RICE UNIVERSITY

THE POSITIVE AND NEGATIVE EFFECTS OF TESTING IN LIFELONG LEARNING

by

Ashley N. D. Meyer

A DISSERTATION SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED DISSERTATION COMMITTEE:

Jessica M. Logan (Committee Chair)
Assistant Professor of Psychology

Margaret E. Beier
Associate Professor of Psychology

James L. Dannemiller
Lynette S. Autrey Professor of Psychology
Chair, Psychology

Nicoletta Orlandi
Assistant Professor of Philosophy

HOUSTON, TX
MAY 2011

Acknowledgements

The completion of this dissertation was made not only possible, but much more enjoyable and less stressful, because of the contributions of multiple people: my advisor, my lab-mates, my committee members, one of our undergraduate research assistants, and my family.

My advisor, Jes Logan, has greatly contributed to the creation of this dissertation. Even before I planned this project, she taught me all about memory and aging and about different learning techniques that affect memory. In addition, her enthusiasm for the topic, spread to me. That knowledge and enthusiasm was the foundation for this project. Additionally, Jes helped guide me throughout the entire dissertation process by helping me to brainstorm about everything from initial topic choice to how to best present and test my ideas. Her most important contribution, however, was her ever-present support and encouragement. This definitely made completing my dissertation less of an assignment and more of an enjoyable journey.

My lab-mates, Sara Haber Halcomb and Gunes Avci, have also been important to me throughout the dissertation process. They have listened to various iterations of my project from initial conception to presentation of the final project and have provided me with much useful feedback and support. Additionally, they have provided valuable diversion on many occasions.

My committee members: Margaret Beier, Jim Dannemiller, and Nico Orlandi, have also helped greatly by contributing their expertise and various points of view to the project. For example, all three committee members helped by making valuable suggestions at the proposal phase of the project that led to a better final project.

Additionally, both Margaret and Jim provided me with important tips about testing and analyzing data from participants of various ages using multiple regression as a statistical technique.

An undergraduate laboratory assistant in our laboratory, Elaine Mahoney, was also an immense help in completing this project. She helped me by collecting a third of the data and by providing me with insightful feedback at the early stages of the project. Her attention to detail certainly helped to catch some errors early on in the project and improved the quality of the work overall.

Lastly, I am extremely appreciative of my family for their love, support, and patience throughout not only my dissertation, but my entire graduate education. I am especially gratefully for my husband's patience and support, as I have spent many nights and weekends working hard on this project. Without all of my family, I may not have pushed myself as hard or have persevered as much when I ran into difficult patches of the dissertation.

So, again, thank you all. If dissertations had co-authors, I would include all of you for your great contributions!

Table of Contents

List of Tables

List of Figures

Abstract

Formal classroom learning is a lifelong pursuit. Many older adults return to school to advance their careers, learn new skills, or simply for personal fulfillment. As such, methods for improving learning should be considered in relation to both younger and older learners in order to properly assess their ultimate usefulness. A technique that has been demonstrably effective at improving learning and memory in younger students is testing. Testing improves memory more than mere exposure to material (e.g., restudying), a benefit known as the positive testing effect. However, recognition tests, where learners are exposed to correct *and* incorrect information (e.g., multiple-choice tests), also introduce false information to test-takers. While evidence shows that testing improves memory for tested material, this can include the *incorrect* material presented on recognition tests manifested as increased reproduction of incorrect answers (lures), a phenomenon known as the negative testing effect. These effects of testing, however, have only been studied in younger learners. Older learners, on the other hand, may show decreased positive testing effects and increased negative testing effects because of poorer long-term episodic and source memory, perhaps making them less receptive to the positive effects of testing and more susceptible to the negative effects of testing. Therefore, this study examined the positive and negative effects of testing on learning in 60 younger university students aged 18-25, 60 younger community adults aged 18-25, and 60 older community adults aged 55-65. This research also scrutinized how individual differences, including intelligence, previous knowledge, initial performance, and source memory were related to the positive and negative effects of testing. All groups showed positive testing effects, but these were larger for younger adults, for individuals with higher initial performance, and for people with more previous knowledge of the topics. Additionally, though no age group showed reliable negative testing effects, they increased for individuals with lower initial performance and previous knowledge and, surprisingly, for learners with higher nonverbal reasoning and verbal intelligence scores. These findings have important implications for the education of people of all ages and show that testing can be a beneficial learning tool for both younger and older learners.

THE POSITIVE AND NEGATIVE EFFECTS OF TESTING IN LIFELONG LEARNING

The Prevalence of Lifelong Learning in Older Adults

Formal classroom learning is a lifelong pursuit. Even older adults (here, adults aged 55 and older) participate in a variety of settings from work-sponsored continuing education courses to personal enrichment classes. Using data from the National Household Education Survey, Creighton and Hudson (2002) reported that adults aged 55 and older participated in formal learning activities in 1991, 1995, and 1999 at a rate of 32, 42, and 53%, respectively. These formal learning activities included any type of continuing education (apprenticeship programs, job- or career-related courses, personal development courses, general education development (GED) preparation classes, English as a second language (ESL) classes, etc.) in addition to the pursuit of a degree at a university or college.

The high and increasing prevalence of older students in formal learning environments is important when one considers the techniques that are being developed and tested to improve learning and memory in students, such as the testing of material to improve long-term retention (Roediger & Karpicke, 2006b). Many of these techniques are normally tested only on younger college students (i.e., students 18 to 25 years old), so it is unclear how or if the benefits of testing would generalize to the large population of older adults who participate in lifelong learning. Interestingly, these learning methods could be especially important for this group of students, considering that older students typically show higher dropout rates in college settings (Bean & Metzner, 1985) and many jobs require the completion of continuing education (e.g., 44% percent of adults who participated in the National Household Education Survey in 1999 were required to complete continuing education courses for their jobs) (Creighton & Hudson, 2002). As such, these techniques should be examined in both younger and older learners in order to assess their usefulness in a more meaningful and generalized way.

The Positive Effects of Testing

A premier technique that has been widely successful in younger student populations is testing. The act of taking a test has benefits beyond simple assessment: testing improves a student's memory for the tested material, meaning that a person will do better on a test if he or she was previously tested on the same material than if he or she was not previously tested. This phenomenon is best known as the "testing effect," but is also known as a retrieval effect, or retrieval enhanced learning (Roediger, Agarwal, Kang & Marsh, 2010).

Many researchers have studied the benefits of testing on subsequent memory, though initial studies compared testing to a lack of testing (i.e., doing nothing), rendering overall exposure to the study material unequal (Myers, 1914; Spitzer, 1939). More recently, however, researchers have corrected this methodological flaw and have compared testing to additional studying of the material in an attempt to equate overall exposure to the studied material. With the use of studying as the control group, testing was still found to be beneficial to long-term performance above and beyond mere studying (studying has only been found to be relatively more beneficial for short-term performance) (Butler & Roediger, 2008; Carrier & Pashler, 1992; Glover, 1989; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008; Kuo & Hirshman, 1996; Roediger, Agarwal, Kang & Marsh, 2010; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978; Toppino & Cohen, 2009).

As a prime example of an experiment where exposure was controlled, Roediger and Karpicke (2006a) had participants study two prose passages on the sun and on sea otters, then, in an experimental learning phase, restudy one of the passages and take a free-recall test on the other passage in which they wrote down as much as they could remember (a free-recall test). Then, in the final testing phase, each participant was given a delayed, final free-recall test on both passages 5 minutes, 2 days, or 1 week later. When final testing occurred only 5 minutes after the experimental learning phase, restudied passages were remembered better than previously tested passages were. Conversely, when final testing occurred after a longer delay (2 days or 1 week), previously tested passages were remembered better than restudied passages were (i.e., a positive testing effect occurred) (see Figure 1). These findings show that the benefits of testing are not attributable to exposure time and are separate from the more cursory benefits of additional study. Furthermore, they cannot be attributable to "transfer-appropriate processing," which theorizes that the benefits of testing occur because the conditions of testing are practiced and this practice will benefit subsequent testing. If this were true, the practice of testing one prose passage would benefit or generalize to the later testing of the other, untested passage because testing conditions are similar for both passages. This experiment, along with various other studies that support a similar finding, makes it evident that the act of testing improves long-term memory of tested information, giving credence to the notion that testing is an effective learning tool and not just a tool for assessment.

