Journal of Experimental Psychology: Applied
2012, Vol. 18, No. 3, 253–264

1076-898X/12/$12.00 DOI: 10.1037/a0029190

The Effectiveness of Test-Enhanced Learning Depends on Trait Test Anxiety and Working-Memory Capacity

Chi-Shing Tse and Xiaoping Pu
The Chinese University of Hong Kong, Hong Kong, China

Despite being viewed as a better way to enhance learning than repeated study, it has not been clear whether repeated testing is equally effective for students with a wide range of cognitive abilities. The current study examined whether test-enhanced learning would be equally beneficial to participants with varied working-memory capacity (WMC) and trait test anxiety (TA). Chinese–English bilingual undergraduates in Hong Kong were recruited as participants. They acquired Swahili–English word pairs (half via repeated study and half via repeated testing) and performed a delayed cued-recall test for all pairs about one week after the acquisition phase. Their WMC and TA were estimated by Unsworth, Heitz, Schrock, and Engle's (2005) operation-span task and the Chinese version of Spielberger's (1980) Test Anxiety Inventory, respectively. We replicated the typical testing effect: Participants performed better for pairs in the repeated-testing condition than those in the repeated-study condition. Regression analyses showed that, (a) relative to other participants, those with lower WMC and higher TA made more intralist intrusion errors (i.e., recalling a wrong English translation to a Swahili word cue) during the acquisition phase, and (b) the testing effect was negatively correlated with TA for participants with lower WMC, but was not correlated with TA for participants with higher WMC. This demonstrates a boundary condition for the use of test-enhanced learning. Implications of these findings for theories of the testing effect (e.g., Pyc & Rawson's, 2010, mediator-effectiveness hypothesis) and their application in classroom settings are discussed.

Keywords: test anxiety, testing effect, working-memory capacity

Tests have often been used to measure learning, rather than to facilitate learning, in spite of the evidence for the *testing effect*. This effect refers to the advantage in long-term retention for materials that are tested over those that are restudied, as reflected by a difference in performance between the repeated-testing condition and repeated-study condition (see Delaney, Verkoeijen, & Spirgel, 2010; Roediger & Karpicke, 2006; Rohrer & Pashler, 2010, for reviews). The robust benefit of testing on memory occurs with various study materials, including paired associates (e.g., Carrier & Pashler, 1992) and pictures (e.g., Wheeler & Roediger, 1992), in different tests such as free recall (e.g., Karpicke & Roediger, 2007) and cued recall (e.g., Carpenter, Pashler, & Vul, 2006), and in various populations like undergraduates (e.g., Carpenter, 2011) and healthy older adults (e.g., Tse, Balota, & Roediger, 2010). Given the solid evidence for the effectiveness of test-enhanced learning, educators have been advised to test students more often to improve their learning outcomes (e.g., Roediger, Agarwal, McDaniel, & McDermott, 2011). However, few testing-effect studies have taken participants' characteristics into account. As the role of individual differences among students has been a critical issue in classrooms, it is important to test whether the benefit of testing could be generalized to students who have diverse abilities before promoting the use of frequent quizzes in schools. The goal of our study was to test whether the testing effect would be modulated by participants' working-memory capacity (WMC) and trait anxiety for test taking (TA, test anxiety). In a multitrial repeated-study versus repeated testing, paired-associate learning task, Chinese–English bilingual undergraduates with varied TA and WMC acquired half of a set of Swahili–English word pairs via repeated testing and the other half via repeated study and performed a final Swahili cued-recall test after about one week.

Defining TA and WMC

We define WMC as the ability to maintain or process task-relevant information and inhibit task-irrelevant information simultaneously, without testing any specific working-memory model in the current study (see, e.g., Baddeley, 2007; Conway, 2007; Cowan, 2005; Engle & Kane, 2004, for reviews). Participants' WMC was quantified by using Unsworth et al.'s (2005) operation-span task, in which they engaged in online arithmetic processing while maintaining letters in memory for later recall. We use *high-WMC* and *low-WMC* participants to refer to those

This article was published Online First July 9, 2012.

Chi-Shing Tse and Xiaoping Pu, Department of Educational Psychology, The Chinese University of Hong Kong, New Territories, Hong Kong, China.

The work described in this paper was supported by Faculty and Departmental Supportive Fund, The Chinese University of Hong Kong. We thank Jeanette Altarriba and Kit W. Cho for their comments on earlier versions of this paper.

Correspondence concerning this article should be addressed to Chi-Shing Tse, 314 Ho Tim Building, Department of Educational Psychology, The Chinese University of Hong Kong, New Territories, Hong Kong, China. E-mail: cstse@cuhk.edu.hk

who have relatively high WMC and relatively low WMC, respectively, within our samples, rather than referring to high versus low WMC participants in an absolute sense. TA is defined as behavioral and physiological responses that occur in individuals who are concerned about negative outcomes in evaluative situations (Zeidner, 1998). We focus on trait anxiety (i.e., stable disposition to react with excessive worry and task-irrelevant thoughts in evaluative situations, e.g., Spielberger, Anton, & Bedell, 1976), rather than state anxiety (i.e., reaction to a present stressor, e.g., Eysenck, 1992). We quantified TA by using the Chinese version of Spielberger's (1980) 20-item Test Anxiety Inventory (TAI; Yue, 1996), in which participants reported symptoms that they typically experience during the test. We use *high-TA* and *low-TA* participants to refer to those who have relatively high TA and relatively low TA, respectively, within our samples.

High-TA individuals perceive that their performance on a test reflects their competence. This self-focused irrelevant thought could occupy their WMC, which would otherwise have been allocated to the concurrent task (e.g., Klein & Boals, 2001). This may in turn reduce their efficiency in performing the ongoing task, increase their susceptibility to distraction from task-irrelevant materials, and in turn impair their performance (e.g., Ashcraft & Kirk, 2001; Cassady, 2004; Eysenck, Derakshan, Santos, & Calvo, 2007; Hayes, Hirsch, & Mathews, 2008; Keogh & French, 2001; Zeidner, 1998). The effect of TA on task performance is larger when the tasks impose high WMC demands, such as under high pressure or with a concurrent task (e.g., Calvo & Eysenck, 1996; Lee, 1999; Zeidner, 1998), and for individuals with lower WMC (e.g., Johnson & Gronlund, 2009). Conversely, high WMC serves as a buffer against the effect due to anxiety. For example, using a dual-task paradigm (primary short-term memory (STM) task + secondary tone-discrimination task), Johnson and Gronlund showed that trait anxiety was negatively correlated with performance for low-WMC and medium-WMC participants, but not for high-WMC participants.

Potential Relationships Among TA, WMC, and the Testing Effect

During the acquisition phase in the testing-effect paradigm, the preoccupation of worry and irrelevant thought in high-TA participants would make them, relative to low-TA participants, less able to encode word pairs in study trials and more likely to produce intralist intrusion errors (i.e., recalling a wrong English translation to a Swahili word cue) on test trials. This would be especially true for low-WMC participants, who do not have enough attentional resources to counteract the interfering effect of TA. Previous studies did report that low-WMC participants made more intralist intrusion errors than high-WMC participants in a recall task (e.g., Unsworth, Spillers, Brewer, & McMillan, 2011). Hence, high-TA, low-WMC participants would presumably make more intralist intrusion errors than other participants in the acquisition phase.

