroom.

Method

Participants

University of Georgia students ($n = 107$) volunteered in exchange for course credit. Each participant completed a computerized battery of tasks that measured the testing effect in paired-associate learning, WMC, AC, EM, and gF. Each participant was tested in two sessions lasting approximately 2 h each.

Materials and procedure

After signing informed consent, all participants completed operation, symmetry, and reading span, cued recall, paired-associate testing (initial), and number series tasks in Session 1. In Session 2, all participants completed a picture-source, gender-source, raven, arrow flankers, psychomotor vigilance, delayed free recall, letter sets, paired-associate testing (final), and antisaccade tasks. Tasks were administered in the order listed above and the testing sessions were separated by a 24 h delay.

Tasks

Paired-associate testing task

The parameters of this task mapped onto those used by Carpenter et al. (2006). In this task participants encoded 40 word pairs for 6 s per pair. Subsequent to the study phase, participants restudied 20 of the cue-target pairs and then took a cued-recall test over the other 20 pairs. This order was chosen to avoid participants carrying over a testing strategy on the restudy pairs. Such a strategy would run the risk of diluting the testing effect. For the restudy pairs, participants were presented with the same pair for an additional 6 s. In this initial test participants were presented with the cue word and instructed to type the target word that it was originally paired with during the encoding phase. For the tested pairs, participants had 5 s to type the target word when presented with the cue. After 5 s elapsed, participants were given the correct target for an additional second. Therefore, even when participants could not recall the target in the initial test they were still presented with the correct target for a brief period of time (i.e., item presentation was roughly matched). In the second experimental session (separated by 24 h) participants were tested over all 40 cue-target pairs. The dependent variable was the difference in proportions of originally tested and restudied cue-target pairs correctly recalled on the final test.

Working memory tasks

Operation span (Ospan)

Participants solved a series of math operations while trying to remember a set of unrelated letters (for full task details see Unsworth, Heitz, Schrock, & Engle, 2005). Participants were required to solve a math operation and after solving the operation they were presented with a letter for 1 s. Immediately after the letter was presented the next operation was presented. At recall, letters from the current set were recalled in the correct order by clicking on the appropriate letters. For all of the span measures, items were scored if the item is correct and in the correct serial position. The dependent variable is the number of correct items recalled in the correct serial position.

Symmetry span (Symspan)

Participants were required to recall sequences of red squares within a matrix while performing a symmetry-judgment task (for full task details see Unsworth, Redick, Heitz, Broadway, & Engle, 2009). In the symmetry-judgment task participants were shown an 8 × 8 matrix with some squares filled in black. Participants decided whether the design was symmetrical about its vertical axis. The pattern was symmetrical half of the time. Immediately after determining whether the pattern was symmetrical, participants were presented with a 4 × 4 matrix with one of the cells filled in red for 650 ms. At recall, participants recalled the sequence of red-square locations in the preceding displays, in the order they appeared by clicking on the cells of an empty matrix. The same scoring procedure as Ospan was used.

Reading span (Rspan)

Participants were required to read sentences while trying to remember a set of unrelated letters (for full task details see Unsworth et al., 2009). Participants read a sentence and determined whether the sentence made sense or not. Half of the sentences made sense while the other half did not. Nonsense sentences were made by simply changing one word from an otherwise normal sentence. After participants gave their response they were presented with a letter for 1 s. At recall, letters from the current set were recalled in the correct order by clicking on the appropriate letters. The same scoring procedure as Ospan was used.

Attention control tasks

Antisaccade

In this task (Kane, Bleckley, Conway, & Engle, 2001) participants were instructed to stare at a fixation point which is onscreen for a variable amount of time (200–2200 ms). A flashing white " = " was then flashed either to the left or right of fixation (11.33° of visual angle) for 100 ms. This cue was followed by the target stimulus (a B, P, or R) onscreen for 100 ms. The target was followed by masking stimuli (an H for 50 ms and an 8 which remains onscreen until a response is given). The participants' task was to identify the target letter by pressing a key for B, P, or R (the keys 1, 2, or 3) as quickly and accurately as possible. In the prosaccade condition the flashing cue (=) and the target appeared in the same location. In the antisaccade condition the target appeared in the opposite location as the flashing cue. Participants received, in order, 10 practice trials to learn the response mapping, 15 trials of the prosaccade condition, and 60 trials of the antisaccade condition. The dependent variable was the proportion of errors on the antisaccade trials.