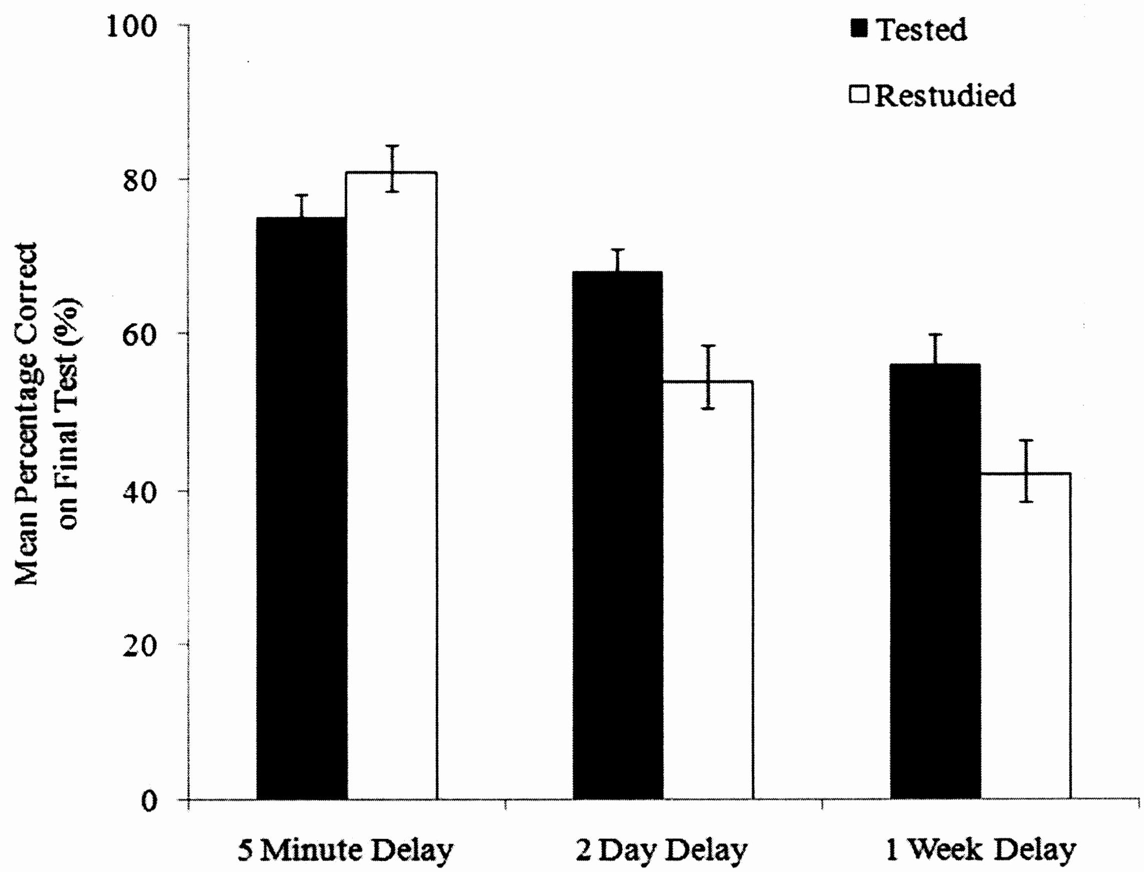

Figure 1. Mean percentage of material recalled on the final test as a function of final test retention interval and experimental learning condition (testing vs. restudying) in Experiment 1 (reproduced from Roediger & Karpicke, 2006a). Error bars represent standard errors of the means.

The positive effects of testing have been found to generalize to different types of tests, including; free-recall tests, both essay (Glover, 1989; Roediger & Karpicke, 2006a) and short answer (Hogan & Kintsch, 1971; Kuo & Hirshman, 1996; Thompson, Wenger, & Bartling, 1978); cued-recall tests (Butler & Roediger, 2008; Carrier & Pashler, 1992; Karpicke & Roediger, 2008; Toppino & Cohen, 2009); and recognition tests, including true/false tests (Sproule, 1934; Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993), old/new recognition tests (Hogan & Kintsch, 1971), and multiple-choice tests (Butler & Roediger, 2008; Marsh, Roediger, Bjork, & Bjork, 2007). Though all types of testing have been shown to improve subsequent performance, Kang, McDermott, and

Roediger (2007) found that regardless of the type of test that was used to assess final performance, the types of test that require more effortful retrieval during the initial test resulted in the best long-term performance, suggesting that retrieval rather than a match between test types in the learning and final test phase, facilitates long-term learning.

It has also been documented that the positive effects of testing also generalize to different types of learning materials including; word lists (Hogan & Kintsch, 1971; Kuo & Hirshman, 1996; Thompson, Wenger, & Bartling, 1978); drawings (Glover, 1989); trigram-number pairs (i.e., a trio of letters made up of a consonant, followed by a vowel, followed by another consonant, and finally followed by a number, e.g., "FOT-24") (Carrier & Pashler, 1992); and more relevant to education, lectures (Butler & Roediger, 2008), prose passages (Glover, 1989; Roediger & Karpicke, 2006a), and foreign language vocabulary (Carrier & Pashler, 1992; Karpicke & Roediger, 2008; Toppino & Cohen, 2009).

The psychological mechanisms behind the positive effects of testing.

Three main theories have been put forth to explain the positive testing effect: the amount of exposure theory, the transfer-appropriate processing hypothesis, and the variable processing theory. Each will be briefly discussed in turn.

Amount of exposure.

The amount of exposure hypothesis asserts that testing is beneficial to subsequent memory performance, because additional processing occurs for the material prior to the final test (i.e., learners are given an additional exposure to the material). More exposure should equate to more learning and more eventual memory when compared to material that does not receive this extra processing (Glover, 1989).

In early studies of the testing effect when researchers compared testing to a lack of testing, this was a reasonable theory. However, when subsequent researchers added a control in which exposure to the material was not increased in the testing condition relative to the untested condition, similar results were still found. Specifically, researchers compared final performance for items that were tested during the learning phase to final performance for items that were restudied during the learning phase. Items in both conditions were processed for an additional amount of time, so exposure was controlled. With exposure controlled, testing still led to increased memory when compared to untested (or restudied) material (Carrier & Pashler, 1992; Kuo & Hirshman, 1996). Therefore, the additional exposure hypothesis could no longer account for the benefit seen from the use of testing as a learning tool.

In addition to the theory's inability to account for the differences seen when exposure is controlled, it also fails to explain why testing and restudying result in different patterns of results depending on the retention interval between the learning session and the final test. Specifically, at short retention intervals (e.g., when 5 minutes pass between initial testing or restudying and the final test), the restudied material is often remembered better than the tested material (Roediger & Karpicke, 2006a). Conversely, at longer retention intervals (e.g., when 1 or 2 days pass between initial testing or restudying and the final test), the tested material is remembered better than the restudied material (i.e., a positive testing effect occurs) (Roediger, Agarwal, Kang, & Marsh, 2010). The theory not only fails to account for these differences, but would predict similar effects at all retention intervals.

Transfer appropriate processing.

The transfer appropriate processing theory claims that taking a test on information during the learning phase is similar in processing to taking a test on information later. Because of this similarity in the two tests, learners will have an advantage on the final test.

This theory, however, would predict that the testing effect would be greater when the type of test during learning and the type of the final test are similar. On the contrary, the positive testing effect is not affected by the match between the learning test and the final test. For example, Kang, McDermott, and Roediger (2007) found that the initial type of test given during the learning phase was indicative of subsequent effects of testing, whereas, the match between the first test and the final test were not indicative of such. To be more specific, learners took an initial test during learning that was either a recognition test or a cued-recall test. The final test was then either a recognition test or a cued-recall test. These were combined such that participants either took an initial recognition test followed by a final recognition or cued-recall test, or they took an initial cued-recall test followed by a final recognition or cued-recall test. Learners who took the cued-recall test during learning had the largest positive testing effects regardless of the type of final test that they took. This shows that the activity done during learning and not the match between learning and test conditions explains the mechanism behind the testing effect.