In the delayed cued-recall test, for pairs acquired via repeated testing, participants need to distinguish their own intralist intrusion errors from the correct answers. Relative to those with low TA, high-TA participants would allocate more WMC toward inhibiting the distraction from their worry and irrelevant thought (e.g., Barrett, Tugade, & Engle, 2004). Whereas high-TA, high-WMC participants only need to allocate *part* of their WMC to overcome the interfering effect of TA in the delayed cued-recall test, high-TA, low-WMC participants might have to allocate *all* of their WMC to do so. This would leave the latter group with fewer attentional resources for distinguishing memory for studied responses versus memory for intralist intrusion errors they produced in the acquisition phase. This problem would occur more frequently in the repeated-testing condition, because participants do not have as many opportunities to produce intralist intrusion errors for pairs in the repeated-study condition. Hence, low-WMC participants would show a larger decrease in performance in the repeated-testing condition as a function of TA than high-WMC participants, leading the high-TA, low-WMC participants to show a smaller testing effect than those who have lower TA and/or higher WMC in the delayed cued-recall test. The relationship between intralist intrusion errors in the acquisition phase and testing effect in the delayed cued-recall test would also be more negative for high-TA, low-WMC participants than those who have lower TA and/or higher WMC.

Present Study and Research Hypotheses

The goal of the current study was to test how the benefit of repeated testing (relative to repeated study) in long-term retention could be modulated by participants' TA and WMC. We manipulated the repeated-study and repeated-testing conditions within subjects so as to yield the testing effect (i.e., the difference in performance between the repeated-testing and repeated-study conditions) for each participant. We predicted an interaction between TA and WMC for intralist intrusion errors in the repeated-testing condition during the acquisition phase, and for the testing effects in the delayed cued-recall test. For low-WMC participants, the intralist intrusion errors would increase as a function of TA, and the testing effects would decrease as a function of TA. In contrast, for high-WMC participants, the intralist intrusion errors and the testing effects would remain constant regardless of their TA level. Moreover, we expected that in the delayed cued-recall test, high-TA, low-WMC participants' testing effects would be negatively correlated with the number of intralist intrusion errors that they produce in the repeated-testing condition in the acquisition phase.

Method

Participants

One hundred sixty (96 female) Chinese–English bilingual undergraduates who reported normal or corrected-to-normal vision participated for monetary compensation (about $6.40 per hour of participation). Their mean age was 20.78 (SD = 1.64).

Materials and Design

Forty Swahili–English word pairs were chosen based on a norming study[1] (see Appendix). The mean word length and log Hyperspace Analogue to Language (HAL) word frequency (Balota et al., 2007; Lund & Burgess, 1996) of English words were 4.83 (SD = 1.32) and 9.29 (SD = 1.32), respectively. They were divided into two sets of 20 and assigned to the repeated-study and repeated-testing conditions, respectively, with this assignment counterbalanced between participants.

Procedure

Personal computers were used to present stimuli and collect data. Participants were tested in groups of one to three in a quiet room and made responses on keyboards. All stimuli appeared in white, lowercase letters in Courier New, 18-point, bold font in a black background at the center of the screen.

Testing-effect task. The main task consisted of acquisition and delayed cued-recall test phases. At the beginning of the acquisition phase, participants were told that they would study a list of Swahili words, together with their English translations, and be tested for their ability to recall the English translations in response to their respective Swahili words in a later cued-recall test. In a study trial, a Swahili–English word pair, with two line spacings inserted in between two words, appeared for 5 s. In a test trial, a Swahili word stayed on the screen until participants typed in their answer for its corresponding English translation. Their typed responses appeared on the screen and they were allowed to use the BACKSPACE key to correct their responses or to skip if they were unable to produce answers. After they finished typing their answers, they pressed the ENTER key to continue. No corrective feedback was provided. A 500-ms blank screen appeared between two trials. There were 12 cycles of 40 study/test trials. In each cycle, each of 40 word pairs appeared either in a study trial or a test trial. Across 12 cycles (S = study trial; T = test trial), the 20 word pairs in the repeated-study condition appeared in the fixed-order S-S-S-T-S-S-S-T-S-S-S-T sequence, and the 20 word pairs in the repeated-testing condition, appeared in the S-T-S-T-S-T-S-T-S-T-S-T sequence (see Table 1).[2] Within each cycle, all items, regardless of being in a study trial or a test trial, were randomly intermixed. As the current study was not designed to test theories of the spacing effect (e.g., Storm, Bjork, & Storm, 2010), we kept constant the lag between every two presentations of study/test trials for pairs in the repeated-study and repeated-testing conditions. This procedure could ensure that the observed testing effect, if any, would not be attributed to the differences in the spacing effect for these two conditions. At the end of each cycle (i.e., after all 40 study/test trials were presented), participants were given a self-paced break before proceeding to the next one. After about a week (M = 7.03 days, SD = .26), participants did a delayed cued-recall test with the procedure being identical to the test trial in the acquisition phase, except that all pairs in the repeated-study and repeated-testing conditions were tested.

Shipley vocabulary test. At the end of the acquisition phase of the testing-effect task, participants completed the 40-item computerized vocabulary subscale test on the Shipley Institute of Living Scale (Shipley, 1940). To control for participants' proficiency in their second language, English, we treated their vocabulary age as a covariate in the regression analyses for the testing effect. The test was reported to have high reliability (Cronbach's α = .87) and strongly correlate with standardized intelligence tests (Zachary, Paulson, & Gorsuch, 1985).

Test Anxiety Inventory (TAI). We adapted the Chinese version of the 20-item TAI (Yue, 1996) to measure TA as a situation-specific trait (i.e., TAI score). On a 4-point scale, participants rated the frequency with which they experienced specific symptoms of anxiety before, during, and after exams. The Cronbach's alpha of our sample (.93) was comparable with those reported in the larger United States samples (Spielberger, 1980). We had half of the participants complete the TAI prior to the acquisition phase, and half after the delayed cued-recall test. According to Zeidner (1991), the effect of test experience could influence TAI ratings, such that participants might have rated their TA to be higher when they completed the TAI after the test than when they did the TAI before the test. To make sure this carryover effect did not distort the pattern of our findings, we counterbalanced the TAI administration time across participants and treated this as a controlling variable in the following analyses. Moreover, for those who received the TAI after the delayed cued-recall test, we attempted to alleviate the effect of test experience on their TAI ratings by inserting several unrelated tasks (Tse, Balota, Yap, Duchek, & McCabe's, 2010, Simon task, Lee, Zhang, & Yin's, 2010, Motivated Strategies for Learning Questionnaire, and Tse & Altarriba's, 2008, Language History Questionnaire) after the delayed cued-recall test and prior to the TAI.

