Journal of Experimental Psychology:
Learning, Memory, and Cognition
2018, Vol. 44, No. 9, 1474–1486

0278-7393/18/$12.00 http://dx.doi.org/10.1037/xlm0000486

RESEARCH REPORT

The Benefits of Retrieval Practice Depend on Item Difficulty and Intelligence

Meredith Minear
University of Wyoming

Jennifer H. Coane, Sarah C. Boland, and Leah H. Cooney
Colby College

Marissa Albat
University of Wyoming

The authors examined whether individual differences in fluid intelligence (gF) modulate the testing effect. Participants studied Swahili–English word pairs and repeatedly studied half the pairs or attempted retrieval, with feedback, for the remaining half. Word pairs were easy or difficult to learn. Overall, participants showed a benefit of testing over restudy. However, almost 1/3 of the sample had a negative testing effect and benefitted more from restudy than testing, as well as performing better overall. These individuals self-reported less use of shallower encoding strategies than positive testing effect participants but did not differ in other dimensions. For individuals with a positive testing effect, difficulty had differential effects on participants who scored high or low on a measure of gF, with high gF participants showing larger testing effects for difficult over easy items, whereas low gF participants showed the opposite. Working memory performance was not related to the magnitude of the testing effect; however, vocabulary knowledge revealed a similar pattern as gF, with higher vocabulary associated with a testing effect for difficult but not easy items. This suggests that the benefit of retrieval practice varies with item difficulty and participant abilities. Thus, recommendations to engage in retrieval practice should take into consideration the interactive effects of to-be-learned materials and individual differences in the learners.

Keywords: retrieval practice, testing effect, individual differences, fluid intelligence

Tests are often viewed as a "necessary evil"—a tool for measuring performance, often associated with negative or aversive feelings. However, retrieving information from memory can be a powerful memory modifier. In other words, taking a test does not simply provide a measure of what has been learned. Rather, it can modify the memory trace in a way that increases its accessibility at a later date, actually improving learning relative to other forms of encoding, such as restudying (e.g., Roediger, Agarwal, Kang, & Marsh, 2010; Roediger & Butler, 2011; Roediger & Karpicke, 2006). The benefit of retrieval practice, or testing, relative to repeated study, is referred to as the *testing effect.*

Although the benefits of retrieval were first studied 100 years ago (Gates, 1917), the past decade has seen a substantial amount of empirical work conducted to specify the conditions under which testing effects arise and to understand the processes involved. In general, the effect is robust and has been reported for a variety of materials, age groups, and across diverse experimental designs (for reviews, see Karpicke, in press; Kornell & Vaughn, 2016; Rowland, 2014). However, the extent to which individuals may vary in how much they benefit from retrieval practice has only recently been explored. Although there are only few studies to date, the initial research provides some evidence that individual differences in cognitive ability can impact the magnitude of the testing effect, although the findings are inconsistent. Both Tse and Pu (2012) and Agarwal, Finley, Rose, and Roediger (2016) reported effects of individual differences in working memory capacity on the testing effect, whereas others found none (Brewer & Unsworth, 2012; Wiklund-Hornqvist, Jonsson, & Nyberg, 2014). Although Brewer and Unsworth (2012) found no effects of working memory or attentional control, they did find a larger testing effect in individuals with lower fluid intelligence (gF) and those with lower episodic memory ability relative to those scoring high on the same measures. However, a recent attempt to replicate the episodic

This article was published Online First April 12, 2018.

Meredith Minear, Department of Psychology, University of Wyoming; Jennifer H. Coane, Sarah C. Boland, and Leah H. Cooney, Department of Psychology, Colby College; Marissa Albat, Department of Psychology, University of Wyoming.

This work was partly funded by Social Sciences Division Grant 01.2222 from Colby College awarded to Jennifer H. Coane. Meredith Minear and Jennifer H. Coane contributed equally to this research. Jennifer H. Coane was supported by James McDonnell Foundation Understanding Human Cognition Award 220020426.

Correspondence concerning this article should be addressed to Meredith Minear, Department of Psychology, University of Wyoming, 1000 E. University Avenue Laramie, WY 82071. E-mail: mminear2@uwyo.edu

memory finding was unsuccessful, even with robust sample sizes and the use of the same materials as the original study (Pan, Pashler, Potter, & Rickard, 2015). Thus, the extent to which individual differences in ability or cognitive performance determines who most benefits from testing remains unclear.

Another factor that complicates examinations of individual differences in the testing effect is the relative lack of a clear understanding of the mechanisms that enhance memory for tested items compared to restudied ones. Clear theoretical models can help researchers identify potential processes to target as sources of interindividual variability. However, as many have noted, theoretical explanations of the mechanisms underlying the benefits of testing have lagged behind the empirical findings (Roediger & Butler, 2011; Rowland, 2014) although a number of hypotheses have been proposed. One key idea is that retrieval is more effortful than restudy and creates conditions favorable to later retention by increasing the amount of effort involved in learning, a desirable difficulty (Bjork, 1994), which subsequently increases the to-be-learned information's availability and accessibility (Tulving & Pearlstone, 1966). Consistent with this hypothesis (Pyc & Rawson, 2009; Rowland, 2014), more difficult initial tests (e.g., recall vs. recognition; Kang, McDermott, & Roediger, 2007) and longer retention intervals result in larger testing effects than easier tests or shorter delays.

In addition to effort or difficulty, variations in the type of processing involved in studying versus testing might play a role in the testing effect. There are different proposals for the forms of additional processing that can occur on retrieval trials, but not on restudy trials, that lead to better recall. The elaborative retrieval hypothesis (Carpenter, 2009) proposes that the act of retrieval results in a more extensive search of memory than restudy and thus greater activation in semantic memory of related concepts. This in turn creates additional retrieval paths and hence more routes for retrieval success in for tested relative to restudied items. Another influential approach focuses on episodic memory processes. In the episodic context account, the act of successful retrieval reinstates the previous temporal context during which items were initially studied and by doing so creates contextual cues that increase the later probability of recalling the needed information (Karpicke, Lehman, & Aue, 2014).

The accounts summarized above are not meant to be exhaustive and are not necessarily mutually exclusive. The benefits seen from retrieval practice potentially could be due to a combination of different factors and even that may depend on the type of memory task and other experimental factors such as the presence of feedback, interval between study and final test, and possibly individual differences (Rowland, 2014). For example, Agarwal et al. (2016) included manipulations of testing delay (10 min v. 2 days), presentation lag (i.e., the number of items presented between an initial encoding trial and a restudy or retrieval practice trial), and the presence of feedback as well as measuring individual differences in working memory. Individual differences in working memory were positively correlated with initial recall success during the encoding phase. Furthermore, there was a negative correlation between working memory and the testing effect, but only for the 2-day delay with feedback condition, suggesting that individuals lower in working memory had a larger testing effect. Presentation lag did not interact with condition or working memory capacity. Thus, working memory capacity did account for some variability in the testing effect, but only when feedback was provided.

As noted, Brewer and Unsworth (2012) did not find differences in the magnitude of the testing effect as a function of working memory or attentional control, but gF and episodic memory ability did predict testing benefits. Again, the relationship was negative so that individuals lower in ability, in this case gF or episodic memory ability, showed a greater testing effect than individuals high in these measures, consistent with Agarwal et al.'s (2016) findings for working memory. Brewer and Unsworth proposed that high capacity individuals are more likely to spontaneously employ better memory strategies such as elaborate encoding or to consciously test themselves on both study and test items resulting in little to no difference between these conditions, although it should be noted they did not directly test this hypothesis. Conversely, lower ability participants are less likely to self-initiate these strategies and therefore are forced to or learn to use more efficient retrieval strategies on the tested items, whereas they might continue to use less effective strategies, such as rote rehearsal, on restudied items. These findings suggested that testing may have the potential to equalize memory performance across ability levels (Brewer & Unsworth, 2012). However, although low capacity individuals in this study showed a benefit on tested items, their performance was still worse overall than that of the high capacity individuals. Therefore, as the authors noted, it is important to examine this difference under conditions where the two groups perform similarly. Furthermore, to demonstrate the effect is broadly generalizable to learners at different levels of ability, it is important to determine whether and under what conditions high gF individuals can still benefit from retrieval practice.