Arrow flankers

Participants were presented with a fixation point for 400 ms. This was followed by an arrow directly above the fixation point for 1700 ms. The participants' task was to indicate the direction the arrow was pointing (pressing the F for left pointing arrows or pressing J for right pointing arrows) as quickly and accurately as possible. On 50 neutral trials the arrow was flanked by two horizontal lines on each side. On 50 congruent trials the arrow was flanked by two arrows pointing in the same direction as the target arrow on each side. Finally, on 50 incongruent trials the target arrow was flanked by two arrows pointing in the opposite direction as the target arrow on each side. All trial types were randomly intermixed. The dependent variable was the reaction time difference between incongruent and congruent trials.

Psychomotor vigilance task

Participants were presented with a row of zeros on screen and after a variable amount of time the zeros began to count up in 1 ms intervals from 0 ms. The participants' task was to press the spacebar as quickly as possible once the numbers started counting up. After pressing the spacebar the RT was left on screen for 1 s to provide feedback to the participants. Interstimulus intervals were randomly distributed and ranged from 1 to 10 s. The entire task lasted for 10 min for each individual (roughly 75 total trials). The dependent variable is the average reaction time from the slowest 20% of trials.

Episodic memory tasks

Delayed free recall unrelated words

Participants attempted to recall 6 lists of 10 words each. All words were common nouns that were presented for 1 s each. After list presentation, participants had a distractor task for 16 s in which a three-digit number appeared for 2 s and then they wrote the digits in ascending order on a separate piece of paper. After the distractor task participants typed as many words as they could remember from the current list in any order they wished. Participants had 45 s for recall. A participant's score was the total number of items recalled correctly.

Cued recall

In this task, participants were given three lists of 10 words pairs each. All words were common nouns and the word pairs were presented vertically for 2 s each. Participants were told that the cue would always be the word on top and the target would be on bottom. After the presentation of the last word, participants saw the cue word and ??? in place of the target word. Participants were instructed to type in the target word from the current list that matched cue and then to press ENTER to indicate their response. The cues were randomly mixed so that the corresponding target words were not recalled in the same order as they were presented. Participants had 5 s to type in the corresponding word. This same procedure was done for all three lists. A participant's score was the proportion of items recalled correctly.

Picture source-recognition

Participants were presented with a picture (30 total pictures) in one of four different quadrants onscreen for 1 s. Participants were explicitly instructed at encoding to pay attention to both the picture as well as the quadrant it was located in. At test, participants were presented with 30 old and 30 new pictures in the center of the screen. Participants indicated if the picture was new or old and, if old, what quadrant it was originally presented in via key press. Participants had 5 s to press the appropriate key to enter their response. A participant's score was the proportion correct.

Gender source recognition

Participants heard words (30 total words) in either a male or a female voice. Participants were explicitly instructed to pay attention to both the word as well as the voice the word was spoken in. At test participants were presented with 30 old and 30 new words and were required to indicate if the word was new or old and, if old, what voice it was spoken in via key press. Participants had 5 s to press the appropriate key to enter their response. A participant's score was the proportion of correct responses.

Intelligence tasks

Raven advanced progressive matrices

The Raven consisted of 18 items presented in escalating degree of difficulty. Each item consisted of a display of 3 × 3 matrices of geometric patterns with the bottom right pattern missing. The task required participants to select, among eight alternatives, the one that correctly completed the overall series of patterns. Participants had 10 min to complete the 18 odd-numbered items. A participant's score was the total number of correct solutions.