Variable processing via successful retrieval hypothesis.

This theory is the prevailing theory. It is the variable processing hypothesis in which successful retrieval during the learning phase purports to increase the number of available cues via variable processing, resulting in better retrieval later. Specifically, since studying then taking a test involves more variable processing than does studying then restudying, the processing creates more retrieval cues that the learner can rely on at later retrieval (Carpenter & DeLosh, 2006; Glover, 1989; Toppino & Cohen, 2009).

This variable processing, however, relies on successful retrieval at the time of the learning phase. If successful retrieval does not occur, the additional, variable processing does not occur, resulting in no benefit. Research has shown that testing *is* more beneficial when successful retrieval occurs and when the retrieval is more complete (i.e., when more information has to be retrieved by the learner, such as in a free-recall test, and less so in a cued-recall or recognition test). Also, when learners do poorly on the initial test, they show little or no testing effect.

The positive effects of testing in older student populations.

Though the positive effects of testing generalize to different types of tests and learning materials, they may or may not generalize with age of the learner, meaning that the positive effects of testing in lifelong learning are uncertain. Though most studies have been done using younger undergraduate students as participants, a few studies have looked at the effects of age on the benefits of testing or retrieval practice. However, evidence is mixed and older adults show poorer long-term episodic memory, which has been linked to decreased testing effects (Marsh, Agarwal, & Roediger, 2009). This evidence will be discussed in turn.

Decreased benefits from retrieval practice?

It is unclear whether older adults benefit as much from retrieval practice as younger adults do. For example, studies executed by Henkel (2007; 2008) have shown that older adults (63-90 years old) benefit less from repeated retrieval practice when compared to younger adults (18-22 years old) when memorizing both seen and imagined images. On the other hand, Rabinowitz and Craik (1986) have shown that older adults (61-75 years old) can benefit as much from retrieval practice as younger adults can. When memorizing words from various categories, there was no difference between the age groups in terms of benefits accrued from retrieval practice, meaning both age groups improved by a similar percent.

Both of these studies however, do not use additional study trials for comparison or even a control condition in which items are not tested. Therefore, it is unclear whether older adults would or would not benefit from testing like younger adults do when compared to the proper controls. As such, it would be advantageous to study the effects of testing in both younger and older adults who might all benefit from testing. Doing so could uncover the potential usefulness of testing in lifelong learning.

Long-term episodic memory impairments.

Older adults often show poor long-term episodic memory relative to their younger counterparts (Balota, Dolan, & Duchek, 2000), meaning that they are less able to accurately remember information when it is no longer in their sensory or short-term/working memory. This impairment in older adults occurs despite spared and even increasing semantic knowledge (e.g., vocabulary) (Park, 2001). Specifically, older adults perform poorer on measures of cued-recall and free-recall (measures of long-term episodic memory) relative to younger adults, but they outperform younger adults on vocabulary tests (a measure of semantic knowledge) (Park, 2001).

Poor episodic memory performance is relevant to the benefits of testing, because previous researchers have found a relationship between initial performance (successful retrieval) and the benefits accrued from testing, such that better initial performance correlates with improved performance (Marsh, Agarwal, & Roediger, 2009). Furthermore, the variable processing through successful retrieval hypothesis purports that having poor episodic memory may adversely affect the occurrence of positive testing effects by limiting variable processing. Combined with the decreased benefits from retrieval practice, older adults may ultimately benefit less from educational testing because they start off performing lower and they improve less from one test to another. Though it would not be surprising if they benefit less from testing, it will be informative to see if older adults benefit at all from testing compared to restudying (i.e., should testing even be used as a learning tool in older adults?).

The Negative Effects of Testing

Though testing can enhance younger college students' long-term memory of tested material, any test that exposes learners to erroneous answers, like a multiple-choice test or even a true/false test, can adversely affect later memory. For example, for each question in a multiple-choice test, a student is typically exposed to one correct answer choice- but three or four incorrect answer choices. Though evidence shows that testing improves memory for the material on a test, this can also include the incorrect material presented on recognition tests, resulting in an increased production of lures (or a reproduction of the wrong answers seen on the test) compared to the production of those same lures by chance. The increased (yet inaccurate) memory of false information presented on a test is known as the negative testing effect (Roediger & Marsh, 2005).

As an example of research on the negative testing effect, Roediger and Marsh (2005) tested students' reading comprehension of passages from the Test of English as a

Foreign Language (TOEFL) and from the Graduate Record Examination (GRE) and found that students performed better on a final, cued-recall test if they had been tested over the material earlier using a multiple-choice test than if they had not been tested earlier (i.e., a positive testing effect). On the other hand, this testing increased the chance that students used the incorrect answers on a subsequent cued-recall test (i.e., a "negative testing effect"). Specifically, the students produced the lures that had been seen on the earlier test more than other students used those same lures by chance (the students who were not tested previously had not seen the lures, so their lure production represented the baseline or chance amount that those lures would be produced without testing). Interestingly, even though subjects were discouraged from guessing, 75% of the errors produced on the final recall test were lures that had been selected on the previous multiple-choice test. Although testing improved students' overall performance on the final test, it also lead them to provide erroneous answers that they were relatively confident about (when told not to guess, they still provided a large rate of false answers). Learners, however, were not given feedback in this particular study, so it is likely that they reproduced lures because they had no indication that their first responses were incorrect.

Using similar methodology as Roediger and Marsh while having students take SAT II subject tests, Huelser and Marsh (as cited in Marsh, Roediger, Bjork & Bjork, 2007) examined students' reasons for selecting multiple-choice answers and found that the lures that persisted to the final test were lures that learners reasoned about, not ones they merely guessed on (15% of guesses persisted, 36% of reasoned answers persisted, and 67% of answers that relied on personal experiences persisted). The authors suggested that the persistence then is not from mere familiarity, but from a reasoning process (they said that if it was familiarity, all types would persist equally because they assume all items are seen equally as per instruction to students to read all answer choices). The authors also found that selecting a lure increased the chance that students would use that lure on a later transfer question, suggesting that the effects were not simply from the priming of the errors, but from integration of the lures into general knowledge. Their recommendation for educators is to deliver immediate feedback (e.g., with Immediate Feedback Test (IFAT) scantrons), to offer a "don't know" option with a penalty for selecting wrong answers, or to switch from definitional to applied questions (which reduces, but does not eliminate negative testing effects). In this study, it is not clear if feedback was given.

Roediger, Agarwal, Kang, and Marsh (Experiment 1, 2010) showed that taking a multiple-choice test can both enhance and diminish performance and that the positive and negative testing effects can be altered by initial performance (which supports the variable processing through successful retrieval hypothesis). In the first study, where general knowledge questions were used as the learning stimuli, Roediger and colleagues found that when multiple-choice performance was high (questions were easier and there were fewer alternatives), the positive effect of testing was high and the negative testing effect was low, but when multiple-choice performance was low, the balance shifted, such that the negative testing effect increased and the positive testing effect decreased. They found that 78% of the answers incorrectly provided on the second test (the short answer test) were lures from the first multiple-choice test, suggesting that the negative testing effect (sometimes referred to as the negative suggestibility effect) comes from erroneously selecting an answer, not from merely reading it. Importantly, negative testing effects were seen in this study despite the fact that feedback was given to all participants after the initial test.