In the current study, we examined the potential roles of individual differences in cognitive ability as measured by working memory, gF, and crystallized intelligence on the magnitude of the testing effect using possibly the largest sample to date. We included working memory measures to address some of the inconsistencies in prior literature (e.g., Agarwal et al., 2016; Brewer & Unsworth, 2012). In addition, we manipulated item difficulty, allowing us to measure the testing effect for high and low capacity individuals at different levels of memory performance. If the lack of a testing effect in high capacity individuals is due to their use of more effective strategies or simply more resource availability, more challenging items or tasks might allow the benefit of testing to emerge. This would suggest that the testing effect depends not only on participant characteristics but also on the material being tested, as suggested by Cronbach and Snow's (1977) Aptitude $\times$ Treatment interactions framework (cf. Brewer & Unsworth, 2012). Furthermore, this would extend the interaction between ability and treatment (i.e., retrieval practice) to include task variables. Although a number of studies have manipulated the type of test (e.g., Kang et al., 2007; Pyc & Rawson, 2009), with more difficult retrieval activities resulting in larger testing effects than less effortful retrieval events, to our knowledge, item difficulty has been examined less thoroughly, in part because it can be challenging to quantify difficulty as a construct.

Overall, one might predict a larger testing effect for the more difficult items, consistent with the desirable difficulties (Bjork, 1994) framework. However, if individual differences in ability modulate the effect, an interaction between item difficulty and

measures of cognitive ability might moderate the effects of testing. Specifically, because retrieval success during the encoding phase is associated with larger testing effects (Rowland, 2014), it is possible that individuals lower in cognitive ability, when faced with difficult items, would experience more retrieval failures or exert less effort in attempting to retrieve the items. Thus, these individuals might only show a testing effect for easier items (cf. Brewer & Unsworth, 2012). It is worth noting that Rowland's meta-analysis showed that retrieval success moderated retrieval practice benefits in the absence of feedback; providing feedback, as was done in the current study, resulted in larger testing effects that were not strongly influenced by retrieval success. However, in an analysis on a subset of studies in the meta-analysis that included feedback, low initial retrieval success was associated with larger testing effects than studies with high retrieval success (Rowland, 2014). Furthermore, according to variants of the retrieval effort accounts (e.g., theory of disuse, Bjork & Bjork, 1992; bifurcation account; Kornell, Bjork, & Garcia, 2011), retrieval increases subsequent availability (i.e., storage strength) to a greater extent for items that are more difficult to retrieve than for items that are easier to retrieve (i.e., more accessible). Restudy, however, presumably results in equivalent increases in storage strength for all items (Kornell et al., 2011). Thus, retrieving more difficult items would strengthen the memory trace selectively for those items, yielding a larger testing effect.

Other theoretical accounts, although they do not explicitly address the role of individual differences in gF or other factors, make similar predictions. For example, according to the elaborative retrieval hypothesis (Carpenter, 2009), retrieval practice engages a search through semantic associates during testing events. One might speculate that individuals higher in crystallized intelligence (e.g., vocabulary knowledge) would have more elaborate semantic networks and this would be most evident for the more difficult items, yielding a larger testing effect on difficult items for individuals high in this measure than those scoring low. gF differences might influence search strategies—high gF individuals might engage a more elaborative search through semantic memory than low gF individuals. Along this vein, if controlled processes (Thomas & McDaniel, 2013) are more engaged in retrieval practice than restudy, one might expect that individuals scoring higher on measures of gF would have more effective inhibitory strategies allowing them to reduce interference from competing information. Finally, according to the episodic context account (Karpicke et al., 2014), because of the emphasis on the role of successful retrieval in updating contexts that become associated with the to-be-remembered information, individual differences in initial test performance might predict final test performance. In this account, retrieval practice serves to restrict a search set through episodic memory and retrieval mode (Tulving, 1983) in particular allows for effective reinstatement of context, which in turn is associated with later retrieval success. One possibility is that individuals higher in some measures (e.g., working memory, gF, inhibitory control) can more effectively update and restrict the search selection and thus benefit more from testing than their counterparts who score lower on the same measures. Thus, although many theoretical accounts do not explicitly predict a role for individual differences, it is likely that most of them could accommodate evidence suggesting specific factors modulate the testing effect.

Method

Participants

Three-hundred forty-three college students were recruited from Colby College (n = 140) and the University of Wyoming (n = 203) with 243 female participants and an average age of 19.8 years. All participants completed all tasks, as described below. Participants earned partial credit toward psychology courses or $30 for completing both sessions. The Institutional Review Boards at both Colby College and the University of Wyoming independently approved the study.

Materials

Paired associates learning task. Stimuli consisted of 48 Swahili–English word pairs from the Nelson and Dunlosky (1994) norms. Nelson and Dunlosky reported the average accuracy for 100 word pairs across three study-test trials. The original 100 pairs were split into ranked groups (ranked 1–8, where 1 refers to the most difficult items and 8 to the easiest items) based on average performance across the three study-test trials. We selected 24 items from the 1 and 3 ranks and 24 from the 6 and 8 ranks to create a set of 24 easier items and 24 more difficult items. These were further divided into two matched sets of easy and difficult items for counterbalancing. The average difficulty of the items in the norms, for difficult and easy items at Test 1, was .07 and .27. Test 3 accuracy was reported as .50 and .79, respectively.

During Session 1, 48 pairs (e.g., *mbwa–dog*) were presented for an initial study phase in random order for 6 s each. This was followed by four cycles of restudy and testing. In the restudy block, 24 pairs were presented again for 8 s each. In the test block, the other 24 pairs were tested by showing the Swahili cue for up to 7 s (e.g., *mbwa–*?) followed by a 1-s presentation of the correct English word (e.g., *dog*). Participants were encouraged to try to retrieve the English translation and type their response using the keyboard in the 7-s window. In all four study-test cycles, the restudy block preceded the test block (cf. Brewer & Unsworth, 2012), to avoid potential concerns with carryover effects from the preceding retrieval attempts. Item order was randomized anew in each block. At Session 2, participants were tested on their memory for all 48 pairs. They were given the Swahili word and then asked to type in the English translation with no time limit and no feedback. Word pairs were counterbalanced across restudy and test conditions.

Raven's advanced progressive matrices. In this task, a measure of general gF, participants saw a 3 × 3 matrix of shapes presented with the last shape missing and chose the item that best completed the pattern both across the rows and columns out of a set of eight options. We used both sets of items so that the dependent measure was the total number of items answered correctly out of a total possible 48 problems (Raven, Raven, & Court, 1998).

Working memory span. We used two measures of working memory, one in the verbal and one in the visuospatial domain as previous studies had only measured working memory span in the

verbal domain. We were interested in whether any relationships between the testing effect using a verbal task and working memory would generalize across working memory span tasks in both verbal and visuospatial domains.

Operation span—computerized version. In this task, which measures verbal working memory capacity, participants were asked to remember a series of three to seven letters presented one at a time. In between letter presentations, participants solved a simple mathematical problem and then clicked on the letters seen in the order presented at the end of each trial. Participants completed 15 trials and the score was the total sum of letters recalled in the correct position with a maximum score of 75 (Unsworth, Heitz, Schrock, & Engle, 2005).