Number series

In this task, participants saw a series of numbers and determined what the next number in the sequence should be (Thurstone & Thurstone, 1962). That is, the series followed some unstated rule which participants were required to figure out in order to determine which the next number in the series should be. Participants selected their answer out of five possible numbers that were presented. Following five practice items, participants had 4.5 min to complete 15 test items. A participant's score was the total number of items solved correctly.

Letter sets

On each problem, participants saw five sets of letters containing four letters each. Participants were instructed to find the rule that applied to four of the five letter sets, and then indicate the letter set that violated the rule. Participants had 5 min to complete 20 items, with their total correct used as the dependent variable.

Results

Descriptive statistics, reliabilities, and correlations for all measures can be found in Tables 1 and 2, respectively. In line with previous research, *z*-score composites were made for the WMC, AC, EM, and gF constructs. Replicating much previous research, these *z*-scores were interrelated and differentially related to the magnitude of the testing effect (Table 3).

Table 1
Descriptive statistics and reliabilities for all of the measures.

Measure	Mean	SD	Skew	Kurtosis	Reliability
Testing task					
PAT testing	0.51	0.22	0.20	−0.73	0.84
PAT nontesting	0.45	0.26	0.43	−0.78	0.89
PAT difference	0.06	0.13	−0.29	−0.27	0.50
Working memory					
Ospan	60.51	12.49	−1.56	2.87	0.85
Sspan	30.12	7.66	−0.51	−0.46	0.76
Rspan	59.98	11.99	−1.33	1.82	0.87
Attention control					
Anti	0.50	0.14	−0.28	−0.31	0.65
Vigilance	504.94	152.24	2.15	5.85	0.72
Flanker	109.54	60.23	1.62	3.47	0.90
Episodic memory					
Gsource	0.62	0.12	−0.15	0.17	0.66
Psource	0.81	0.11	−1.37	3.01	0.78
CR	0.49	0.23	0.07	−0.87	0.87
DFR	0.54	0.19	0.08	−1.90	0.84
Intelligence					
Nseries	0.72	0.15	−0.97	2.17	0.60
Lseries	0.69	0.17	−1.25	1.12	0.76
Raven	0.59	0.13	−0.52	0.52	0.64

Note: PAT Testing = paired-associate testing condition; PAT Nontesting = paired-associate nontesting condition; PAT Diff = Difference Score; Ospan = operation span; Sym = symmetry span; Rspan = reading span; Anti = antisaccade; Vigilance = psychomotor vigilance; Flanker = arrow flanker; Gsource = gender source; Psource = picture source; CR = cued recall; DFR = delayed free recall; Nseries = number series; Lseries = letter series; Raven = Raven advanced progressive matrices.

The testing effect

Participants recalled 47% (SE = .02) of the 20 targets in the initial testing phase. Moreover, the testing effect in the paired-associate testing task was replicated in the current study, $F(1,106) = 25.16$, $p < .001$, $\eta_p^2 = .19$. That is, more targets were successfully recalled if they had previously been tested ($M = 51\%$) as compared with restudied ($M = 44\%$). Notably, this effect is considerably smaller than that reported by Carpenter et al. (2006). Additionally, there was a great deal of variability in participants' susceptibility to the effects of testing with 67% showing a positive effect, 12% showing no effect, and 21% showing a negative effect.

Relations with external measures

The broad correlations amongst these measures are found in Table 2 and the correlations amongst the *z*-composites are found in Table 3. To investigate the relationships between the paired-associate testing effect and external measures of WMC, AC, EM, and gF a difference score between the tested and restudied pairs was simultaneously regressed on the composite *z*-scores. This approach was chosen to evaluate any independent relations between the cognitive ability measures and the magnitude of the testing effect. The results from the simultaneous regression analysis can be found in Table 4. Neither the WMC nor AC constructs significantly predicted the magnitude of the testing effect (i.e., the difference score). There was, however, a significant relation between the EM com-

Table 2
Correlations for all of the measures.