In another study using educationally relevant materials and multiple-choice tests, Marsh, Agarwal, and Roediger (2009) had students take SAT II subjects tests (or not) without initial studying (unlike most studies that have an initial studying period). Though the experiment did not include an initial study session, the material tested on the SAT II subject tests was material that the students were exposed to during their recent education, so the experimenters did not see a need for an initial study period. Again, both positive and negative testing effects were found. Answering multiple-choice questions first increased recall on a final cued-recall test, but also resulted in more intrusions of lure items compared to a baseline condition. They also found that the top 25% of performers on the initial multiple-choice tests improved much more from the initial multiple-choice test to the final cued-recall test as a result of testing than did the bottom 25% of performers, such that they had more answers that were correct and fewer lure intrusions (a larger positive testing effect and a smaller negative testing effect). In the first experiment, subjects could refrain from answering if they were not sure of the correct answer, but in the second experiment, they had to answer every question, regardless of how sure they were of the correct response. This forced responding increased multiple-choice test errors and lures on the final cued-recall test. Additionally, compared to undergraduates (Experiment 3), high-school students answered more questions incorrectly, selected more distractors, skipped fewer questions, improved less from testing, and made more lure intrusions, lending credibility to the idea that testing effects can affect populations of students differently. It is unclear why the age difference would matter in this case, but it may have to do with a disparity in previous knowledge, which might lead the less knowledgeable group to have poorer initial performance, or it might even have to do with study strategy differences.

The negative effects of testing in older student populations.

The negative testing effect has only been studied in younger college populations (i.e., 18- to 25-year-old university students) and, narrowly, in high school students. Older students, on the other hand, may show increased negative testing effects for a few reasons. Older adults have associative deficits and poorer long-term episodic memory, which may enhance the negative effects of testing, both of which may contribute to increased negative testing effects (Erngrund, Mäntylä, & Nilsson, 1996; Park, 2001). These deficits, along with how each may affect the negative effects of testing will be discussed.

Associative deficits and source memory problems.

When compared to younger adults (here, students aged 18-25), older adults have been shown to have associative deficits, or difficulties in making connections between single ideas or ideas and the circumstances around them. Specifically important to the current question of interest is the finding that older adults have impaired source memory and source monitoring abilities, a particular type of associative deficit (Naveh-Benjamin, 2000). Source memory consists of ascribing a memory, knowledge, or a belief to its origin or source. This origin includes the context under which a memory or other piece of information is acquired and can include, among other things, spatial characteristics, temporal features, and the modality in which it was first encountered. The ability to determine the source of a memory can allow a person to determine where that given memory began. For example, they would more easily be able to determine whether a memory occurred with one friend or another, whether it was read from a credible newspaper or in a tabloid, or even whether they dreamt it or whether it really happened (Johnson, Hashtroudi, & Lindsay, 1993).

Older adults have been shown to have a deficit in their source memory relative to younger adults, making it difficult for older adults to monitor the source of a memory and the context surrounding it (Dywan & Jacoby, 1990; Hashtroudi, Johnson, & Chrosniak, 1989; Johnson, Hashtroudi, & Lindsay, 1993; McIntyre & Craik, 1987; Mitchell, Johnson, & Mather, 2003; Rabinowitz, 1989). For example, McIntyre and Craik (1987) tested younger ($M = 19$ years old) and older adults ($M = 69$ years old) from Canada on trivia about their country and found that older adults were less likely to attribute the source of their knowledge correctly. Specifically, when both groups answered a question incorrectly during the first session and were then given feedback, the older adults came back during the second experimental session a week later and were much less likely to ascribe their knowledge of these items to the previous experimental session. Instead, they claimed to have previously learned the information from a book, magazine, newspaper, class, television program, etc. (even though they answered incorrectly the week before). Though it is possible that the experiment merely reminded older adults about where they had learned information previously, a follow-up experiment showed similar results when testing these different age groups over fabricated trivia, ruling out previous knowledge as a confound. Furthermore, when older adults were able to assess that the knowledge had been attained in the experiment, they were still less likely than the younger adults to be able to determine whether that information had been presented verbally or on the overhead projector.

As another example, Mitchell, Johnson, and Mather (2002) found that older adults (*M* = 76 years old) showed decreased source memory for the events in a video of a burglary when compared to younger adults (*M* = 20 years old) after both age groups were exposed to subsequent misinformation. Specifically, older adults were less able to ascertain if information related to the videotaped scenario was actually shown in the video or whether it was merely suggested to them after watching the video. The presentation of false information after the video adversely affected older adults more than it did younger adults. This presentation of the misinformation may be similar to when one is presented with incorrect answer choices or lures in a multiple-choice test after the encoding of the previously studied material. In the case of the multiple-choice test, the presentation of lures could then be considered a type of misinformation. As such, one might find similar detriments in older adults' abilities to identify the original source of lures from a multiple-choice test when compared to younger adults' abilities.

Erngrund, Mäntylä, and Nilsson (1996) conducted a cross-sectional study of 1000 individuals between the ages of 35 and 80 years old and found that source memory performance decreased as a function of age starting around the age of 50. Specifically, participants were less able to remember the source of studied information about known and unknown people the older the participants were (even when item memory was controlled and the source memory was only assessed for correctly remembered items).

If older adults are less able to monitor the source of their memories or knowledge (or, put more simply, where they have encountered the information before, either in the study session only, on the test only, or on both the study session and on the test), they may be more susceptible to the negative testing effect. The negative testing effect occurs when one integrates incorrect information provided on a test into one's knowledge. If older adults cannot determine whether information was studied or whether it was merely an incorrect answer or lure provided on a multiple-choice test, they may be more likely than younger adults to acquire this false information as fact. As such, older adults may benefit less from multiple-choice tests in a classroom.

Long-term episodic memory impairments.

As discussed above, older adults often show poor long-term episodic memory relative to their younger counterparts (Balota, Dolan, & Duchek, 2000). The episodic memory impairments that older adults experience may affect the extent to which they are influenced by the negative testing effect, because they will likely perform more poorly than younger adults on an initial multiple-choice test that follows an initial study period and will, thus, choose relatively more lures as initial answers. Thus, poorer episodic memory may exacerbate the previous problem of having poorer source memory. Previous researchers have found that lower initial performers obtained a higher negative testing effect and improved less overall than higher performers (Marsh, Agarwal, & Roediger, 2009). As such, it is of interest to see if older learners perform more poorly than younger learners initially and, if so, if this low initial performance leads them to incur a higher negative testing effect when using testing as a learning tool. Furthermore, older adults' relatively poor performance may worsen over a longer retention interval, since they show quicker rates of forgetting (Park, Royal, Dudley, & Morrell, 1988). Quicker forgetting may lead to the incorporation of false information more after a longer delay than after a shorter delay.

Summary of Key Issues

Testing has been shown to be beneficial to younger students and their learning. However, it has not been studied in older student populations and may not be as beneficial to older students who may show poorer episodic and source memory. This may result in a decreased positive testing effect and an increased negative testing effect relative to younger students. On the other hand, some studies do report that older adults benefit from retrieval practice to the same extent as younger adults, leaving open the question of whether students of different ages will benefit differently from testing.

While assessing this question, the current study also assessed whether initial performance and previous knowledge is related to the positive effects of testing, which the variable processing hypothesis would predict. Also, it asked if the size of a learner's negative testing effect was related to his or her initial test performance, previous knowledge of the tested and restudied topics, and/or source monitoring abilities. It may be that these factors, in addition to or instead of age, contributed to a student's enhancement in performance and learning as a result of testing.

Overview of Experiment

The following experiment looked at the effects of testing in both younger and older students and assessed two main questions: 1) were the positive and negative effects of testing affected by age such that both younger and older students benefited from testing to different extents in immediate and longer-term learning? and 2) were the positive and negative effects of testing related to one's initial performance, previous knowledge, source monitoring abilities, or some combination of these factors? To address the first question, the following experiment examined both the positive and negative effects of testing in both younger and older adults at different retention intervals. Additionally, the second question was addressed by regressing age, final test retention interval, intelligence, initial performance, previous knowledge, and source monitoring performance onto the positive and negative effects of testing.

Briefly, the experiment followed this general set-up: an initial study phase, a brief distractor phase, the experimental learning phase where participants restudied some material and took an initial test on other material, another brief distractor phase, a retention interval manipulated between-subjects, a final cued-recall test, and an intelligence test (see Figure 2).