Symmetry span—computerized version. In this task, which assesses visuospatial working memory, participants were asked to remember a sequence of two to five locations presented one at a time in a 4 × 4 grid. Between each location presentation, participants were shown a shape and decided whether it was symmetrical or not. At the end of each trial, participants indicated the locations seen in order using a 16-location array. Participants completed 12 trials. The score was the sum of locations recalled in the correct position with a maximum score of 42 (Unsworth, Redick, Heitz, Broadway, & Engle, 2009).

Shipley vocabulary task. This is a standardized measure of a participant's vocabulary. It is a 40-item vocabulary test that requires the respondent to choose which of four listed words "means the same or nearly the same" as a specified target word (Shipley, 1940).

Self-reported strategy. Participants were asked to write a response to the following question at the end of the final testing session: "What are some ways in which you tried to learn the English translation of the Swahili word? In other words, what are/is the method(s) you used to help you remember the word pairs?" and were given a page to write their answer. These self-reports were then coded by two separate raters as involving deep, intermediate, or shallow processing strategies. Deep processing strategies involved attending to imagery, meaning, or creating meaningful sentences (e.g., relating the word for garden, *bustani*, with Boston Garden, where a favorite team plays). Shallow processing strategies focused on shared sounds or number of letters, or simply repeating the pairs out loud. Intermediate strategies were those that combined sound based encoding with some more semantic processing (e.g., pronouncing a Swahili word and making a sentence with it). Raters also separately noted whether participants described using the specific strategy of self-testing as well as the use of the keyword method (McDaniel & Pressley, 1984). The keyword method, frequently used in foreign language courses, involves finding a known word neighbor to a cue and creating a mental image connecting it to the known translation (e.g., *wingu* means cloud, so one might connect *wingu* to *wing* and bird and visualize a bird flying in the clouds; Pyc & Rawson, 2012). Interrater reliability was .9.

Procedure

There were two 90-min sessions scheduled 48 hr apart. Session 1 began with the paired associates task in which participants completed an initial study phase followed by four cycles of restudy and testing. Raven's was then administered. Session 2 started with the final test of the paired associate task, followed by the Operation Span, Shipley, Symmetry Span, and a series of personality survey measures (not reported here[1]) as well as basic demographic information and an open-ended question asking participants to describe the method or methods they used to help remember the word pairs. Tasks were administered across the two days in a way to equate session duration. Debriefing took place after Session 2 was complete.

Results

Two participants' operation span data and four participants' Shipley scores were lost due to computer error. Descriptive data for all measures are reported in Table 1.

Testing Effect

We began our analyses by determining the size of the overall testing effect as the magnitude of this effect can vary considerably across studies. For example, Brewer and Unsworth (2012) reported an average testing effect of 7%, while Pan et al. (2015) found a testing effect of 17% using the same materials. In our data, there was a significant testing effect with tested items (M = 52%) recalled more than studied items (M = 45%) for a testing benefit of 7%, $F(1, 342) = 54.9$, $p < .001$, $\eta_p^2 = .14$. Easy items (M = 62%) were recalled more than difficult items (M = 35%), $F(1, 342) = 1{,}149.2$, $p < .001$, $\eta_p^2 = .77$, and a trend toward an interaction between difficulty and item type with a testing effect of 8% for easy items and 6% for difficult items, $F(1, 342) = 3.8$, $p = .053$, $\eta_p^2 = .01$ To determine the generality of the testing effect in our sample, we examined which participants showed a positive (testing > restudy), negative (testing < restudy), or null (testing = restudy) testing effect. Sixty-one percent of our participants had a positive testing effect, 8% had a null testing effect, and 31% had a negative testing effect.

Positive Versus Negative Testers

Given the size of our overall sample and that over a third of our participants did not show a benefit from testing, we first investigated possible individual differences between participants showing a positive testing effect (hereafter, *positive testers*, n = 210) and those who showed a negative testing effect (*negative testers*, n = 106). Because the testing effect is a difference score, it could simply be that negative testers are overall poor performers and are not actively engaging in the opportunity to practice recall during initial encoding. However, an initial comparison of overall final recall performance found that negative testers (M = .54, 95% confidence interval [CI]: .50, .58) significantly outperformed positive testers (M = .45, 95% CI [.42, .46]) $t(314) = -3.2$, $p < .001$. This suggests that the benefits on studied items for negative testers are larger in magnitude than the benefit for tested items for positive testers. A further examination of final performance broken down by study condition found that, although positive

[1] The additional measures, which are not reported here, assessed personality (NEO), grit, academic entitlement, stress, need for cognition, academic self-efficacy, and test anxiety. A self-report strategy measure on study habits was also administered.

Table 1
Descriptive Statistics and Reliabilities for All Measures

Measure	*M*	*SD*	Skew	Kurtosis	Reliability
Easy studied	.57	.28	−.12	−1.0	.83
Easy tested	.66	.24	−.44	−.37	.72
Easy testing effect	.08	.21	−.01	−.22	.45
Difficult studied	.32	.28	.75	−.49	.85
Difficult tested	.38	.28	.54	−.69	.81
Difficult testing effect	.06	.22	.02	.26	.57
Operation span	40.8	18.1	−.26	−.54	.74
Symmetry span	19.6	9.2	.42	.73	.54
Ravens	28.1	8.8	−.28	−.65	.91
Shipley	29.1	4.4	−.21	−.17	.74

testers had a significantly 9% better average performance on tested items than negative testers, the average performance difference between positive and negative testers for studied items was 23%. This did not interact with difficulty, $F < 1$. These data are summarized in Table 2.

We then compared the groups' Session 1 performance to assess whether the group differences on final recall of tested items were present at initial encoding. We used a repeated measures analysis of variance (ANOVA) with encoding block and item difficulty as within-subjects factors and positive versus negative testing effect as a between-subjects factor. There were significant main effects of encoding block and difficulty with improving recall performance from Block 1 to Block 4, $F(1, 314) = 1523.4$, $p < .001$, $\eta_p^2 = .83$, and better recall performance for easy compared to difficult items, $F(1, 314) = 1282.5$, $p < .001$, $\eta_p^2 = .80$. There was also an interaction between difficulty and block with a larger improvement from Block 1 to Block 4 for easy items than difficult items, $F(1, 314) = 74.1$, $p < .001$, $\eta_p^2 = .19$. However, there was no main effect of group nor were there any interactions between group and difficulty or block, all Fs < 1. Therefore, individuals experiencing a negative testing effect did not show any differences in retrieval success during initial encoding compared to positive testers.

We then turned to our two remaining possible sources of differences, cognitive ability (as measured by working memory, gF and crystallized intelligence) and differences in self-reported strategy. There were no group differences on any of the cognitive measures between positive and negative testers, suggesting the two groups did not differ systematically in ability. These results are summarized in Table 3. Turning to the strategy data, there was evidence of some differences between groups. A greater percentage of positive testers reported using shallow processing strategies whereas a greater percentage of negative testers reported using self-testing. These strategy data are summarized in Table 4.

Table 3
Means (SDs) of Individual Difference Measures and Encoding Performance for Positive and Negative Testers

Measure	Positive testers	Negative testers	*t*	*p*
Ospan	41.2 (18.5)	40.0 (18.4)	.56	.58
Symspan	19.9 (9.3)	18.8 (9.2)	1.1	.29
Ravens	28.0 (8.7)	28.3 (8.3)	−.34	.73
Shipley	29.3 (4.5)	28.8 (4.3)	1.0	.32
Session 1 recall	59 (.23)	.57 (.22)	1.2	.25

Note. Ospan = operation span; Symspan = symmetry span.