Working-memory task. The automated operation-span task (see Unsworth et al., 2005, for the details of its task structure, scoring, and validity) was given immediately after the Shipley vocabulary test. This task shows good internal consistency (.78) and test–retest reliability (.83). In each trial, participants were presented with a series of letters, each of which was followed by a two-operator arithmetic problem. Across trials, the number of letters for memorization varied randomly from two to seven. At the end of a letter-arithmetic-problem sequence, participants recalled letters in the same order as appeared before. High scores are achieved by holding more letters in memory, while maintaining prespecified accuracy (>85%) in the arithmetic task. We used the

[1] We conducted a norming study on a pool of 175 Swahili–English words from Nelson and Dunlosky (1994), or created by using *http://translate.google.com/* to translate the category exemplars in Van Overschelde, Rawson, and Dunlosky (2004) from English to Swahili. We asked another group of 32 Chinese–English bilingual undergraduates to give a familiarity rating on a 5-point scale to each of 175 English words and an ease-of-learning rating on a 5-point scale to each of 175 word pairs. All English words and Swahili–English word pairs appeared in random order. The rating tasks were blocked and their order was counterbalanced between participants. To ensure that participants understood the English words, and the perceived difficulty of Swahili–English word pairs was low enough to avoid ceiling effects, we chose 40 word pairs with all English words having higher scores than 4.50 (M=4.88, SD=.11) on the familiarity rating, and all pairs having lower scores than 2.30 (M = 1.89, SD = .15) on the ease-of-learning rating.

[2] We did not use the S-T-T-T-S-T-T-T-S-T-T-T sequence, since a pilot study showed that this sequence yielded much lower cumulative proportions of recall in the acquisition phase.

Table 1
The Sequence of Events for the 20 Word Pairs in the Repeated-Study Condition and the 20 Word Pairs in the Repeated-Testing Condition in the 12 Cycles of Study/Test Trials in the Acquisition Phase

Cycle	The 20 word pairs in the repeated-study condition appeared as	The 20 word pairs in the repeated-testing condition appeared as
1st	Study trials	Study trials
2nd	Study trials	Test trials
3rd	Study trials	Study trials
4th	**Test trials**	**Test trials**
5th	Study trials	Study trials
6th	Study trials	Test trials
7th	Study trials	Study trials
8th	**Test trials**	**Test trials**
9th	Study trials	Study trials
10th	Study trials	Test trials
11th	Study trials	Study trials
12th	**Test trials**	**Test trials**

Note. As only in the 4th, 8th and 12th cycles that word pairs from *both* repeated-study and repeated-testing conditions were tested (i.e., as test trials) in the acquisition phase, in order to compare the cumulative proportions of recall between the repeated-study and repeated-testing conditions, we analyzed participants' mean performance in the test trials for these three cycles only.

absolute operation-span scoring method to yield a working-memory score (i.e., *WM score*).

Results

Unless otherwise specified, significance level was set at .05. The stimuli counterbalance scheme for the repeated-study and repeated-testing conditions did not interact with any variable in the analyses (all Fs < 1), so it was not considered further. Regarding TAI administration time, relative to the participants who rated before the acquisition phase, those who did so after the delayed cued-recall test showed numerically higher TAI scores (45.78 versus 43.05, $t(158) = 1.57$), partially consistent with Zeidner (1991), in which participants reported higher TAI scores when they completed the TAI after the test than when they did so prior to the test. Participants' mean vocabulary age was 14.96 ($SD = 1.09$). Their mean WM score was 51.54 ($SD = 14.00$) and was very weakly correlated with their TAI score in the expected direction ($-.11$).

Cumulative Proportions of Word Pairs Recalled in the Acquisition Phase

Figure 1 shows the cumulative proportions of word pairs recalled in the acquisition phase in the repeated-study and repeated-testing conditions. For pairs in the repeated-study condition, participants received test trials in the 4th, 8th, and 12th cycles, whereas for pairs in the repeated-testing condition, they received test trials in the 2nd, 4th, 6th, 8th, 10th, and 12th cycles (see Table 1). Hence, only in the 4th, 8th, and 12th cycles were word pairs from both repeated-study and repeated-testing conditions tested (i.e., as test trials) in the acquisition phase. To compare the cumulative proportions of recall between the repeated-study and repeated-testing conditions, we analyzed participants' mean performance in the test trials for these three cycles only. We conducted a 2 (condition) $\times$ 2 (TAI administration time) $\times$ 3 (cycle: 4th, 8th, or 12th) mixed-factor ANOVA, using a Greenhouse-Geisser correction for the potential violation of sphericity. Condition and cycle were within-subject variables, whereas TAI administration Time was a between-subjects variable. Only the main effect of Cycle and the Cycle $\times$ Condition interaction were significant, $F(1.35, 213.37) = 750.74$, $MSE = .03$, $\eta_p^2 = .83$ and $F(1.88, 297.57) = 4.78$, $MSE = .01$, $\eta_p^2 = .03$. The difference between repeated-study and repeated-testing conditions was significant only in the 4th cycle, .44 versus .42, $t(159) = 2.00$, but not in the 8th, .77 versus .78, $t(159) = 1.49$ or the 12th cycle, .887 versus .894, $t(159) = .90$. The absence of a difference in the last (12th) cycle meant that participants acquired the word pairs in the repeated-study and repeated-testing conditions to the same degree at the end of the acquisition phase.

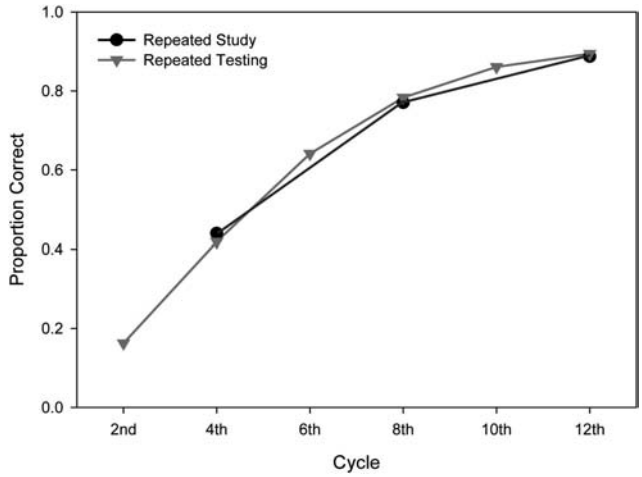

Figure 1. Mean cumulative proportions of Swahili–English word pairs recalled in the repeated-study and repeated-testing conditions across the 12 cycles in the acquisition phase.

Proportion of Word Pairs Recalled in the Delayed Cued-Recall Test

In the delayed cued-recall test, the mean proportions of correctly recalled word pairs in the repeated-study and repeated-testing conditions were .47 and .54, respectively. The 7% difference (i.e., the testing effect) was small yet significant, $F(1, 158) = 65.95$, $MSE = .01$, $\eta_p^2 = .29$.

Regression Analyses on the Testing Effect in the Delayed Cued-Recall Test

We conducted regression analyses to examine how individual differences of WM score and TAI score modulated the benefit of repeated testing in the delayed cued-recall test. The testing effect (i.e., the difference in the proportions of cued-recall performance in the repeated-testing vs. repeated-study conditions) was the dependent variable. In the first step of regression models, we entered participants' sex, age, vocabulary age, TAI administration time, and cumulative proportions of recall in the repeated-study and repeated-testing conditions in the acquisition phase. Vocabulary age was entered to control for individual differences in English proficiency among Chinese–English bilingual participants. Cumulative proportions of recall were entered to control for the extent to which participants learned the word pairs in the acquisition phase.[3] In the second step, we entered mean-centered TAI score and WM score. In the third step, we entered the multiple of mean-centered TAI score and WM score (i.e., the interaction term). There was no problem of multicollinearity, as indicated by a low variance–inflation ratio (<1.16).