	PATTest	PATNon	PATDiff	Ospan	Svm	Rspan	Anti	Vigilance	Flanker	Gsource	Psource	CR	DFR	Nseries	Lseries	Raven
PATTest	1.00															
PATNon	0.86	1.00														
PATDiff	−0.01	−0.50	1.00													
Ospan	0.24	0.26	−0.10	1.00												
Sym	0.14	0.23	−0.22	0.38	1.00											
Rspan	0.41	0.34	0.04	0.6S	0.33	1.00										
Anti	−0.10	−0.11	0.04	−0.07	−0.21	0.00	1.00									
Vigilance	−0.27	−0.28	0.09	−0.16	−0.15	−0.09	0.39	1.00								
Flanker	−0.12	−0.13	0.06	−0.22	−0.13	−0.18	0.13	0.10	1.00							
Gsource	0.23	0.27	−0.15	0.15	0.19	0.12	−0.20	−0.24	−0.08	1.00						
Psource	0.4:	0.4S	−0.23	0.34	0.21	0.47	−0.08	−0.11	−0.22	0.33	1.00					
CR	0.51	0.51	−0.14	0.17	0.34	0.26	−0.05	−0.04	−0.04	0.33	0.27	1.00				
DFR	0.45	0.54	−0.23	0.40	0.34	0.37	−0.12	−0.35	−0.13	0.24	0.30	0.58	1.00			
Nseries	0.15	0.23	−0.19	0.18	0.27	0.09	−0.12	−0.15	−0.10	0.17	0.27	0.11	0.13	1.00		
Lseries	0.23	0.35	−0.31	0.13	0.26	0.05	−0.19	−0.46	−0.18	0.1S	0.20	0.13	0.28	0.46	1.00	
Raven	0.19	0.26	−0.18	0.10	0.35	0.07	−0.21	−0.37	−0.18	0.29	0.17	0.16	0.19	0.41	0.52	1.00

Note: PATTest = paired-associate testing condition; PATNon = paired-associate nontesting condition; PATDiff = Difference Score; Ospan = operation span; Sym = symmetry span; Rspan = reading span; Anti = antisaccade; Vigilance = psychomotor vigilance; Flanker = arrow flanker; Gsource = gender source; Psource = picture source; CR = cued recall; DFR = delayed free recall; Nseries = number series; Lseries = letter series; Raven = Raven advanced progressive matrices.

Table 3
Correlations for the composite scores.

	PAT Diff	WMC	AC	EM	gF
PAT diff.	1.00				
WMC	−0.12	1.00			
AC	0.09	−0.23	1.00		
EM	−0.29	0.49	−0.26	1.00	
gF	−0.28	0.26	−0.36	0.33	1.00

Note: WMC = *z*-composite for three complex-span tasks; AC = *z*-composite for three attention-control tasks; EM = *z*-composite for four episodic memory tasks; gF = *z*-composite for three intelligence tasks.

posite and the magnitude of the testing effect (see Table 4). The slope coefficient is negative indicating that higher EM scores were associated with smaller testing effects. Additionally, there was a significant negative relation between the magnitude of the testing effect and gF indicating that higher ability students benefited less from testing.[1] To reiterate, above and beyond the correlations amongst the *z*-composite measures, EM and gF had independent contributions to predicting the magnitude of the testing effect.