Study phase

(Study 4 topics)

⇩

Distractor phase

(Multiplication problems)

⇩

Experimental learning phase

(Initial recognition test of 2 topics with feedback + Restudying of other 2 topics)

⇩

Between-subjects final test retention interval of 5 minutes or 2 days

Distractor phase 2

(Multiplication problems)

⇩

Final test phase

(Recall test over all topics + Previous knowledge questions + Source memory questions)

⇩

Intelligence test

Figure 2. General outline of experiment.

Experiment

Method

Participants.

Participants consisted of 180 adults from the following three populations: 60 younger adults (aged 18-25) (*M* = 19.3, *SEM* = 0.16) from the Rice University undergraduate population, 60 younger adults (aged 18-25) (*M* = 22.4, *SEM* = 0.25) from the Houston community, and 60 older adults (aged 55-65) (*M* = 59.8, *SEM* = 0.43) from the Houston community (see Table 1 for education levels of each group). Though it would be ideal to test a large range of ages, testing the age of young adults who are typically tested in similar studies (i.e., 18-25 year olds) and testing older adults who are still young enough that they are still likely to participate in lifelong learning, but old enough to show memory decline will allow one to see if age differences are worth looking at over a range of ages in a future study. According to the Education survey by Creighton and Hudson (2002), 55-65 year-olds are more likely to participate in lifelong learning than adults aged over 65 are. Additionally, Park (2001) shows that 55-65 year olds show a significant decline in various memory processes. Participants were recruited from Rice through Momentum™ Experiment Scheduling System (known as Experimetrix to Rice faculty and students), a web-based experiment scheduling and tracking program (Sona Systems, Ltd., 2000-2006). Community participants were recruited through craigslist and through flyers posted in the Houston community near Rice University (craigslist, 2010). Younger adults from Rice University are similar to the types of students included in previous studies of the testing effect and were used to provide a good comparison between previous studies and the current study. Younger adults from the community, on the other hand, were used to provide a more appropriate comparison for the older adult population as those two populations were more likely to have similar backgrounds (e.g., education and intelligence). The participants from Rice University were compensated with partial course credit for their participation while volunteers from the community received $20 for their efforts. Inclusion criteria for all participants included: normal or corrected to normal vision, some college experience, and fluency in English (the latter two criteria were to ensure that all participants could easily read and understand the study materials). Additionally, all participants in this study were treated according to the ethical standards of the American Psychological Association (APA) and Rice University's Institutional Review Board (IRB) concerning the use of human participants in research. A priori power analyses suggested that this number of participants should have allowed for sufficient power (.8+) to detect any experimental differences.

	Young Adults from Rice (%)	Young Adults from Community (%)	Older Adults from Community (%)
Some College	60 (100%)	35 (58.3%)	14 (23.3%)
Completed College	0 (0.00%)	18 (30.0%)	29 (48.3%)
Graduate School	0 (0.00%)	5 (8.33%)	9 (15.0%)
PhD/MD/JD/DDS	0 (0.00%)	2 (3.33%)	8 (13.3%)

Table 1. Number of people with each education level within each age group.

Materials.

Study phase.

The study materials included four passages obtained from National Geographic online covering the topics of tsunamis, black holes, the human heart, and armadillos. Each of these passages is approximately three-quarters to one full page in length (single spaced, Times New Roman, size 12 font). These passages were chosen because they seemed to be topics that people did not necessarily know a lot about, yet were interesting and engaging to read. Pilot testing done on undergraduates showed this to be the case (see Appendix A).

Distractor phases.

The multiplication problems that participants solved during the distractor phases were generated on an education website (see Appendix B) (SuperKids, 1998-2010). The multiplication problems consisted of two factors and each factor was between 10 and 50 (e.g., 14 X 49 = _____). This math distractor was used instead of a verbal task to ensure that there was no overlapping material in the distractor phases and the other phases of the experiment.

Experimental learning phase.

Initial recognition test.

Questions for the initial multiple-choice tests were created for this study such that there were one correct and three incorrect answers per question (see Appendix C). For example, participants might see the following:

1. The last stage before a black hole is formed, a detonation occurs, known as a(n)
    a. starburst
    b. explosion
    c. blastula
    d. supernovae

There were 10 such questions for each of the four passages, resulting in 40 questions total (though each participant saw only 20 questions: 10 from each of 2 topics). The 20 questions administered to each participant were presented in a random order. In a pilot test of the materials, there was no significant difference between participants' performance on the topics and performance was around 60%.

Restudying.

The same passages used for the initial study phase were re-used for restudying.

Final cued-recall test with source memory questions.

The final test questions were created by taking the questions in the initial multiple-choice test and merely providing blank to fill in the answers rather than providing the participants with four multiple-choice alternatives for each question (see Appendix D). As an example, participants saw the following question in the final test, which corresponds to the question above:

1. The last stage before a black hole is formed, a detonation occurs, known as a(n) ____________.

There were 40 cued-recall or fill-in-the-blank questions (again, 10 questions per reading passage). Each of the cued-recall questions was then followed by a multiple-choice source memory question. The question asked the participants about the source of their knowledge as it pertained to the experiment. For example, after seeing the following question:

1. The last stage before a black hole is formed, a detonation occurs, known as a(n) ____________.

participants saw the following source memory multiple-choice question and answer choices:

> Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
> a. readings
> b. previous test
> c. readings AND previous test
> d. previous knowledge outside of the experiment
> e. I'm just guessing

These answer choices were the same for each source memory question (see Appendix D). There were also additional questions that covered material not studied or tested in the experiment during either the study or experimental learning phases. These questions were related to the four topics studied by participants, but were not facts included in the readings (e.g., questions about tsunami they could not have answered based on the readings or on the initial test). There were 20 of these additional questions (5 for each topic studied). These were used to assess previous knowledge about the studied topics. All 60 questions were presented in the same random order to all participants.

Intelligence test.

The two subtest version of the Wechsler Abbreviated Scale of Intelligence (WASI) was used to measure each participant's intelligence. The two subtests consisted of a vocabulary section with 34 words and a matrix reasoning section with 25 reasoning puzzles in which participants engaged in abstract reasoning processes. (However, number of items administered to each participant depended on the performance of each participant at the time of the test, since rules for administering the tests dictate that the test administrator stops after a predetermined number of consecutively missed questions) (Wechsler, 1999).

Follow-up questionnaire.

Follow-up questions included questions about each participant's education level, sex, and questions about knowledge of the positive and negative testing effects. The questions were included to ensure that all participants had college experience and that participants included were not knowledgeable about the positive and negative effects of testing (all participants had college experience and no participants had knowledge of the effects of interest). Also, though sex of each participant was collected, it was not used in any analyses, since there were no theoretical reasons to believe that sex would affect positive and negative testing effects (see Appendix E).

Procedure.

The procedures used were similar to those used previously by researchers who studied the positive and negative effects of testing in younger adults (Butler & Roediger, 2008; Marsh, Agarwal, & Roediger, 2009; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005). After participants read and signed the consent forms, they were randomly assigned to one of two between-subjects experimental conditions for the final test retention interval. Again, the phases of the experiment consisted of a study phase in which participants read all of the passages, a math distractor task, the experimental learning phase in which participants completed an initial recognition test on two of the studied topics (and received feedback) and restudied the other two topics, a second math distractor task, a final cued-recall test that included source memory questions, and the intelligence test (see Figure 2). All experimental phases were given on paper except for the intelligence test, which was administered orally by the experimenter.

Study phase.

During the study phase, participants were instructed to read the four passages selected from National Geographic (see Appendix A). They were informed that they would be tested on the information later in the experiment, but were not told specifics about the upcoming test. Each participant read the passages in the same order and were given a total of 15 minutes to read them (all participants in this study finished reading all 4 passages within the allotted time frame).

Distractor phase 1.

After the study phase, all participants were given 50 multiplication problems to work on for 5 minutes in an effort to clear the contents of their working memory, so that the subsequent tests measured long-term memory instead of working memory (see Appendix B).