Individual Differences in the Magnitude of Retrieval Benefits

We then proceeded to test whether our individual difference measures affected the magnitude of the testing effect as reported by Brewer and Unsworth (2012; see also Pan et al., 2015). Given the initial difference between positive and negative testers reported above, we chose to analyze the possible effects of individual differences using only individuals showing a positive testing effect. Our logic for this was simple: (a) Our individual-differences variables did not differ between groups and do not appear to contribute to whether a participant showed a positive or negative testing effect. Instead, strategy use data suggest this may depend on the type of strategies one uses during encoding; and (b) the negative testing effect observed in our participants was not simply due to these participants performing worse on tested items than studied. Instead they had demonstrably better performance on studied items than positive testers as well as showing equivalent encoding performance at the end of Session 1. Therefore, if we are interested in whether individual differences interact with the benefits of retrieval practice that lead to better recall of tested items over restudied items, it is then logical to only include those individuals who demonstrate this effect in the analysis.

We first examined whether any of our individual differences measures were related to the overall testing effect and then moved to analyses taking item difficulty (easy vs. difficult) into account. None of our measures were significantly correlated with the overall testing effect. Correlations between all measures are reported in Table 5.

We then conducted a repeated-measures analysis of covariance (ANCOVA) with type of item (restudy vs. test) and level of difficulty (easy vs. difficult) as within-participants variables and

Table 2
Final Session 2 Recall Means (SDs) for Positive and Negative Testers as a Function of Item Type (Restudied vs. Tested) and Difficulty of Item (Easy vs. Difficult)

	Positive testers			Negative testers		
Pair	Study	Test	Testing effect	Study	Test	Testing effect
Easy pairs	.49 (.26)	.69 (.21)	.20	.74 (.22)	.62 (.24)	−.12
Difficult pairs	.24 (.24)	.41 (.29)	.17	.45 (.27)	.33 (.24)	−.12

Table 4
Percentage of Self-Reported Strategy Use Across Positive and Negative Testers

Strategy	Positive testers	Negative testers	χ^2	*p*
Shallow	82.1%	68.8%	6.6	.01
Intermediate	9.7%	12.5%	.51	.55
Deep	37.9%	44.8%	1.3	.31
Keyword	16.4%	21.9%	1.3	.26
Self-testing	2.6%	8.3%	5.0	.03

Note. Because multiple responses were possible, totals are higher than 100%.

our individual difference measures as covariates. There was a main effect of difficulty, $F(1, 205) = 21.8$, $p < .001$, $\eta_p^2 = .10$, with easy items remembered better overall than difficult items (easy $M = .59$, 95% CI [.56, .62], difficult $M = .32$, 95% CI [.29, .36]). There was a main effect of item type, $F(1, 205) = 7.81$, $p < .01$, $\eta_p^2 = .03$, (restudied items $M = .36$, 95% CI [.33, .40], tested items $M = .55$, 95% CI [.51, .58]) and an interaction between the type of item and level of difficulty, $F(1, 205) = 18.6$, $p < .001$, $\eta_p^2 = .08$, with a larger difference between tested and restudied items for easy items (mean difference = .20, 95% CI [.18, .22]) than difficult items (mean difference = .17, 95% CI [.14, .19]). Both our working memory measures significantly interacted with level of difficulty (operation span, $F(1, 205) = 7.6$, $p < .01$, $\eta_p^2 = .04$; symmetry span, $F(1, 205) = 5.8$, $p < .05$, $\eta_p^2 = .03$), but not with item type. Only Shipley and Raven's significantly interacted with both difficulty and item type, Shipley, $F(1, 205) = 6.2$, $p < .05$, $\eta_p^2 = .03$ and Raven's, $F(1, 205) = 6.4$, $p < .05$, $\eta_p^2 = .03$. These results are consistent with the correlational data, in which both Shipley and Ravens were positively correlated with the testing effect using difficult items and negatively correlated with the testing effect using easier items. Because Shipley and Raven's scores were significantly correlated, we used hierarchical regression to test whether these measures made separable contributions to the testing effect calculated using easy items and difficult items. The results of these analyses are shown in Tables 6 and 7. It appears that Raven's was a unique negative predictor of the testing effect for easy items, whereas both Raven's and Shipley were separable positive predictors of the testing effect for difficult items.

To better characterize the relationship between Raven's and difficulty, we used a quartile analysis similar to Brewer and Unsworth (2012) comparing the performance of participants in the top and bottom quartiles of Raven's using a mixed factorial ANOVA with Raven's quartile as a between-subjects factor and item type (restudy vs. test) and difficulty (easy vs. difficult) as within-subjects factors. Tested items ($M = .54$, 95% CI [.50, .58]) were remembered better than restudied items ($M = .35$, 95% CI [.31, .38]), $F(1, 106) = 334.1$, $p < .001$, $\eta_p^2 = .76$. Easier items ($M = .58$, 95% CI [.54, .61]) were remembered better than difficult items ($M = .31$, 95% CI [.27, .35]), $F(1, 108) = 397.4$, $p < .001$, $\eta_p^2 = .79$, and there was an effect of group with high gF participants ($M = .58$, 95% CI [.52, .63]) correctly recalling more items than low gF participants ($M = .31$, 95% CI [.26, .36]), $F(1, 106) = 53.8$, $p < .001$, $\eta_p^2 = .34$. There was also a Group × Item Type × Difficulty Interaction, $F(1, 106) = 19.1$, $p < .001$, $\eta_p^2 = .15$. The low gF group had a larger testing effect for easy than difficult items, whereas the high gF group had a significantly larger testing effect for difficult than easy items.[2] These data are shown in Figure 1.

We then examined the Session 1 encoding performance for the two groups as a function of difficulty. For the items that were tested during the acquisition phase, a 2 (group: high vs. low gF) × 2 (difficulty level: easy vs. difficult) × 4 (encoding block) mixed factorial ANOVA revealed better memory performance for high gF ($M = .46$, 95% CI [.42, .50]) than low gF ($M = .28$, 95% CI [.23, .30]) individuals, $F(1, 106) = 37.1$, $p < .001$, $\eta_p^2 = .26$. Participants improved their performance across the four encoding blocks ($Block_1 = .15$, 95% CI [.13, .17]; $Block_2 = .32$, 95% CI [.29, .35]; $Block_3 = .47$, 95% CI [.43, .51]; $Block_4 = .53$, 95% CI [.50, .57]), $F(3, 106) = 453.8$, $p < .001$, $\eta_p^2 = 81$, and performance on easier items ($M = .51$, 95% CI [.47, .54]) was higher than on difficult items ($M = .27$, 95% CI [.24, .30]), $F(1, 108) = 440.6$, $p < .001$, $\eta_p^2 = .80$. There were two significant two-way interactions: Block × Group, $F(3, 106) = 11.3$, $p < .001$, $\eta_p^2 = .10$, with greater improvements in accuracy across blocks by the high gF group than by the low gF group, and Block × Difficulty, $F(3, 106) = 18.6$ $p < .001$, $\eta_p^2 = .15$, and a significant three-way interaction between group, block, and difficulty, $F(3, 106) = 13.5$ $p < .001$, $\eta_p^2 = .11$. For easy items, the high gF group recalled more items than the low gF group, but the size of this difference stayed the same across blocks. However, for difficult items, the difference in performance between the groups grew larger from the first encoding block to the last block, with high gF individuals showing a steeper learning curve between Blocks 2 and 3 and overall higher performance than low gF individuals. These data are shown in Figure 2.