Table 2 summarizes the findings of multiple regression analyses. The TAI Score × WM Score interaction significantly predicted the testing effect in the delayed cued-recall test. We then divided participants into high versus low WMC groups by median split ($N = 80$ each), based on their WM scores (high: $M = 62.78$, $SD = 6.26$; low: $M = 40.31$, $SD = 10.01$) and performed multiple regression analyses on each group. The TAI scores were statistically equivalent for low WMC group ($M = 44.64$, $SD = 10.62$) versus high WMC group ($M = 44.19$, $SD = 11.52$), $t(158) = .26$. Given that participants' WMC was quantified by their performance on the operation-span test, rather than directly manipulated (e.g., performing a secondary task), in the following analyses we compared participants who have relatively high WMC versus those who have relatively low WMC within our samples, rather than the performance of high versus low WMC participants in an absolute sense. In these two-step analyses, the first step was the same as the one mentioned above and the second step was the mean-centered TAI score. The simple main effect of TAI score was significant for low-WMC participants, $\beta = -.47$, $t(72) = 4.37$, but not for high-WMC participants, $\beta = -.16$, $t(72) = 1.38$.[4] Figure 2 shows the proportion correct in the repeated-study and repeated-testing conditions, and the testing effect as a function of TAI scores for low-WMC participants versus high-WMC participants. TA modulated the testing effect for low-WMC participants, with those who had higher TA showing smaller testing effects, but not for high-WMC participants. The decrease of the testing effect as a function of TAI scores for low-WMC participants could be attributed to a larger decrease in their performance in the repeated-testing condition than in the repeated-study condition. This was supported by regression analyses done on proportions correct in the repeated-study and repeated-testing conditions in the delayed cued-recall test. In the repeated-testing condition, the proportion correct decreased as a function of TAI scores for low-WMC participants, $\beta = -.18$, $t(72) = 2.09$, but not for high-WMC participants, $\beta = .04$, $t(72) = .43$. In the repeated-study condition, the proportion correct did not change with TAI scores for low-WMC, $\beta = -.06$, $t(72) = .62$ or high-WMC participants, $\beta = .12$, $t(72) = 1.17$.[5]

[3] To examine whether participants' performance in the acquisition phase depended on TA and WMC, we conducted regression analyses on the cumulative proportions of recall in the repeated-study and repeated-testing conditions in the acquisition phase by entering participants' sex, age, vocabulary age, and TAI administration time in the first step, mean-centered TAI score and WM score in the second step, and their interaction term in the third step. The WM Score × TAI score interaction was significant in the repeated-study condition, $\beta = .18$, $t(152) = 2.31$, but marginally so in the repeated-testing condition, $\beta = .14$, $t(152) = 1.82$, $p = .07$. Follow-up analyses showed that in both repeated-study and repeated-testing conditions, the simple main effect of TAI score was marginally significant for low-WMC participants, $\beta = -.20$, $t(71) = 1.68$, $p = .10$ and $\beta = -.20$, $t(71) = 1.70$, $p = .09$, but not for high-WMC participants, $\beta = -.04$, $t(71) = .32$ and $\beta = -.07$, $t(71) = .62$. As low-WMC participants, but not high-WMC participants, showed a trend towards a lower cumulative proportion as a function of TA in the acquisition phase, we covaried out the cumulative proportions of recall in the repeated-study and repeated-testing conditions in the regression analyses of delayed cued-recall performance. Following a reviewer's suggestion, we also re-analyzed the delayed cued-recall data in the regression models by including only the pairs that had been correctly recalled in the acquisition phase. The overall pattern of findings remained unchanged, so only the unconditionalized data are reported here.

[4] We also divided participants into three WMC groups, low ($N=53$) vs. mid ($N=54$) vs. high ($N=53$). By using the performance of mid-WMC participants as a reference, we could test whether the negative relationship between TA and testing effect could be strengthened by lower WMC (relative to mid WMC), alleviated by higher WMC (relative to mid WMC), or both. We performed similar regression analyses for all three WMC groups (as were done when there were two groups). The simple main effect of TAI score was significant for low-WMC and mid-WMC participants, $\beta = -.55$, $t(45) = 4.06$ and $\beta = -.38$, $t(46) = 2.68$, respectively, but not for high-WMC participants, $\beta = .10$, $t(45) = .64$. As indicated by standardized regression coefficients, relative to mid-WMC participants ($-.38$), the negative relationship between TA and the testing effect was stronger for low-WMC participants ($-.55$), but high-WMC participants showed a numerically reversed (i.e., positive) relationship between TA and the testing effect (.10). Thus, an increase in WMC may eliminate the negative relationship between TA and the testing effect, whereas a decrease in WMC may exacerbate the negative relationship between TA and the testing effect.

[5] Apart from controlling for the effect of TAI administration time in regression analyses (see Table 2), we also conducted separate regression analyses for participants who received the TAI at the beginning and those who received the TAI at the end of the experiment. The overall patterns of findings were highly similar for both groups, indicating that the conclusion reported in the main text held, regardless of TAI administration time.

Table 2
Standardized Regression Coefficients in the Full Model of Regression Analyses of Test Anxiety Inventory (TAI) Score and Working Memory (WM) Score on Predicting the Testing Effect (i.e., Subtracting the Proportion of Cued-Recall Performance in the Repeated-Study Condition From the Proportion of Cued-Recall Performance in the Repeated-Testing Condition) and the Proportions of Intralist Intrusion Errors in the Repeated-Study and Repeated-Testing Conditions in the Acquisition Phase

				Proportion of intralist intrusion errors in the					
	Testing effect			Repeated-study condition			Repeated-testing condition		
Variable	β	*t*(152)	*p*	β	*t*(152)	*p*	β	*t*(152)	*p*
Sex	−0.04	0.45	0.66	0.10	1.35	0.18	0.03	0.44	0.66
Age	0.05	0.69	0.49	0.04	0.48	0.63	0.06	0.72	0.47
Vocabulary Age	−0.08	0.97	0.34	0.17	2.12*	<0.05	0.14	1.71	0.09
TAI Administration Time	0.06	0.73	0.47	0.04	0.53	0.60	0.04	0.57	0.57
Cumulative-Repeated-study	−0.08	0.55	0.58	—	—	—	—	—	—
Cumulative-Repeated-testing	−0.17	1.14	0.26	—	—	—	—	—	—
Main Effect of TAI score	**−0.30**	**−3.75***	<0.01	0.07	0.92	0.36	0.09	1.11	0.27
Main Effect of WM score	0.15	1.81	0.07	**−0.25**	**−3.17***	<0.01	**−0.29**	**−3.74***	<0.01
TAI score × WM score Interaction	**0.21**	**2.69***	<0.01	−0.15	−1.88	0.06	**−0.16**	**−2.07***	<0.05

Note. The full models from left to right were all significant, $F(9, 150) = 3.00$, $MSE = .01$, $p < .01$, $R^2 = .15$; $F(7, 152) = 2.50$, $MSE = .02$, $p = .02$, $R^2 = .11$; and $F(7, 152) = 2.91$, $MSE = .02$, $p = .01$; $R^2 = .12$.
* $p < .05$ (two-tailed).