Experimental learning phase: Initial recognition test and restudying.

After the first distractor task, all participants took an initial recognition test in the form of a multiple-choice test over two of the studied topics and also restudied the other two topics. The order that these two parts of the experiment occurred was counterbalanced across participants and the particular topics tested and restudied were also counterbalanced across participants. For example, participant 1 took an initial multiple-choice recognition test over the tsunami and black hole articles, received feedback on that test, and then restudied the human heart and armadillo articles. Participant 2, on the other hand, restudied the human heart and armadillo articles, took a multiple-choice recognition test on the other topics, and then received feedback on that test. For the multiple-choice test each problem contained one correct answer choice and three incorrect answer choices (see Appendix C for full list of multiple-choice questions). Participants worked on answering all 20 questions and then notified the experimenter when they were finished (it was self-paced). Immediately after answering all of the multiple-choice test questions, the experimenter graded the multiple-choice test and showed the results to the participants so that participants could see which answers they got correct and incorrect. For the part of the learning phase in which participants were restudying two topics, they reread each passage one more time, and then notified the experimenter (pilot testing showed that rereading two passages took about the same amount of time as answering 20 multiple-choice questions). This part was also self-paced.

Distractor phase 2.

The second distractor task was like the first distractor task, but included different questions (see Appendix B). It also was used in an attempt to clear the contents of the participants' working memories prior to the final recall test.

Final cued-recall test with source memory questions.

For each participant, the final cued-recall test occurred either directly after the second distractor task or two days later depending on which between-subjects condition of retention interval they were randomly assigned to. Retention intervals were manipulated to assess whether the different positive and negative testing effects for younger and older adults vary over time. In the test each participant was tested over the entire list of 40 cued-recall questions derived from the multiple-choice, recognition questions as well as the 20 additional questions that were not covered in the readings. The questions were presented in a randomized order. The participants were told to answer each question by filling in the blank as best as they could and to answer the source memory question that followed each cued-recall, fill-in-the-blank question by choosing the appropriate answer (see Appendix D). This phase was also self-paced.

Intelligence test.

The Wechsler Abbreviated Scale of Intelligence was administered orally to each participant according to the scale's instructions.

Follow-up questionnaire.

At the end of the study, participants completed a quick follow-up questionnaire (see Appendix E) and were then debriefed about the experiment.

Results

Measures

Correct memory performance and the positive testing effect.

The main dependent variable assessed for correct memory performance was the percentage correct on the final cued-recall test, which was compared as a function of experimental learning condition (testing vs. restudying). The size of the positive testing effect was defined as the difference between these two conditions. Specifically, the positive testing effect equaled the percentage correct on the final cued-recall test for items previously tested minus the percentage correct on the final cued-recall test for items that were restudied (or those that were *not* previously tested). If there was a significant difference between the two conditions, such that testing resulted in better final performance than restudying did, a positive testing effect occurred.

Lure production and the negative testing effect.

The main dependent variable assessed for lure production was the percentage of answers produced on the final test that were lures, which was compared as a function of experimental learning condition (testing vs. restudying). Since learners were not exposed to lures during restudy, the lures in the latter condition represent the baseline amount that these answers are produced when people are not exposed to them. The size of the negative testing effect was defined as the difference between the two measures. Specifically, the negative testing effect equaled the percentage of lures produced on the final cued-recall test for items previously tested minus the percentage of lures produced on the final cued-recall test for items that were restudied (or those that were *not* previously tested and were thus, produced by chance). The former measure represented how often participants reproduced wrong answers from the multiple-choice test, whereas the latter measure represented the number of lures that would be produced by chance, since participants had not seen these lures while restudying. If there was a significant difference between the former two measures, such that testing resulted in an increased amount of lures produced as compared to restudying, a negative testing effect occurred.

Other measures.

In addition to final test performance and lure production, two measures of intelligence, initial performance on the multiple-choice recognition test, final test performance on previous knowledge questions, and source memory performance on the final test were also assessed.

Intelligence: Measures of verbal ability and nonverbal reasoning.

A combined intelligence score was calculated by transforming the raw scores of the vocabulary and matrix reasoning sections of the Wechsler Abbreviated Scale of Intelligence into z-scores, and then adding them together. Additionally, the raw scores from the vocabulary and matrix reasoning sections were utilized for analyses looking at each subsection of the test separately. The traditional IQ scores were not calculated according to age, as this would minimize or alter age differences between the age groups.

Initial recognition test performance.

Initial recognition performance on the multiple-choice test was calculated as a percentage correct on all multiple-choice test items.

Final cued-recall test performance on previous knowledge questions.

Final cued-recall test performance on previous knowledge questions was also assessed to ascertain how much people knew about all of the topics coming into the study. This was calculated as a percentage. Additionally, the relative amount of knowledge each learner had for tested versus restudied topics was measured in relation to positive and negative testing effects. For example, if a participant came into the study knowing much about the two topics that they restudied, but little about the two topics that they were tested on before the final test, it would appear that restudying helped performance on the final test more than testing did because the participant came into the study with more knowledge about those topics. To assess the effect of previous knowledge, each participant's performance on previous knowledge questions related to the restudied topics was subtracted from his or her performance on the previous knowledge questions related to the two previously tested topics. A positive number for a participant indicates better performance on previous knowledge questions related to the tested topics compared to the restudied topics (meaning they likely came into the experiment with more knowledge about the topics in the testing condition). A negative number signifies better performance on previous knowledge questions related to the restudied topics (suggesting that more was known about these topics coming into the experiment). A number close to zero indicates similar performance on topics from both conditions, (suggesting similar knowledge on both tested and restudied topics prior to the experiment).

Source memory performance.

Source memory summarized a participant's ability to assess the source of information he or she was tested on in both phases of the experiment (whether that information was studied, studied and tested previously, or not seen) and was calculated as a percentage of the total number of times that they accurately assessed the source of the information. For example, if a participant was not previously tested over a piece of studied information prior to the final test, he or she should have selected *a) readings*. If the participant was previously tested over a particular piece of studied information in the initial recognition test, he or she should have identified this by selecting *c) readings AND previous test*. If, on the other hand, a participant had not seen a piece of information prior to the final test (i.e., it was not covered in any of the reading passages or tested on the previous test), he or she should have chosen *d) previous knowledge outside of experiment*. This latter choice would correspond to all previous knowledge questions.

Analyses

All analyses were considered significant if the *p*-value fell below .05.

Correct memory performance and the positive testing effect.

Final test performance was assessed using a 2 x 2 x 3 repeated measures ANCOVA, covarying for intelligence (using the combined raw scores from the vocabulary and matrix reasoning sections of the WASI), with the within-subjects factor of experimental learning condition (testing vs. restudying), the between-subjects factor of final test retention interval (5 minutes vs. 2 days), and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows if testing resulted in better performance than restudying did in both short- and longer-term learning and if this changed based on one's age.

The repeated measures ANCOVA shows that the covariate, intelligence, was significantly related to the dependent measure, final performance, $F(1, 173) = 103, p < .001$, MSE = 254. As such, intelligence was controlled for. After controlling for intelligence, the ANCOVA also revealed a significant main effect of experimental learning condition (i.e., testing vs. restudying), $F(1, 173) = 186, p < .001$, MSE = 133, such that testing led to better final performance (M= 73.7, SEM= 1.03) than restudying did (M= 57.1, SEM= 1.05), meaning there was an overall positive effect of testing on performance (see Figure 3). There was also a significant main effect of final test retention interval on final test performance, $F(1, 173) = 66.7, p < .001$, MSE = 254, such that participants performed better on the final test when the delay was only 5 minutes after the experimental learning phase ($M = 72.3$, $SEM = 1.19$) compared to when it was 2 days after the experimental learning phase ($M = 58.5$, $SEM = 1.19$) (see Figure 3). This is not surprising: longer retention intervals typically result in less remembered than shorter ones do. Additionally, there was a significant interaction between experimental learning condition (testing vs. restudying) and the retention interval before the final test, $F(1, 173) = 4.65$, $p = .03$, MSE = 133. A post-hoc analysis using the positive testing effect as the dependent variable reveals that the difference between the test and restudy conditions (i.e., the testing effect) increased with an increasing retention interval, $F(1, 177) = 4.62$, $p = .03$, MSE = 266 (see Figure 3), such that testing improved performance more than restudying did and this improvement increased with an increasing retention interval. Specifically, the positive testing effect (or the advantage of testing compared to restudying) was 14 points with a 5 minute retention interval ($M = 13.9$, $SEM = 1.72$) and 19 points with a 2 day retention interval ($M = 19.2$, $SEM = 1.72$).