Careful examination of Fig. 1a shows that individuals with negative EM *z*-scores typically exhibited bigger testing effects than participants with positive EM *z*-scores. To further investigate this effect, participants falling in the upper and lower quartiles of the distribution of EM scores were selected. As can be seen in Fig. 1b, participants with impoverished EM performance exhibited significantly larger testing effects than participants with high performance, $F(1,52) = 6.01, p < .05, \eta_p^2 = .11$. Follow up *t*-tests confirmed that participants falling in the lower quartile of the EM scores exhibited a significant testing effect whereas participants in the upper quartile did not exhibit a significant effect, $t(27) = 5.41$, $p < .01$ and $t(27) = 1.61$, *ns*.[2] Similar effects were found when examining the relation between gF and the paired-associate testing effect (see Fig. 2a and b). Specifically, participants in the lower quartile of gF scores showed the largest testing effects, $F(1,52) = 6.54$, $p < .05$, $\eta_p^2 = .11$. Follow up *t*-tests confirmed that participants with lower gF scores exhibited a significant testing effect whereas participants with higher scores did not exhibit a significant effect, $t(27) = 4.97$, $p < .01$ and $t(27) = 1.39$, *ns*. Thus, of the four external cognitive correlates examined in the current study, only EM and gF correlated with the paired-associate testing effect.[3] Furthermore, these relations indicated that low ability students benefit more from paired-associate testing than high ability students.

Discussion

The results from this study speak to the nature of individual differences in the testing effect in several important ways. First, the testing effect reported by Carpenter and

[1] Critically, to examine individual differences in the testing effect one must explore difference scores between initially tested and restudied pairs. The reliability of the difference score sets the upper bound on the correlation that measure can have with other measures. Table 1 shows that the reliability of the difference score computed by Lord's (1963) method was .50. Despite the poor reliability of the difference score, we still managed to find relations with other cognitive constructs (EM and gF). Moreover, the correlations were more impressive between the difference score and EM ($r = -.49$) and gF ($r = -.46$) after correcting for unreliability of the measures.

[2] A similar effect was found when performance on the criterion task was investigated. Participants who recalled less on the paired-associate testing task on day two generally exhibited the largest testing effects ($r = -.28$).

[3] The astute reader will notice in Figs. 1 and 2 that there was a wide range of scores on the paired-associate testing task suggesting that scaling issues may be at play. To mitigate this concern we created a new dependent measure by dividing the difference between nontested and tested averages by average memory for the nontested pairs. The rationale behind this transformation was to reduce the impact of baseline cued recall differences. This transformed measure of the testing effect was regressed on the same four constructs as before (WMC, AC, EM, & gF) with no change in the qualitative pattern of results. Both EM ($\beta = .33$, $p < .05$) and gF ($\beta = .17$, $p < .10$) predicted the magnitude of the testing effect even when accounting for baseline differences in the nontested pairs.

Table 4
Summary of the simultaneous regression analysis predicting the magnitude of the testing effect from working memory, attention control, episodic memory, and general-fluid intelligence constructs ($n = 107$).

Predicted: nontesting mean – testing mean		B	SE(B)	β	t	P
Working Memory Capacity	WMC	0.01	0.02	0.05	0.47	0.64
Attention Control	AC	−0.01	0.02	−0.05	−0.50	0.62
Episodic Memory	EM	**−0.05**	**0.02**	**−0.25**	**−2.27**	**0.03**
General-Fluid Intelligence	gF	**−0.04**	**0.02**	**−0.23**	**−2.22**	**0.03**

Note: $R^2 = .13$, $F(4,102) = 3.66$, $p < .05$, bolded values are significant at $p < .05$ level.