Regression Analyses on the Proportion of Intrusion Errors in the Acquisition Phase

To test the relationship between intrusion errors and the testing effect, we examined participants' mean proportions of intralist intrusion errors in the repeated-study and repeated-testing conditions in the acquisition phase in regression analyses. We entered participants' sex, age, vocabulary age, and TAI administration time in the first step, the mean-centered TAI score and WM score in the second step and their interaction term in the third step. The WM score × TAI score interaction was significant in the repeated-testing condition, but only marginally so in the repeated-study condition (see Table 2). Follow-up analyses showed that in the repeated-testing condition, the simple main effect of TAI score was significant for low-WMC participants, $\beta = .24$, $t(71) = 2.00$, but not for high-WMC participants, $\beta = -.04$, $t(71) = .34$, but in the repeated-study condition, the simple main effects of TAI score were not significant for low-WMC, $\beta = .18$, $t(71) = 1.53$ or high-WMC participants, $\beta = -.03$, $t(71) = .23$. In the repeated-testing condition, low-WMC participants with higher TA made more intralist intrusion errors than those with lower TA, but for high-WMC participants, there was no relationship between TA and the proportion of intralist intrusion errors.[6] Nor was there any relationship between proportion of intralist intrusion errors and TA in the repeated-study condition, regardless of participants' WMC. Similar analyses were conducted for extralist intrusion errors, but none of the predictors of WM score, TAI score, or their interaction term was significant.

Correlations Between the Proportion of Intralist Intrusion Errors in the Acquisition Phase and the Testing Effect in the Delayed Cued-Recall Test

We further tested the relationship between intralist intrusion errors in the repeated-testing condition in the acquisition phase and the testing effect in the delayed cued-recall test, after dividing our participants into four groups ($N = 40$ each): high WMC with high TA, high WMC with low TA, low WMC with high TA, and low WMC with low TA. For each of these groups, we computed a correlation between participants' intralist intrusion errors in the repeated-testing condition and their testing effect in the delayed cued-recall test, after partialing out their sex, age, vocabulary age, TAI administration time, and the intralist intrusion errors in the repeated-study condition. The correlation between the intralist intrusion errors and the testing effects was only significant for low-WMC/high-TA group (−.37, $p < .05$), but not for high-WMC/high-TA group (−.19), high-WMC/low-TA group (−.27), or low-WMC/low-TA group (−.29).

Discussion

The goal of our study was to investigate whether the advantage of repeated testing (as compared with repeated study) could be modulated by one's trait test anxiety (TA) and working-memory capacities (WMC). We had participants with varied TA and WMC acquire half of a set of Swahili–English word pairs via repeated study, and half via repeated testing in the acquisition phase, and after about a week, perform a delayed cued-recall test for all pairs. We replicated the typical testing effect in the delayed cued-recall test: Participants showed better performance for pairs learned via repeated testing than for those learned via repeated study. We also identified two individual difference markers (TA and WMC) that modulated the testing effect. In the acquisition phase, whereas low-WMC participants with higher TA made more intralist intrusion errors than those with lower TA, high-WMC participants'

[6] Although one could attribute the intralist intrusion errors to participants' guessing, the pattern of intralist intrusion errors in the regression analyses remained the same after taking participants' extralist intrusion errors into account. Therefore, the findings of intralist intrusion errors could not be explained by the use of guessing strategy, per se.

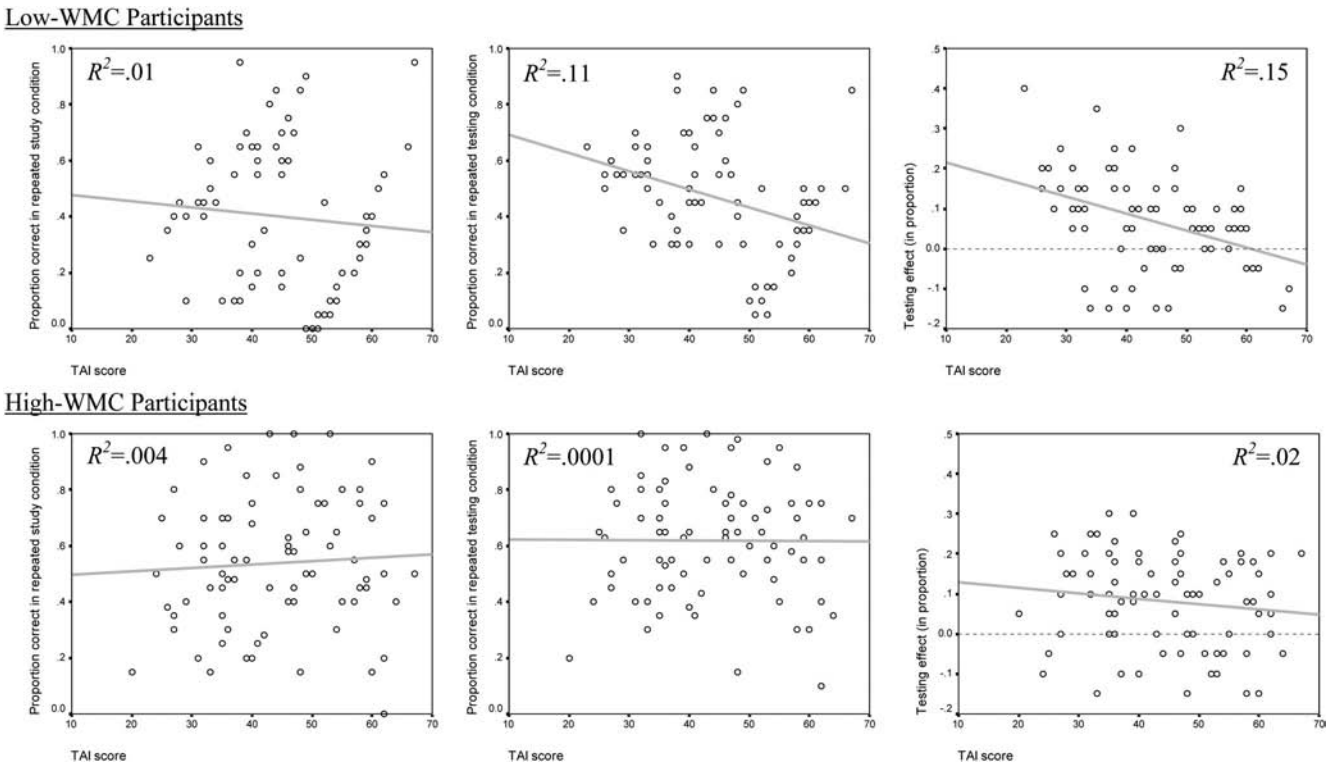

Figure 2. Scatterplots of participants with high vs. low WMC for the relationships for TAI scores vs. proportion correct in repeated-study condition (left panel), TAI scores vs. proportion correct in repeated-testing condition (middle panel), and TAI scores vs. testing effects (right panel).