Conversely, the main effect of age group, $F(2, 173) = 1.10$, $p = .34$, MSE = 254, and the following interactions; the interaction between learning condition and intelligence, $F(1, 173) = .167$, $p = .68$, MSE = 133, the interaction between learning condition and age group, $F(2, 173) = 1.41$, $p = .25$, MSE = 133, the interaction between age group and final test delay, $F(2, 173) = .272$, $p = .76$, MSE = 254, and the interaction between learning condition, age group, and final test delay, $F(2, 173) = .967$, $p = .38$, MSE = 133, were not significant (see Figure 3). Interestingly, age did not interact with any of the variables, meaning both younger and older adults benefited similarly from testing over both retention intervals when compared to restudying. However, though age did not affect the positive testing effect, subsequent analyses show how age and other individual differences related to the size of the positive effects of testing.

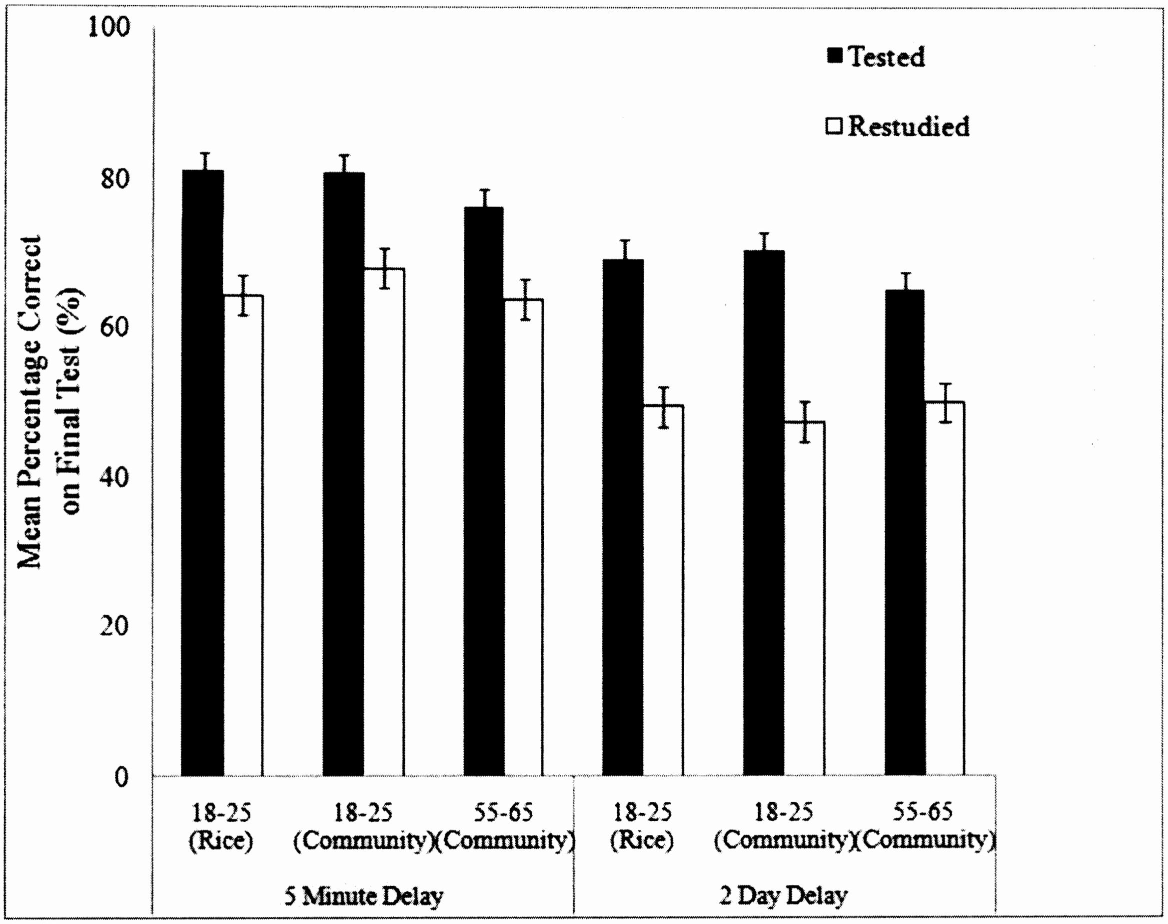

Figure 3. Mean percentage of material recalled on final cued-recall test as a function of learning condition, final test retention interval, and age group after being adjusted for intelligence. Error bars represent standard errors of the means.

Lure production and the negative testing effect.

Lure production in the final cued-recall test was examined using a 2 x 2 x 3 repeated measures ANCOVA, covarying for intelligence (using the combined raw scores from the vocabulary and matrix reasoning sections of the WASI), with the within-subjects factor of experimental learning condition (testing vs. restudying), the between-subjects factor of final test retention interval (5 minutes vs. 2 days), and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows if testing in this paradigm resulted in an increased production of lures when compared to restudying in both short- and long-term learning and if this changed with age.

The repeated measures ANCOVA shows that the covariate, intelligence, was significantly related to the dependent measure, lure production, $F(1, 173) = 5.06, p = .03$, MSE = 44.1. As such, intelligence was controlled for. After controlling for intelligence, the ANCOVA also revealed a significant main effect of final test retention interval on lure production, $F(1, 173) = 4.95, p = .03$, MSE = 44.1, such that participants produced fewer lures on the final test when it occurred only 5 minutes after the experimental learning phase ($M = 7.76$, $SEM = .50$) compared to when it occurred 2 days after the experimental learning phase ($M = 9.31$, $SEM = .50$) (see Figure 4). This is consistent with previous research in which longer delays lead to increased lure intrusions and decreased memory for source.

Conversely, the main effects of experimental learning condition (testing vs. restudying), $F(1, 173) = 1.68, p = .20$, MSE = 41.0, and age group, $F(2, 173) = .351, p = .71$, MSE = 44.1, as well as the following interactions; the interaction between learning condition and age group, $F(2, 173) = .049, p = .95$, MSE = 41.0, the interaction between learning condition and final test delay, $F(1, 173) = 1.08, p = .30$, MSE = 41.0, the interaction between learning condition and intelligence, $F(1, 173) = 1.37, p = .24$, MSE = 41.0, the interaction between age group and final test delay, $F(2, 173) = 1.59, p = .21$, MSE = 44.1, and the interaction between learning condition, age group, and final test delay, $F(2, 173) = 2.53, p = .08$, MSE = 41.0, were not significant (see Figure 4). Surprisingly, no negative effects of testing were found in this particular paradigm (at least, when comparing group means across all topics). Participants produced as many lures after restudying (or, by chance) as they did after testing. Perhaps even more surprising, older adults did not produce more lures than younger adults from either sample even after a longer retention interval. Subsequent analyses, however, looked at individual differences to determine when negative testing effects do occur more or less in relation to individual differences.