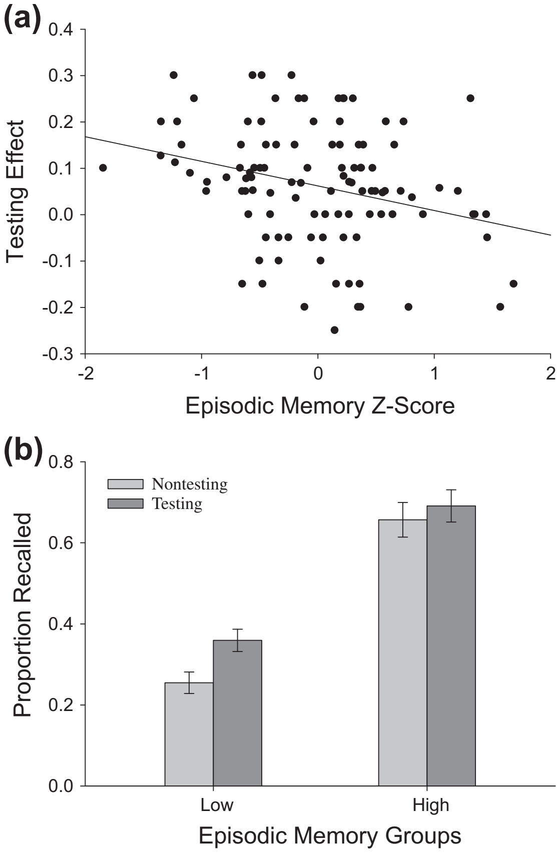

Fig. 1. A scatterplot (a) and bar graph (b) showing the relation between episodic memory abilities and the magnitude of the testing effect.

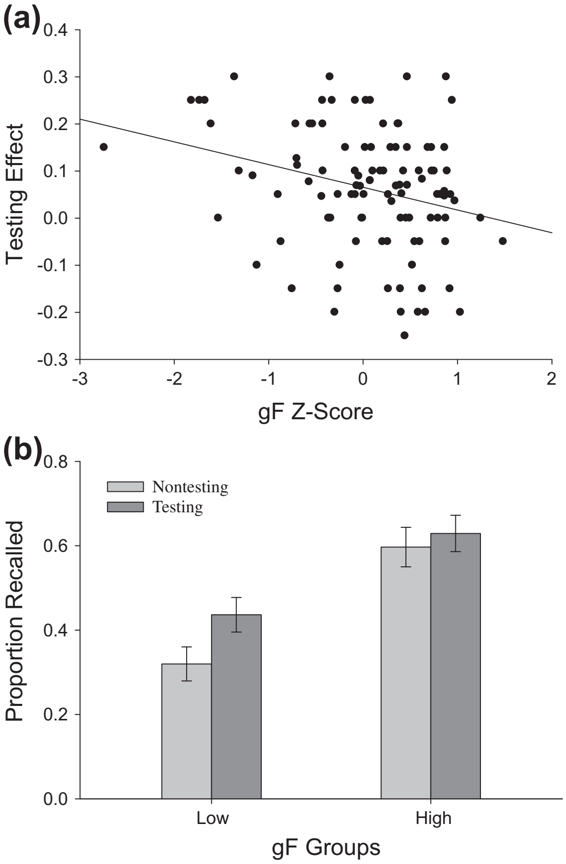

Fig. 2. A scatterplot (a) and bar graph (b) showing the relation between general-fluid intelligence and the magnitude of the testing effect.

colleagues (2006) was replicated. Second, there was variability in the magnitude of the testing effect with some participants demonstrating large effects, some demonstrating no effects, and others demonstrating negative effects. Thus, despite the robustness of the testing effect in the literature, it appears that not all participants get a memorial benefit from retrieving information from memory. Third, external measures of WMC, AC, EM, and gF were interrelated (replicating previous research; Unsworth & Spillers, 2010), related to performance on the criterion paired-associate testing task, but only the EM and gF constructs were related to the magnitude of the testing effect. Collectively, the results from the current study inform extant research on the beneficial effects of retrieval from long-term memory.

In the current study, neither the WMC nor AC constructs were reliably related to the magnitude of the testing effect (see Chan, 2009 for similar findings with a single measure of WMC). These results stand in opposition to previously reported results on the relation between WMC and the effects of retrieval from long-term memory. Brewer, Unsworth, and Spillers (in preparation) found that students with lower WMC can benefit from free recall testing to reduce proactive interference that accrues across multiple study-test trials on semantically related word lists. Taken together, the results from the current study and Brewer et al. (in preparation) highlight the notion that WMC is related to the testing effect only under certain conditions. As described earlier, Roediger and Karpicke

(2006a) proposed that testing leads to better memory performance by several different means. In Brewer et al. (in preparation), low WMC students used testing to sharpen their cue-driven retrieval strategies (i.e., an indirect effect). The current results demonstrate that direct effects of testing, at least in the paired-associate testing paradigm, were not related to WMC or AC. Effortful retrieval may be an important component of the testing effect, but variation in the control processes necessary for performing WMC and AC tasks does not influence the magnitude of the effect above and beyond either EM or gF. Thus, it remains an open question as to when external measures of WMC or AC will be related to the testing effect.