intralist intrusion errors did not vary with their TA level. In the delayed cued-recall test, the testing effect decreased as a function of TA for low-WMC participants, but was not related to TA level for high-WMC participants.[7]

The current findings were consistent with those reported in Hinze and Rapp (2011), who investigated how instruction-induced state anxiety modulated the benefit of testing on memory. They used science texts as study materials and multiple-choice and open-ended application questions in their delayed-memory test. In the acquisition phase, they manipulated participants' state anxiety by inducing performance pressure via offering them bonus money for their high test performance. Those who received this bonus instruction showed higher state anxiety in the acquisition phase. In the delayed test, these participants showed lower proportion recall and higher forgetting rates (relative to their initial-test performance in the acquisition phase) than those who did not receive the bonus instruction. Hence, performance pressure boosted state anxiety, and in turn minimized the benefit of testing in the delayed test. Despite the procedural differences between the two studies (e.g., trait anxiety vs. state anxiety; manipulating repeated-study and repeated-testing conditions within vs. between subjects), the current findings were in line with Hinze and Rapp's findings and further showed that individuals with lower WMC and higher TA could be less aided by repeated testing (relative to repeated study). Before elaborating on the theoretical and educational implications, we first delineate some potential limitations of our study that could generate future research questions.

Potential Limitations of the Present Study

First, similar to other studies that involved correlational/regression analyses, it is not easy to infer a causal relationship between the factors examined within the current study. TA might have a negative impact on the benefit of repeated testing for

[7] The scores estimated by Spielberger's (1980) TAI can be divided into two components: emotionality and worry, which are differentiated by their temporal patterns and impact on academic performance (Zeidner, 1998). Emotionality is related to physiological symptoms that stem from arousal of the autonomic nervous system, and worry is related to debilitating thoughts and concerns that students have about evaluative situations. The negative relationship between anxiety and test performance was reported to be weaker for emotionality scores than for worry scores (e.g., Liebert & Morris, 1967). After worry scores were controlled, emotionality scores were no longer correlated with test performance (Cassady, 2004; Cassady & Johnson, 2002). We computed emotionality and worry scores in our dataset and found that the results for these two indices were similar to those for the overall TAI scores. Regression analyses with mean-centered TAI scores replaced by mean-centered worry or emotionality scores showed that the WM Score × Emotionality Score interaction was significant after partialing out worry score, and the WM Score × Worry Score interaction was significant after partialing out emotionality score. However, these findings should be interpreted with caution because the factor structures of emotionality and worry components were not always clear cut in the TAI, with some of the items reflecting a mixture of both dimensions (e.g., Zeidner, 1998).

low-WMC participants, but it could also be that the relationship between anxiety and performance is reciprocal. High levels of TA produce some aversive patterns of motivation and task strategies that interfere with learning. The fact that performance suffers may lead to further anxiety over time and generate a vicious cycle of increasing anxiety and degrading performance (e.g., Wells & Matthews, 1994), especially for participants with lower WMC. More research should be done to test the causal relationship between anxiety and performance (e.g., by embedding the testing-effect paradigm in TA intervention programs).

Second, we focused on trait anxiety, a personality trait whose effect is generalized to daily life-evaluative situations (e.g., exam) as reflected by TAI scores, rather than state anxiety, a mood state that occurs naturally, but can also be induced temporarily using an induction task (e.g., Brodish & Devine, 2009). State anxiety, which has also been reported to affect WMC (e.g., Eysenck, 1992), could be experienced in an evaluation situation when the nature of a person's vulnerability (high evaluative trait anxiety) is congruent with the nature of the situation (evaluative; e.g., Zeidner & Matthews, 2005). Both situational factors and individual differences in test anxiety may modulate the occurrence of interfering thoughts. Future studies may consider measuring participants' state anxiety immediately after the acquisition phase and take that into account when evaluating performance on a subsequent test. State anxiety could even be directly manipulated to test whether the anxiety induced right *before* the acquisition phase could modulate the benefit of repeated testing. By measuring naturally occurring state anxiety or inducing state anxiety at different time points (e.g., before the acquisition phase or before the delayed-memory test), one could investigate whether the testing effect would be weakened when test anxiety acts on the encoding stage (the acquisition phase), the retrieval stage (the delayed-memory test), or both.

Third, we kept constant the lag between every two presentations of study/test trials for pairs in the repeated-study and repeated-testing conditions. This ensured that the testing effect we observed might not be differentially influenced by the difference in spacing between these two conditions, as a positive effect of spaced retrieval on memory has been well-documented in the literature (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Logan & Balota, 2008; Maddox, Balota, Coane, & Duchek, 2011; Roediger & Karpicke, 2011). Future research should orthogonally manipulate the lag between the study versus test trials and the repeated-study versus repeated-testing conditions to test (a) whether the testing effect would be boosted when a more optimal interval between the two study/test trials is used, and (b) whether the magnitude of this boost would also depend on participants' levels of TA and WMC.

Theoretical Implications of the Current Findings

Consistent with our hypotheses, we found a significant WM Score × TAI Score interaction on the testing effects in the delayed cued-recall test. There was also a significant relationship between the intralist intrusion errors in the repeated-testing condition in the acquisition phase, and the testing effect in the delayed cued-recall test for high-TA, low-WMC participants. Because these participants' WMs were likely preoccupied by a high level of irrelevant thought and worry, they did not have sufficient WMC left to resolve the source confusion between correct responses (as acquired from study trials) and incorrect responses (as produced by these participants during the acquisition phase) corresponding to Swahili word cues in the delayed cued-recall test. Therefore, relative to those with lower TA and/or higher WMC, high-TA, low-WMC participants demonstrated lower performance in the repeated-testing condition. As the group difference in the repeated-study condition was not as large as in the repeated-testing condition (see Figure 2), the high-TA, low-WMC participants showed a smaller testing effect than the other participants. This account can explain how the testing effect was modulated by TA and WMC. In the following discussion, we also consider how our findings could be accommodated by other accounts of testing effects that focus on why retrieval practice in the acquisition phase would trigger the testing effect in the delayed cued-recall test.

According to the encoding-variability account (e.g., McDaniel & Masson, 1985; Szpunar, McDermott, & Roediger, 2008), an intervening test, but not an extra study opportunity, increases the number of retrieval cues (or contextual elements) encoded with an item's memory trace. This allows more new routes to access the encoded association between Swahili and English words and provides tested items with a mnemonic advantage over restudied items in a delayed memory test. Carpenter's (2009) elaborative retrieval account suggests that retrieval involves a search in memory for a specific target that activates a network of related concepts. The generation of this elaborative structure provides multiple retrieval routes to the tested items, which can facilitate their retrieval from memory during subsequent tests. In contrast, as an item is directly available in a restudy trial, participants are less likely to generate such an elaborative structure. Thus, in a subsequent memory test, tested items are better remembered than restudied items. Specifying the type of information activated during retrieval practice in the acquisition phase that is helpful for subsequent retention, Pyc and Rawson's (2010) mediator-effectiveness hypothesis (see also Carpenter, 2011) postulates that relative to restudying, testing can strengthen the link between a cue and a target via spontaneously activated mediating information (e.g., *wing* for the word pair, *wingu–cloud*). In the delayed-memory test, this mediator can be used to help in retrieving the English target in response to the Swahili cue (*wingu* → *wing* → *cloud*).