We also compared the encoding strategies reported between our high and low gF groups. Although there were no significant differences in reported use of shallow, intermediate and self-testing strategies, high gF individuals were more likely to report using deep strategies and the keyword strategy. These results are summarized in Table 8.

In sum, high gF individuals benefitted more from testing on difficult items than low gF individuals whereas the opposite was true for easy items. Interestingly, in both the encoding phase and final test phase, the high gF group's performance on difficult items was equivalent to the low gF group's performance on easy items, highlighting the importance of examining item difficulty as well as individual differences in ability.

A Preliminary Exploration of Individual Differences and Negative Testing Effects

Our initial examination of the testing effect revealed that over a third of our participants showed an advantage for restudied over tested items (i.e., a negative testing effect) and overall better memory performance at final test. Although our individual differences measures did not reveal any differences between these

[2] To rule out the possibility that this interaction was "removable" and due to measurement scaling issues, we used the logit (p) transformation as recommended in Wagenmakers, Krypotos, Criss, and Iverson (2012) and reran the analysis. The interaction was still present, $F(1, 60) = 11.5$, $p = .001$, $\eta_p^2 = .16$.

Table 5
Correlation Matrix for All Measures

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Studied	1.00												
2. Tested	.76**	1.00											
3. TE	−.45**	.24**	1.00										
4. Easy studied	.94**	−.71**	−.44**	1.00									
5. Easy tested	.71**	.93**	.22**	.68**	1.00								
6. Easy TE	−.46**	.11	.83**	−.56**	.22**	1.00							
7. Diff. studied	.94**	.72**	−.42**	.77**	.65**	−.30	1.00						
8. Diff. tested	.72**	.95**	.23**	.65**	.76**	−.003	.70**	1.00					
9. Diff. TE	−.30**	.29**	.84**	−.17**	.14**	.39	−.40**	.38**	1.00				
10. Ospan	.18**	.20**	.01	.14**	.14**	−.02	.19**	.23**	.05	1.00			
11. Symspan	.16**	.20**	.03	.18**	.20**	−.003	.13**	.16**	.05	.42**	1.00		
12. Ravens	.40**	.40**	−.05	.40**	.37**	−.11*	.35**	.38**	.03	.29**	.34**	1.00	
13. Shipley	.18**	.23**	.05	.18**	.18**	−.04	.15**	.24**	.11*	.18**	.16**	.22**	1.00

Note. TE = testing effect; Diff. = difficult; Ospan = Operation span; Symspan = symmetry span.
* Correlation is significant at the .05 level. ** Correlation is significant at the .01 level.

negative testers and individuals showing the traditional positive testing effect, we were curious to explore whether the extent of this study benefit for negative testers was predicted by any of our individual difference measures in the same way that retrieval benefits were in our positive testing individuals.

We performed the same repeated-measures ANCOVA with type of item (restudy vs. test) and level of difficulty (easy vs. difficult) as within-participants variables and our individual difference measures as covariates. There was a main effect for difficulty with easier items (M = .68, 95% CI [.64, .72]) recalled more than difficult items (M = .40, 95% CI [.36, .45]), $F(1, 101) = 316.8$, $p < .001$, $\eta_p^2 = .75$. Of our individual difference items, only Raven's was a significant covariate, $F(1, 101) = 9.2$, $p < .01$, $\eta_p^2 = .08$. There was a significant interaction between type of item and Raven's, $F(1, 101) = 5.8$, $p < .05$, $\eta_p^2 = .06$. No other effects were significant, minimum $F = 2.7$. The correlation between Raven's and the size of the negative testing effect was significant, $r = -.26$, $p < .01$, with larger benefits for restudied over tested items with increasing Raven's scores. These data are shown in Figure 3.

To test whether there were differences in reported strategy between high and low gF individuals in our negative testing sample, we also identified our top and bottom gF quartiles (high gF, n = 24; low gF, n = 26) and tested for differences in self-reported strategies. There were no differences, all $\chi^2 < 1$.

Discussion

The present study significantly extends our understanding of the relationship between individual differences in fluid abilities and the testing effect and raises important new questions about who may benefit most from testing. Furthermore, it raises the possibility that a significant percentage of the population benefits more from restudying items than being tested, even after a relatively long delay. First, we discuss our results pertaining to the positive testing effect and the interaction between difficulty and individual differences in gF on the effectiveness of retrieval practice. We will then move to an exploration of our negative testing data and conclude with an examination of the extent to which our results are consistent with different theories of the testing effect.

In our positive testing data, we found no relationship between the overall testing effect and our individual-differences variables. However, when the difficulty of the items was included as a factor, difficulty interacted with gF such that lower gF individuals showed a testing effect of 25% for easy items compared to 14% for difficult items, whereas high gF individuals had a testing effect of

Table 6
Summary of Hierarchical Regression Analysis for Predicting the Testing Effect for Easy Items

Variable	B	$SE(B)$	ß	t	p
Model 1, $F(1, 208) = 3.5$, $p = .06$, $R^2 = .02$					
Shipley	−.005	.002	−.13	−1.9	.06
Raven's					
Model 2, $F(2, 207) = 4.3$, $p < .05$, $R^2 = .04$					
Shipley	−.004	.002	−.10	−1.5	.14
Raven's	−.003	.001	−.15	−2.2	.03

Note. Shipley scores were entered as a predictor in the first model then Raven's was added in the second model.

Table 7
Summary of Hierarchical Regression Analysis for Predicting the Testing Effect for Difficult Items

Variable	B	$SE(B)$	ß	t	p
Model 1, $F(1, 208) = 6.8$, $p < .01$, $R^2 = .03$					
Shipley	.003	.18	2.6	.01	.003
Raven's					
Model 2, $F(2, 207) = 6.1$, $p < .01$, $R^2 = .06$					
Shipley	.006	.003	.15	2.2	.03
Raven's	.003	.001	.16	2.3	.03

Note. Shipley scores were entered as a predictor in the first model then Raven's was added in the second model.

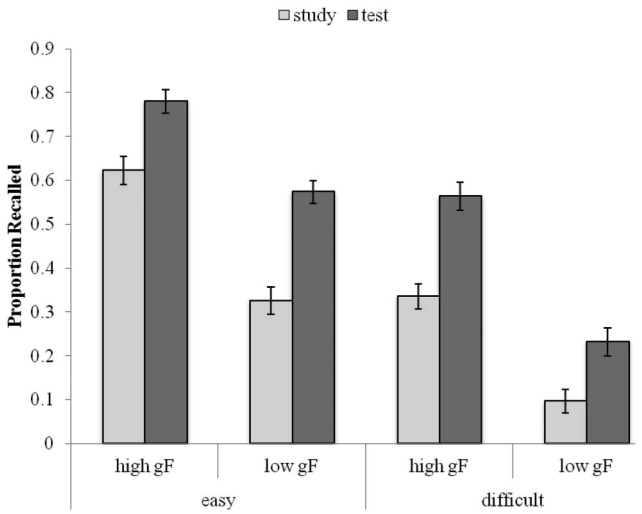

Figure 1. Final test performance as a function of condition (study vs. test), item difficulty (easy vs. difficult), and group (high vs. low fluid intelligence [gF]). Error bars represent standard error.

23% for difficult items and 16% for easy items. The data from our high gF participants is consistent with the desirable difficulties framework, such that a more difficult initial test leads to a larger testing effect (cf. Rowland, 2014). However, our low gF participants had a smaller testing effect for difficult compared to easy items.