Long-term memory ability was, however, related to the magnitude of the paired-associate testing effect. This relationship was driven by low ability students gaining larger benefits from testing than high ability students. This result is consistent with the previous finding that poor performers on the criterion task (paired-associate testing) showed the biggest benefits from testing. This difference could arise for a number of reasons. For instance, participants with more efficient EM processes may not benefit as much from testing because they elaborately encode information leading to multiple retrieval routes whereas participants with poor EM ability need intermediate retrieval to build these routes (Bjork, 1975; Carpenter, 2009). Another possibility is that participants who were in the lower range of EM performance may have been forced to use more efficient retrieval strategies during initial testing. These retrieval strategies may have benefited the retrieval of the tested pairs during the final criterion test. Thus, future work is needed to tease apart these, and other, competing explanations for the current results. Nevertheless, students with poor EM abilities benefit more from testing on material in paired-associate tasks and this finding is clearly of great importance for applied and educational psychologists.

Perhaps the most compelling finding reported in the present work was that measures of higher-order intelligence were related to the magnitude of the paired-associate testing effect when controlling for variability in related measures of WMC, AC, and EM. Currently, it is not yet known why gF is correlated with the testing effect but there are several potential hypotheses to be tested in future research. Perhaps this correlation is driven by a general *g*-factor that extends across a variety of mental abilities including the testing effect (Jensen, 1998). Alternatively, there may exist some component tapped by the common variance amongst the tasks that is responsible for the relationship (e.g., metacognitive processes). With regards to the hypotheses proposed in the introduction, the results from the current study are most consistent with the idea that testing homologizes performance across the ability range; although, intermediate retrieval did not completely equate students with low and high gF (the astute reader will recall the quote from the head of this article).

To view these results in a larger context one should consider Cronbach and Snow's (1977) aptitude × treatment interactions framework. Clearly, the finding that testing improves low ability students more so than high ability students fits nicely within this framework and is informative in several regards. Researchers interested in the theoretical basis for the testing effect must design research aimed at understanding why individuals who perform well in general intelligence tasks or have above average episodic memory abilities do not benefit from the testing effect in certain situations. Moreover, the intelligence-testing effect relation demonstrated herein should resonate with researchers who are actively implementing testing procedures in classrooms and other applied settings (McDaniel et al., 2007). The current results indicate that future work in theoretical and applied psychological settings begin examining and accounting for individual differences in the testing effect. Cronbach and Snow's (1977) framework is important for interpreting individual differences in the testing effect in these settings.

Taken together, the results of the current study are indicative that high-ability students capitalize on controlled processes such as cue-driven search processes and fluid abilities to encode information. These results are consistent with McCabe's (2008) demonstration that participants engage in covert retrieval during span tasks and that covert retrieval benefits subsequent memory. Perhaps participants in the upper EM quartile engaged in more elaborative encoding by sneaking in additional retrieval attempts while encountering the restudy pairs. If so, this would mitigate the testing effect for high-ability students. Clearly, future research should be directed at examining why individuals with better EM abilities have diminishing returns from testing. In terms of gF processes, perhaps participants in the upper intelligence quartile used strategies to better elaborate, or abstract, both the studied and tested cue-target pairs. These encoding strategies would have a similar mitigating effect as covert retrievals may have caused for participants in the upper EM quartile. Of course, these claims are speculative at this point but they suggest several important lines of research for psychologists who are interested in the testing effect and individual differences in the effects of retrieval from long-term memory.