All these accounts could accommodate the current finding that the relationship between TA and the testing effect was mediated by WMC. Previous research showed that TA influences memory performance via affecting the acquisition (encoding), organization rehearsal (study skills), and retrieval for a given test (e.g., Naveh-Benjamin, 1991). Compared with low-TA participants, high-TA participants are less likely to perform elaborative rehearsal and so they engage in more restricted encoding (e.g., Mueller, Carlomusto, & Marler, 1978). This could be especially true for those who had low-WMC in the current study, because most of their WMC might have already been allocated toward inhibiting the distraction from their worries during the tests. Compared with those who have higher WMC and/or lower TA, high-TA, low-WMC participants were less able to spontaneously come up with mediators, develop retrieval cues, and establish elaborative structures when they acquired Swahili–English word pairs in the repeated-testing condition during the acquisition phase. Their delayed cued-recall performance was then less benefited by repeated

testing, and so these participants demonstrated a smaller testing effect.

Could these cue/mediator accounts also explain why high-TA, low-WMC participants' intralist intrusion errors in the repeated-testing condition were negatively correlated with their testing effect in the delayed cued-recall test? During the acquisition phase, high-TA, low-WMC participants produced more intralist intrusion errors, such that they would be more likely to develop retrieval cues/mediators to link up the cues with the wrong targets in the repeated-testing condition. In the delayed cued-recall test, these wrong cues/mediators would lead the participants to encounter source confusion (i.e., a failure to distinguish between the correct and wrong targets) and to fail to recall as many correct targets in the repeated-testing condition as participants with higher WMC and/or lower TA. As a result, these high-TA, low-WMC participants would show a smaller testing effect. However, this proposal is not consistent with the above explanation for the relationships between TA, WMC, and testing effect. Specifically, even if high-TA, low-WMC participants produced more intralist intrusion errors in the test trials during the acquisition phase, they would presumably not be able to develop retrieval cues/mediators for the wrong targets due to their having insufficient WMC to tackle the task. Hence, they would presumably show no correlation between intralist intrusion errors in the acquisition phase and the testing effect in the delayed cued-recall test, contradicting the current results. To be fair, these testing-effect accounts were not originally developed to explain the modulating roles of TA and WMC in producing the testing effect. More research should be done to examine how these accounts could be expanded to accommodate individual differences in testing effects.

Apart from testing-effect theories, it is worth noting the connection between the current findings and research on the interplay of emotion and attentional control, such as the modulating role of WMC on the negative effect of stereotype threat on performance. Stereotype threat refers to a performance decline in a task due to the fear of confirming an existing negative stereotype about one's social, gender, or ethnic group (cf. Steele & Aronson, 1995, e.g., women in math). Beilock, Rydell, and McConnell (2007) proposed that stereotype threat creates a state of imbalance between one's concept of self (e.g., being a woman) and one's expectation of success (e.g., excel in math) that interferes with WM. When a task demands heavy WMC, those who are induced with stereotype threat would show worse performance than those who are not. The current study demonstrated that low-WMC participants showed a stronger negative relationship between TA and the testing effect than high-WMC participants. However, unlike Beilock et al., we did not manipulate TA (e.g., using high-stake vs. low-stake tests, see Hinze & Rapp, 2011) or WM demand in the task (e.g., acquiring word pairs under divided vs. full attention). Future studies should manipulate these variables to test the causal relationships between WMC, TA, and the benefit of testing on memory performance, so as to shed more light on the interaction between affective and memory processing.

Conclusion and Educational Implications of the Current Findings

By examining individual differences in undergraduate participants' TA and WMC in acquiring Swahili–English word pairs in a testing-effect paradigm, we found that the testing effect decreased as a function of TA for low-WMC participants, but was not associated with TA level for high-WMC participants. To our knowledge, we are among the first to show a boundary condition for using repeated testing to enhance long-term retention for young adults (see Hinze & Rapp, 2011, for another example and Tse, Balota, & Roediger, 2010, for a boundary condition for healthy older adults). Regarding educational implications, although the effectiveness of test-enhanced learning has been reported in classroom settings (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger et al., 2011), few researchers have taken potential individual differences and the benefit of testing into account. The current research showed that the benefit of repeated testing was very small, if not absent, in undergraduate participants with lower WMC and higher TA. This conceptually echoes previous findings that TA was negatively correlated with students' academic performance (e.g., Seipp, 1991). Even when students with high TA possess efficient study skills, they still suffer from anxiety blockage, fail to handle stress in evaluative situations, and thus find it difficult to retrieve relevant information during an exam.

It is noteworthy that the current task was a "low-stakes" test, in that we did not impose any evaluative pressure, nor did we provide feedback to our participants regarding their performances (e.g., Cassady, 2004), nor did monetary compensation depend on their performances (e.g., Hinze & Rapp). That is, the current paradigm has already biased us against observing any relationships between TA and the testing effect. The negative correlation between TA and the testing effect that we observed in low-WMC undergraduates in a laboratory might already underestimate what would occur for students in classrooms, in which they are often given high-stakes tests under high-evaluation pressure. Thus, when promoting the use of frequent testing in the classrooms, teachers should recognize that the benefit of repeated testing may depend on students' working memory and the extent to which they are anxious about test taking. To boost the benefit of testing for students with high TA, teachers should first provide them with effective intervention programs (e.g., Nelson & Knight, 2010; Orbach, Lindsay, & Gray, 2007; Ramirez & Beilock, 2011) before having them perform more tests and quizzes. To generalize the current findings, future research should test the effectiveness of test-enhanced learning in classroom settings by also taking into account the role of individual differences in cognitive abilities and personality traits among participating students.

References

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130,* 224–237. doi:10.1037/0096-3445.130.2.224

Baddeley, A. D. (2007). *Working memory, thought and action.* Oxford, UK: Oxford University Press. doi:10.1093/acprof:oso/9780198528012.001.0001

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging, 21,* 19–31. doi:10.1037/0882-7974.21.1.19

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130,* 553–573. doi:10.1037/0033-2909.130.4.553

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spill over. *Journal of Experimental Psychology: General, 136,* 256–276. doi:10.1037/0096-3445.136.2.256

Brodish, A. B., & Devine, P. G. (2009). The role of performance-avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology, 45,* 180–185. doi:10.1016/j.jesp.2008.08.005

Calvo, M. G., & Eysenck, M. W. (1996). Phonological working memory and reading in test anxiety. *Memory, 4,* 289–306. doi:10.1080/096582196388960

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. doi:10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1547–1552. doi:10.1037/a0024140

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. doi:10.3758/BF03194004

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14,* 569–592. doi:10.1016/j.learninstruc.2004.09.002

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27,* 270–295. doi:10.1006/ceps.2001.1094

Conway, A. R. A. (2007). *Variation in working memory.* New York, NY: Oxford University Press.