To further explore the roots of the different magnitude of the testing effect in our sample, we examined whether final performance was contingent upon encoding success during the learning phase. The largest testing effects were observed in both groups when retrieval success in Session 1 was around 55% correct. Specifically, for low gF individuals the Session 1 retrieval success for easy items was .58 and the corresponding testing effect was 25%, whereas for high gF individuals, the Session 1 retrieval success for difficult items was .52 and the testing effect was 23%. Smaller sized testing effects were observed under conditions of

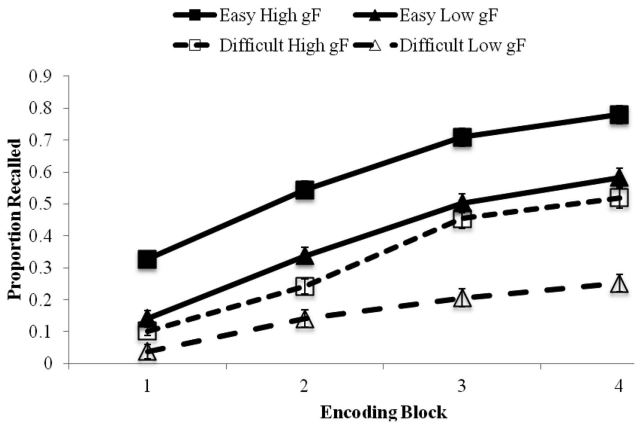

Figure 2. Mean performance for tested items during the encoding phase as a function of block (1–4), item difficulty (easy vs. difficult), and group (high vs. low fluid intelligence [gF]). Error bars reflect standard error.

Table 8
Percentage of Self-Reported Strategy Use Across High and Low gF Groups

Strategy use	High gF	Low gF	X^2	*p*
Shallow	75%	73%	.11	.75
Intermediate	9%	9%	.004	.95
Deep	57%	20%	15.4	.001
Keyword	28%	7%	8.2	.004
Self-testing	4%	0%	2.1	.15

Note. gF = fluid intelligence. Because multiple responses were possible, totals are higher than 100%.

both higher (.78 for the high gF group recalling easy words) and lower (.25 for the low gF group recalling difficult words) retrieval success at encoding. Thus, when performance is equivalent between high and low performers, a similar sized testing effect is present. This pattern suggests a sweet spot for initial retrieval success in this paradigm, such that both high and low success at encoding leads to smaller testing effects than an intermediate amount of success. When this level of initial retrieval success is reached appears to depend on both the materials used and individual differences in gF.

One important question is then: What was it about some word pairs that made them more difficult than others and might that relate to differences in gF? In the development of the Swahili–English word pairs used in this study, Nelson and Dunlosky (1994) based their difficulty designation on empirically based participant recall. In speculating on what factors may have contributed to some word pairs being easier to remember than others, they noted a positive relationship between the frequency of occurrence in the English language of the target word and the probability of recall. They proposed that baseline familiarity with the English word could influence how easily it could be linked with its Swahili translation, either by requiring less processing time to read the English word and thereby freeing up more time to work on the association or by providing more information in long-term memory that could be used to form associations. Our high and low gF groups differed significantly on the Shipley (1940) vocabulary measure, suggesting differences in word familiarity and experience. In addition, Shipley appeared to predict additional variance in the testing effect using difficult items beyond that accounted for by gF. One possibility is that individuals possessing larger vocabularies have a larger semantic network affording a greater number of possible retrieval paths, especially for difficult items. This would be consistent with the elaborative encoding hypothesis (e.g., Carpenter, 2009), which proposes that the act of retrieval leads to greater activation of semantic memory than restudy. Although other studies have not found the role of prior knowledge in retrieval practice to have an effect, they did not include individual differences in gF or strategies as factors (Xiaofeng, Xiao-e, Yanru & AiBao, 2016).

We compared the strategies reported by our high and low gF participants and, whereas the two groups did not differ in their reported use of shallow strategies, high gF participants did report greater use of more elaborative strategies and specifically use of the keyword strategy. It may be that high gF individuals used shallower strategies on easier items but switched to more elabo-

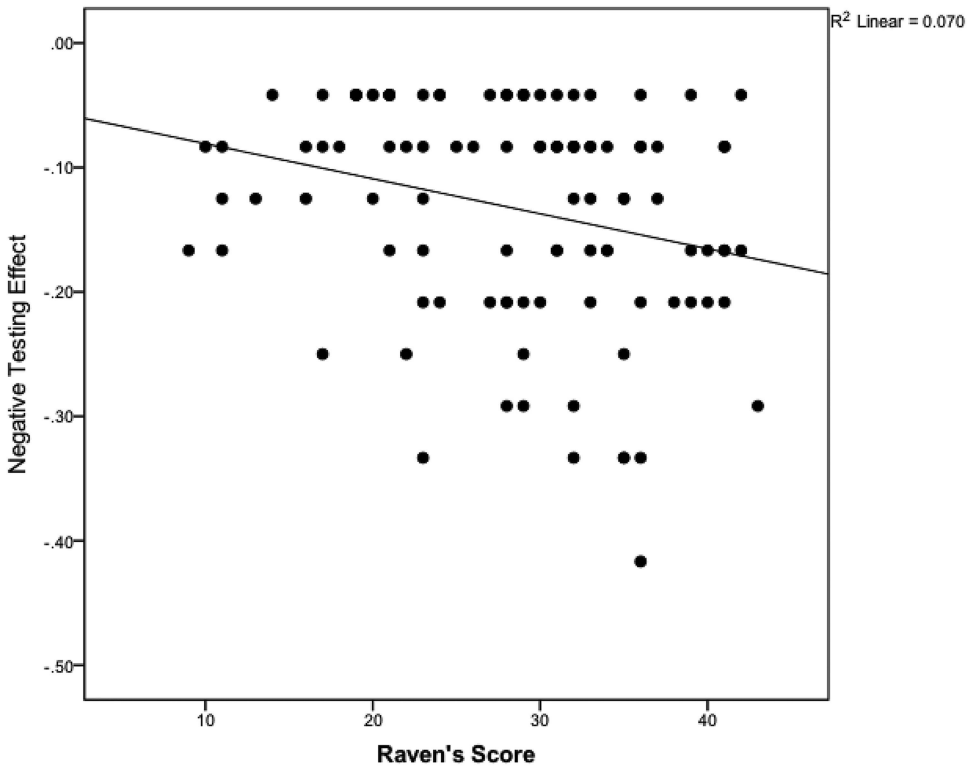

Figure 3. The correlation between Raven's score and the negative testing effect.

rative strategies for difficult items or that high gF individuals began with shallower strategies and then switched over to more elaborative strategies across encoding blocks. An examination of the interaction in between block, difficulty, and gF quartile suggests a combination of these possibilities, as there is a jump in performance specifically on difficult items between encoding Blocks 2 and 3 for high gF participants, but not for low gF participants. There is evidence that strategy shifts from less to more effective can occur over the course of learning in paradigms such as spacing and directed forgetting (e.g., Delaney & Knowles, 2005). It may be that the extent to which such shifts occur may be dependent, in part, on individual differences in in gF. One possibility is that such a shift in strategy was involved in how participants processed and used the feedback following an unsuccessful retrieval event. However, because all our self-reported strategy data were collected retrospectively at the end of Session 2, this remains a speculation that needs to be addressed in future work.

We will now turn briefly to the results of negative testers, who comprised approximately one third of our sample. It is important to note that many studies do not address or even report the percentages of participants who did not show a benefit from testing. Brewer and Unsworth (2012) reported that 12% of their sample showed no testing effect and 21% showed a negative testing effect. However, they did not examine these participants' performance separately. It might be easy to dismiss these negative testers as spurious save for four important results: (a) negative testers scored better than positive testers overall; (b) negative testers had equivalent performance at encoding and did not differ in working memory capacity, gF, or crystallized intelligence from positive testers; (c) the size of the benefit of study over test was positively related to gF; and (d) this did not interact with difficulty, unlike retrieval benefits.