Future research can also shed light on several methodological issues inherent in examining individual differences in the testing effect. That is, the possibility exists that high and low ability students may scale memorial information differently. In the current study, we included additional analyses to mitigate the effect of scaling issues. However, to fully alleviate this concern, it would be optimal to find testing paradigms that equated baseline memory performance between the two groups. A related and unexplored issue in the testing literature concerns the magnitude of testing effects at various levels of memory performance. Future research can elucidate exactly how testing benefits subsequent retrieval for weak versus strong memories (see Carpenter (2009) for empirical work with weak and strong associations). Given the robustness of the testing effect across multiple paradigms and stimulus sets (Roediger et al., 2010), another useful direction will be to examine whether individual differences in the testing effect are domain general across materials and tasks, or whether they are paradigm specific. Researchers who are interested in exploring individual differences in the testing effect should begin exploring these issues with more experimental and theoretical rigor. Feedback may be another crucial aspect driving individual differences in the testing effect. In the current work, we provided feedback and found aptitude × treatment interactions consistent with testing homologizing memory across the ability range. In Tse et al. (2010), participants where given no feedback (Experiment 1) and feedback (Experiment 2). In their Experiment 1 older adults failed to show benefits from testing whereas in Experiment 2 they showed larger effects than younger adults. Therefore, future individual differences research examining the testing effect should attempt to specify the conditions under which aptitude × treatment interactions emerges.

Conclusion

The primary goal of the current research was to implement a large-scale individual differences study of the testing effect in a college sample to elucidate any relations with higher-order cognition. Given the profusion of recent research investigating the testing effect, it is a significant question whether the testing effect extends to all students in the same manner, helps high ability students more than low ability students, or helps low ability students more than high ability students. Certainly, the restricted sample of participants in terms of age (18 year old college students) and intelligence scores supports the idea that these patterns should only strengthen when a more representative sample is examined. With regards to Neisser et al. (1996) quote at the head of this report, the current research points to a specific role of the testing effect in ameliorating the correlation between intelligence and test scores in a group of college students. However, the correlation has not yet been fully eliminated. This research demonstrates an important relation between the testing effect and episodic memory and intelligence abilities. Clearly much more research of this nature is needed.

Acknowledgments

This research was completed in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Georgia. This work was funded by the University of Georgia Dean's Award. We thank Greg Spillers, Justin Knight, Richard Marsh, and Brett Clementz for their thoughtful commentaries throughout this project. We also thank Henry Roediger, Andrew Butler, and Douglas Detterman for their thoughtful comments on a previous draft of this manuscript.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11*, 159–177.

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

Brewer, G. A., Unsworth, N., & Spillers, G. J. (submitted for publication). Working memory, interference, and the testing effect. *Manuscript Currently*.

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

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*, 826–830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153–170.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Irvington.

Dudukovic, N. M., DuBrow, S., & Wagner, A. D. (2009). Attention during memory retrieval enhances future remembering. *Memory & Cognition, 37*, 953–961.

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

Izawa, C. (1967). Function of test trials in paired-associate learning. *Psychological Reports, 75*, 194–209.

Jensen, A. (1998). *The G factor: The science of mental ability*. Greenwood Publishing Group.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*(2), 169–183.

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

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in Measuring Change* (pp. 22–38). Madison: University of Wisconsin Press.

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language, 58*, 480–494.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200–206.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

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

Rizzuto, D. S., & Kahana, M. J. (2001). An auto associative neural network model of paired-associate learning. *Neural Computation, 13*, 2075–2092.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory*. Brighton, UK: Psychology Press.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Thurstone, L. L., & Thurstone, J. (1962). *Test of primary mental abilities* (Revised ed.). Chicago: Chicago Science Research Association.

Tse, C. S., Balota, D. A., & Roediger, H. L. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging, 25*(4), 833–845.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30*, 128–134.

Unsworth, N., Spillers, G. J. (2010). Working memory capacity: Attention, memory, or both? A direct test of the dual-component model. *Journal of Memory & Language*.

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

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent variable analysis of the relationship between processing and storage. *Memory, 17*, 635–654.