Cowan, N. (2005). *Working memory capacity.* New York, NY: Psychology Press. doi:10.4324/9780203342398

Delaney, P. F., Verkoeijen, P. P., & 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,* 64–111. doi:10.1016/S0079-7421(10)53003-2

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* (pp. 145–199). New York, NY: Academic Press. doi:10.1016/S0079-7421(03)44005-X

Eysenck, M. W. (1992). *Anxiety: The cognitive perspective.* Hillsdale, NJ: Erlbaum.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7,* 336–353. doi:10.1037/1528-3542.7.2.336

Hayes, S., Hirsch, C., & Mathews, A. (2008). Restriction of working memory capacity during worry. *Journal of Abnormal Psychology, 117,* 712–717. doi:10.1037/a0012908

Hinze, S. R., & Rapp, D. N. (2011, May). *How does test anxiety influence testing effects?* Paper presented at the Eighty-Third Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Johnson, D. R., & Gronlund, S. D. (2009). Individuals lower in working memory capacity are particularly vulnerable to anxiety's disruptive effect on performance. *Anxiety, Stress, & Coping, 22,* 201–213. doi:10.1080/10615800802291277

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Keogh, E., & French, C. C. (2001). Test anxiety, evaluative stress, and susceptibility to distraction from threat. *European Journal of Personality, 15,* 123–141. doi:10.1002/per.400

Klein, K., & Boals, A. (2001). The relationship of life event stress and working memory capacity. *Applied Cognitive Psychology, 15,* 565–579. doi:10.1002/acp.727

Lee, J. C.-K., Zhang, Z., & Yin, H. (2010). Using multidimensional Rasch analysis to validate the Chinese version of the Motivated Strategies for Learning Questionnaire (MSLQ-CV). *European Journal of Psychology of Education, 25,* 141–155. doi:10.1007/s10212-009-0009-6

Lee, J. H. (1999). Test anxiety and working memory. *Journal of Experimental Education, 67,* 218–240. doi:10.1080/00220979909598354

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20,* 975–978. doi:10.2466/pr0.1967.20.3.975

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal spaced retrieval practice in healthy young and older adults. *Aging, Cognition, and Neuropsychology, 15,* 257–280. doi:10.1080/13825580701322171

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, 28,* 203–208.

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging, 26,* 661–670. doi:10.1037/a0022942

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414. doi:10.1037/a0021782

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. doi:10.1037/0278-7393.11.2.371

Mueller, J. H., Carlomusto, M., & Marler, M. (1978). Recall and organization in memory as a function of rate of presentation and individual differences in test anxiety. *Bulletin of the Psychonomic Society, 12,* 133–136.

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology, 83,* 134–139. doi:10.1037/0022-0663.83.1.134

Nelson, D. W., & Knight, A. E. (2010). The power of positive recollections: Reducing test anxiety and enhancing college student efficacy and performance. *Journal of Applied Social Psychology, 40,* 732–745. doi:10.1111/j.1559-1816.2010.00595.x

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

Orbach, G., Lindsay, S., & Grey, S. (2007). A randomised placebo-controlled trial of a self-help Internet-based intervention for test anxiety. *Behaviour Research and Therapy, 45,* 483–496. doi:10.1016/j.brat.2006.04.002

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science, 331,* 211–213. doi:10.1126/science.1199427

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17,* 382–395. doi:10.1037/a0026252

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

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), Successful remembering and successful forgetting: Essays in honor of Robert A. Bjork. (pp. 23–48). New York, NY: Psychology Press.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39,* 406–412. doi:10.3102/0013189X10374770

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4,* 27–41. doi:10.1080/08917779108248762

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

Spielberger, C. D. (1980). *Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Anton, W., & Bedell, J. (1976). The nature and treatment of test anxiety. In M. Zuckerman & C. D. Spielberger (Eds.), *Emotion and anxiety: New concepts, methods, and applications* (pp. 317–345). Hillsdale, NJ: Erlbaum.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test-performance of African-Americans. *Journal of Personality and Social Psychology, 69,* 797–811. doi:10.1037/0022-3514.69.5.797

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38,* 244–253. doi:10.3758/MC.38.2.244

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399. doi:10.1037/a0013082

Tse, C.-S., & Altarriba, J. (2008). Evidence against linguistic relativity in Chinese and English: A case study of spatial and temporal metaphors. *Journal of Cognition and Culture, 8,* 335–357. doi:10.1163/156853708X358218

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,* 833–845. doi:10.1037/a0019933

Tse, C.-S., Balota, D. A., Yap, M. J., Duchek, J. M., & McCabe, D. P. (2010). Effects of healthy aging and early-stage dementia of the Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology, 24,* 300–315. doi:10.1037/a0018274

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. doi:10.3758/BF03192720

Unsworth, N., Spillers, G. J., Brewer, G. A., & McMillan, B. (2011). Attention control and the antisaccade task: A response time distribution analysis. *Acta Psychologica, 137,* 90–100. doi:10.1016/j.actpsy.2011.03.004

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An update and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289–335. doi:10.1016/j.jml.2003.10.003

Wells, A., & Matthews, G. (1994). *Attention and emotion: A clinical perspective.* Hove, UK: Lawrence Erlbaum.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245. doi:10.1111/j.1467-9280.1992.tb00036.x

Yue, X. D. (1996). Test anxiety and self-efficacy: Levels and relationship among secondary school students in Hong Kong. *Psychologia: An International Journal of Psychology in the Orient, 39,* 193–202.

Zachary, R. A., Paulson, M. J., & Gorsuch, R. L. (1985). Estimating WAIS IQ from the Shipley Institute of Living Scale using continuously adjusted age norms. *Journal of Clinical Psychology, 41,* 820–831. doi:10.1002/1097-4679(198511)41:6<820::AID-JCLP2270410616>3.0.CO;2-X

Zeidner, M. (1991). Test anxiety and aptitude test performance in an actual college admission testing situation: Temporal considerations. *Personality and Individual Differences, 12,* 101–109. doi:10.1016/0191-8869(91)90092-P

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York, NY: Plenum Press.

Zeidner, M., & Matthews, G. (2005). Evaluation anxiety. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141–163). New York, NY: Guilford Press.

(*Appendix follows*)

Appendix

Experimental Stimuli

Swahili	English	Swahili	English
ndizi	banana	pombe	beer
kitabu	book	mashua	boat
fagio	broom	ndugu	brother
ndoo	bucket	siagi	butter
nafaka	corn	zulia	carpet
kaa	crab	ngombe	cow
tabibu	doctor	pazia	curtain
yai	egg	mbwa	dog
bustani	garden	jicho	eye
bunduki	gun	chakula	food
farasi	horse	kaburi	grave
ziwa	lake	kofia	hat
tumbili	monkey	godoro	mattress
chungwa	orange	kipanya	mouse
sumu	poison	kitunguu	onion
viazi	potato	nguruwe	pig
mwamba	rock	malkia	queen
kiatu	shoe	hariri	silk
theluji	snow	nyanya	tomato
dirisha	window	ukuta	wall

Received October 16, 2011
Revision received April 20, 2012
Accepted May 22, 2012 ■