Recent work has demonstrated that negative testing effects can be produced under specific experimental conditions (Peterson & Mulligan, 2013; Mulligan & Peterson, 2015). Mulligan and Peterson (2015) have argued that testing effect theories should be able to account for situations in which restudy leads to better memory than retrieval practice. In their item-specific-relational account, item-specific features are those features that make a target distinctive and distinguishable from other targets in a list as well as features related to the relationship between a specific target and its cue (Mulligan & Peterson, 2015). There are also interitem relationships (i.e., associations that can be formed between all the target items). Retrieval is proposed as strengthening item-specific relationships, which in turn leads to better performance at final recall. However, in situations where processing interitem relationships (e.g., where the target words are semantically related to each other) could lead to substantial benefits, restudy leads to better performance as the act of retrieval focuses the participant on item-specific information, leaving fewer resources available for interitem processing. This account does not seem particularly applicable in the present situation, because we did not specifically include strong interitem relationships in our stimulus set.

However, there may be other types of beneficial processing that can occur during restudy. The opportunity for restudy has been hypothesized to affect the storage strength of an item (i.e., how well it has been learned; Bjork & Bjork, 1992), whereas testing has its effects on the retrieval strength of an item (i.e., how easily it can

be recalled; Wheeler, Ewers, & Buonanno, 2003). Individual differences in the types of processes or strategies used during restudy would then have a strong effect on the magnitude and direction of the testing effect. Delaney, Verkoeijen, and Spirgel (2010) described the results of a testing effect study of free recall in which study strategy was manipulated between groups. One group was assigned to use a simple rehearsal strategy, whereas the other group used a story-based strategy. In the rehearsal strategy group, the typical testing effect was observed. However, for the strategy group, restudied words were recalled better than tested words, even after a delay of 7 days. Our negative testers were significantly less likely to report using a shallow strategy and they clearly experienced less forgetting of restudied items (61% compared to 30% recall by positive testers) at Session 2. Whereas they did show more evidence of forgetting on the tested items compared to the positive testing group, the magnitude of the difference was considerably smaller than that observed for restudied items, with 48% recall of tested items compared to 55% for positive testers. Thus, consistent with Delaney et al. (2010), under some circumstances, negative testing effects occur, either because of directed strategy use, as in Delaney et al., or other, as of yet unclear, factors.

Interestingly, there were no differences in cognitive abilities between our negative and positive testers. However, within each group gF was positively related to the magnitude of the respective benefit. gF was positively related to the magnitude of the benefit of restudy over test, but this did not interact with item difficulty as it did with the positive testing effect. There also was no effect of vocabulary and self-reported strategy did not differ between high and low gF individuals who showed a negative testing effect.

Recent individual differences work by Carpenter et al. (2016) also reported a situation in which individuals benefited more from restudy than retrieval. This study was conducted in a university level introductory biology course and used student achievement (operationalized as course performance) as the individual difference measure. During in-class exercises, students either were asked to recall course definitions and diagrams or to recopy the information without retrieval demands. Learning was later assessed using a multiple-choice quiz. They found that although high achieving students benefited from retrieval, low- and middle-performing students actually benefited more from recopying information. These data combined with our own and those reported by Delaney et al. (2010) suggest that negative testing effects can be found in multiple paradigms and may be related to individual differences in strategy use. However, the extent to which an individual consistently benefits more from restudy across paradigms or materials is unknown. It may be that negative testing effects arise from a match between an especially effective strategy and a particular learning paradigm such as cued recall.

Overall, the results from this study have important implications for future work on understanding the role of individual differences in the effectiveness of testing to enhance learning. The first is that the extent to which an individual benefits more from testing or restudy may be an individual difference in itself perhaps linked to individual differences in strategy or organization (Ozier, 1980) rather than cognitive ability. GF did not differentiate between positive and negative testers, but within each group was predictive of the overall benefit (i.e., restudy over retrieval or retrieval over restudy). Fluid ability may predict how well an individual uses an effective strategy or the ability to flexibly adopt better strategies in response to difficulty. This is consistent with recent work suggesting that learning and retrieval processes are significantly related to gF even when controlling for working memory (Unsworth, 2010; Unsworth, Brewer, & Spillers, 2009; Wang, Ren, & Schweizer, 2017). The second implication is that, within individuals showing a positive testing effect, differences in fluid and crystallized intelligence interact with the difficulty of the material.

The present study contributes to an ongoing analysis of individual differences in the retrieval practice effect. Variation in sampling, the difficulty of the materials to be learned and individual differences in the ability to learn through restudy were all raised by Pan et al. (2015) as important issues to be addressed in researching individual differences in the effectiveness of testing to enhance learning. The results from our study can help explain how two studies using very similar materials and methods such as Brewer and Unsworth (2012) and Pan et al. (2015) can produce conflicting results even when using similar methods and the same materials. First, the inclusion of individuals showing a negative testing effect in overall analyses of individual differences would appear to be problematic: In our data the magnitude of the negative testing effect was positively related to gF. This has the potential to mask the effects of individual differences if negative and positive testing effects are simply characterized as a continuous variable indexing the same processes. It is also unclear how much the percentage of negative testers as well as individuals with no testing effect can vary from sample to sample. Within our data, they constituted 39% of our participants. In Brewer and Unsworth (2012) they were 32% of their sample while in Pan et al. (2015), they were only 18% of participants tested (S.Pan, personal communication, November, 11, 2017). Second, the effect of gF on the testing benefit interacts with the difficulty of items, which can vary based on the levels of education and experience present in the sample tested.

Finally, it is important to consider how the findings here can or cannot yet be accommodated by current theories of the testing effect. The desirable difficulties account has already been proposed as a useful framework for the consideration of individual differences and our results are consistent with that approach and the Aptitude $\times$ Treatment interaction proposed by Brewer and Unsworth (2012). However, our data suggest a boundary effect with difficulty for lower ability participants even with the presence of feedback on trials where the item is not retrieved. As noted, when high and low gF participants had approximately 50% retrieval success during Session 1, we saw a similar sized benefit of testing over restudy. Kornell and Vaughn (2016) argued that successful retrieval and unsuccessful retrieval attempts with feedback both enhance learning, but it appears that a larger percentage of unsuccessful retrievals can lead to smaller testing effects than when the probability of success and failure is roughly equal. Our negative testing data did not show an effect of difficulty. However, the broader of approach of the theory of disuse (Bjork & Bjork, 1992, 2011) may provide some theoretical framework for the existence of a negative testing effect. According to this theory, memory performance relies on two sources, retrieval and storage strength. Retrieval strength is based on how easily an item can be recalled from memory while storage strength refers to how well an item is stored and integrated in memory. As noted above, it is possible that negative testers are employing better restudy techniques than simply rereading or repeating the word pairs that then enhanced the items' storage strengths. Unlike retrieval strength,

storage strength has not been proposed as being affected by difficulty and we did not find an effect of item difficulty on the negative testing effect. An interesting manipulation for future research would be to extend the retention period. According to the theory of disuse, retrieval strength decays with time whereas storage strength does not. Therefore, the performance advantage of negative testers compared to positive testers on restudied items hypothetically should grow larger with time.

The second theory described in the introduction was that of elaborative encoding (Carpenter, 2009). In this account of the testing effect, the act of retrieval leads to additional meaning based processing or elaboration of the item that then provides additional retrieval cues at the final test. Individual differences would then seem to arise based on whether participants are or are not consciously employing elaboration-based strategies. Those who are not would therefore have larger testing effects as they would benefit the most from such processing arising as a result of retrieval practice. This would be consistent with our data as the only evidence of an individual difference in positive versus negative testers was that positive testers were more likely to report use of shallower strategies. For negative testers, more effortful elaboration may have been disrupted by retrieval leading to greater benefits of restudy over test. A second point that would appear consistent with elaborative encoding was the additional variance for difficult items in our positive testers accounted for by individual differences in vocabulary beyond that predicted by gF. In the elaborative encoding hypothesis, greater difficulty leads to more elaboration of weaker cues and a greater search of semantic memory than easier retrieval. Individuals with a larger vocabulary would have more words that could become retrieval cues, which would especially advantageous under greater difficulty. Therefore, elaborative encoding would appear to be consistent with several points in our results.

The episodic context account (Karpicke et al., 2014) proposes that attempting retrieval results in an enhanced memory representation due to updates in the context surrounding the item to be retrieved. The reinstatement of prior contexts then leads to fewer retrieval paths (in contrast to the increased number proposed by elaborative encoding) for more efficient searching of memory and better performance for items that were previously retrieved. The effect of difficulty in this account is more closely tied to the extent to which contextual updating is needed. Greater variability across different retrieval contexts including temporal context leads to a more elaborate memory trace and a larger testing effect. The act of restudying is less effective because it does not lead to contextual updates. Our study did not provide much variation in context as retrieval practice was massed in one session so our results are difficult to interpret in terms of episodic context. Brewer and Unsworth (2012) did find evidence that the testing effect is affected by individual differences in episodic memory performance and based on our data, we might predict a similar crossover interaction with difficulty as defined by manipulating the hetero or homogeneity of the retrieval contexts. In terms of negative testing effects, the episodic context account would not appear to make specific predictions although its proposers have directly criticized the role of elaborative encoding as the source of retrieval benefits (Lehman, Smith, & Karpicke, 2014) as well as examining the benefits of testing against more active restudy conditions involving elaborative encoding. Karpicke and Blunt (2011) directly compared memory performance using elaborative encoding (i.e., concept mapping) to retrieval practice and found that the vast majority of students benefited more from retrieval. However, 16% of their participants showed performance equivalent to or better than repeated testing using conceptual mapping. Although this is only half the percentage of non- or negative testers seen in our study, it is worth noting that the concept mapping technique was taught to the participants on the first day of the experiment and it is unclear whether any had had prior experience in using this study technique. It is possible that individual differences in cognitive ability such as gF as well as prior experience in spontaneously applying a particular strategy affect the extent to which it is competitive with the benefits of retrieval.

There are a number of questions that will be important to address in future research. Although a few other studies, as described above, have reported negative testing effects, it is unknown whether this is a stable individual difference that is reliable across materials and memory paradigms such as recognition and free recall. Our retrospective self-reported strategy data pointed toward strategy differences between positive and negative testers, but it will be important to replicate this finding with more sophisticated measures of strategy as well as experimental manipulations of strategy such as those used by Delaney et al. (2010) or Karpicke and Blunt (2011). In addition, future studies with stronger manipulations of retrieval contexts can help assess whether contextual updating plays any role in the individual differences seen in the testing effect to date.

Conclusion

These results have important implications for educational practice. The benefits of testing may have an optimal effect at a specific range of difficulty, which may be related to individual differences in study strategies, knowledge, or ability. It is important to understand the possible boundaries of the testing effect especially if some students may not benefit as much or at all from testing. There may be a level of knowledge where testing is most effective, but for students for whom the material is too easy or too difficult or for those with effective self-directed strategies, testing may be ineffective or frustrating when the expectation is that testing will enhance performance. One important caveat is that, although it appears that some participants' performance is disrupted by testing, there are additional benefits of retrieval practice that go beyond the direct effects on retention. In educational and applied practice, testing also yields important indirect effects (e.g., Roediger, Putnam, & Smith, 2011), including more frequent review and study of material, feedback on performance, and improved metacognition (Karpicke & Blunt, 2011). Thus, we recommend that students and instructors continue to incorporate retrieval practice as an effective tool toward greater learning.

References

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2016). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory, 25,* 764–771.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In

M. A. Gernsbacher, R. W. Pew, & L. M. Hough (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In S. M. Kosslyn & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Routledge.

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

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

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28,* 353–375. http://dx.doi.org/10.1007/s10648-015-9311-9

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

Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in the free recall of unmixed lists. *Journal of Memory and Language, 52,* 120–130. http://dx.doi.org/10.1016/j.jml.2004.09.002

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63–147. http://dx.doi.org/10.1016/S0079-7421(10)53003-2

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives de Psychologie, 6,* 1–104.

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

Karpicke, J. D. (in press). Retrieval-based learning: A decade of progress. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2. Learning and memory: A comprehensive reference* (J. H. Byrne, Series Ed.). http://dx.doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. http://dx.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). San Diego, CA: Elsevier Academic Press.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. http://dx.doi.org/10.1016/j.jml.2011.04.002

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation, 65,* 183–215. http://dx.doi.org/10.1016/bs.plm.2016.03.003

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,* 1787–1794. http://dx.doi.org/10.1037/xlm0000012

McDaniel, M. A., & Pressley, M. (1984). Putting the keyword method in context. *Journal of Educational Psychology, 76,* 598–609. http://dx.doi.org/10.1037/0022-0663.76.4.598

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 859–871. http://dx.doi.org/10.1037/xlm0000056

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

Ozier, M. (1980). Individual differences in free recall: When some people remember better than others. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory: Advances in research and theory* (pp. 309–364). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60164-4

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language, 83,* 53–61. http://dx.doi.org/10.1016/j.jml.2015.04.001

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1287–1293. http://dx.doi.org/10.1037/a0031337

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. http://dx.doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 737–746. http://dx.doi.org/10.1037/a0026166

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for raven's progressive matrices and vocabulary scales*. New York, NY: Psychological Corporation.

Roediger, H. I., Agarwal, P. K., Kang, S. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *Current issues in applied memory research* (pp. 13–49). New York, NY: Psychology Press.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford, UK: Elsevier. http://dx.doi.org/10.1016/B978-0-12-387691-1.00001-6

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

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology: Interdisciplinary and Applied, 9,* 371–377. http://dx.doi.org/10.1080/00223980.1940.9917704

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 437–450. http://dx.doi.org/10.1037/a0028886

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, 253–264. http://dx.doi.org/10.1037/a0029190

Tulving, E. (1983). *Elements of episodic memory*. New York, NY: Oxford University Press.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior, 5,* 381–391. http://dx.doi.org/10.1016/S0022-5371(66)80048-8

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica, 134,* 16–28. http://dx.doi.org/10.1016/j.actpsy.2009.11.010

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity-fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review, 16,* 931–937. http://dx.doi.org/10.3758/PBR.16.5.931

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. http://dx.doi.org/10.3758/BF03192720

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & 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. http://dx.doi.org/10.1080/09658210902998047

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40,* 145–160. http://dx.doi.org/10.3758/s13421-011-0158-0

Wang, T., Ren, X., & Schweizer, K. (2017). Learning and retrieval processes predict fluid intelligence over and above working memory. *Intelligence, 61,* 29–36. http://dx.doi.org/10.1016/j.intell.2016.12.005

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580. http://dx.doi.org/10.1080/09658210244000414

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

Xiaofeng, M., Xiao-e, Y., Yanru, L., & Zhou, A. (2016). Prior knowledge level dissociates effects of retrieval practice and elaboration. *Learning and Individual Differences, 51,* 210–214. http://dx.doi.org/10.1016/j.lindif.2016.09.012

Received December 26, 2015
Revision received July 25, 2017
Accepted July 25, 2017 ■