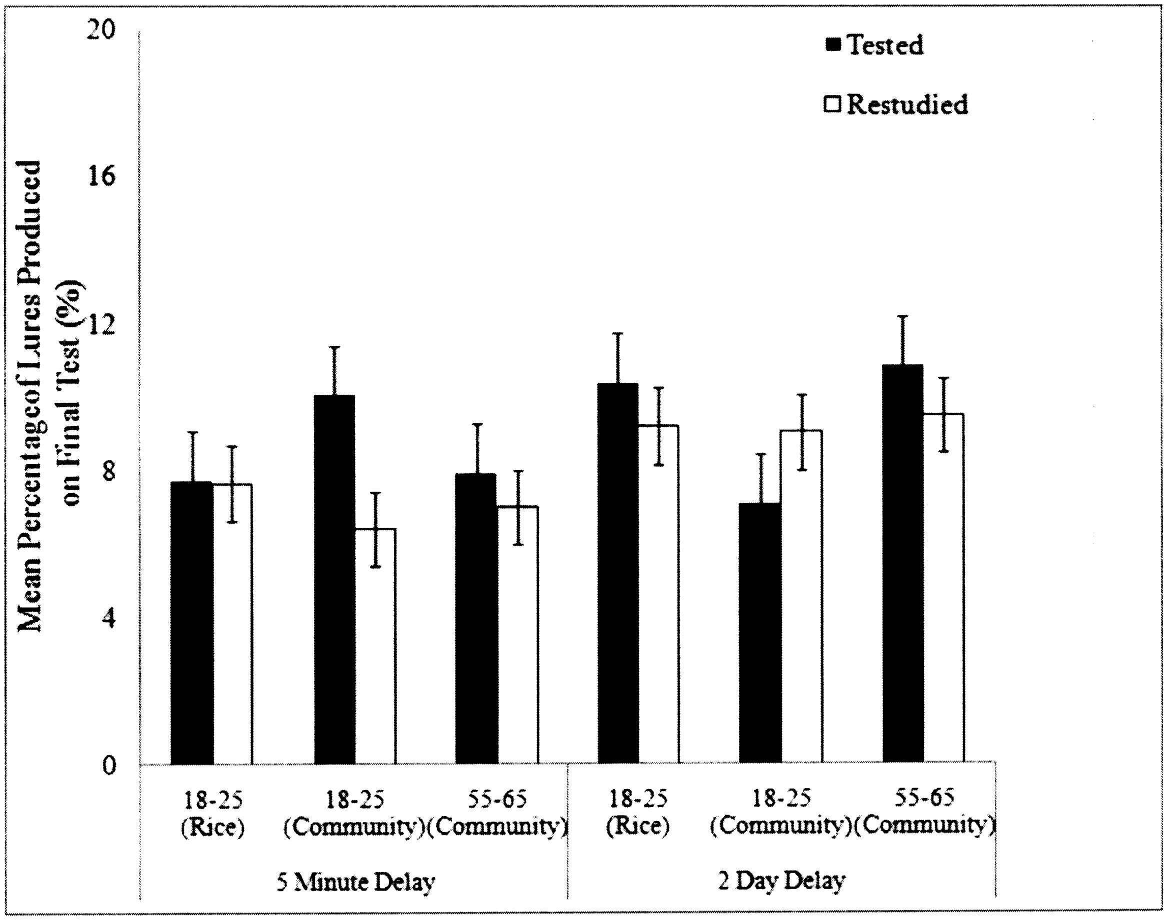

Figure 4. Mean percentage of lures produced on final cued-recall test as a function of learning condition, final test retention interval, and age group after being adjusted for intelligence. Error bars represent standard errors of the means.

Other measures.

Intelligence: Measures of verbal ability and nonverbal reasoning.

Intelligence was assessed using the combined raw vocabulary and matrix reasoning scores using a 2 x 3 between-subjects ANOVA with the between-subjects factor of final test retention interval (5 minutes vs. 2 days) and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows how the between-subjects groups differed based on intelligence and shows why intelligence was used as a covariate in other analyses (i.e., to control for this variable).

The ANOVA revealed a significant main effect of age group on intelligence, $F(2, 174) = 16.9$, $p < .001$, MSE = 2.29. Post-hoc pairwise comparisons (LSD) revealed that the younger adults from Rice ($M = .916$, $SEM = .195$) had higher combined intelligence scores than both the younger adults from the community ($M = -.577$, $SEM = .195$) and the older adults from the community did ($M = -.339$, $SEM = .195$) ($p < .001$ for both comparisons). The latter two groups, however, did not differ significantly on these combined intelligence scores ($p = .39$) (see Figure 5). Retention interval prior to the final test, $F(1, 174) = .017$, $p = .90$, MSE = 2.29, and the interaction between age group and retention interval, $F(2, 174) = .429$, $p = .65$, MSE = 2.29, was not significantly related to learners' combined intelligence scores (see Figure 5).

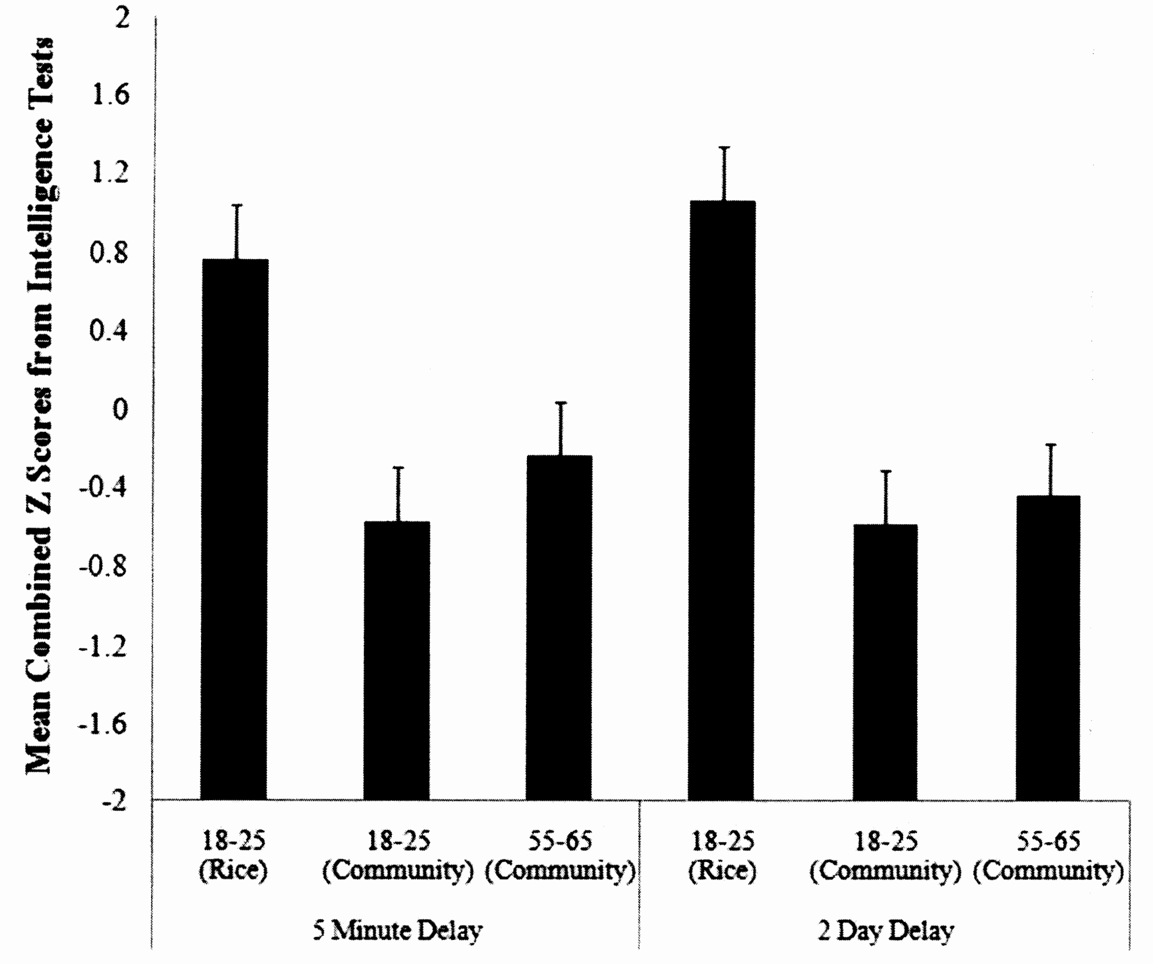

Figure 5. Mean combined z-scores from vocabulary and matrix reasoning sections of the intelligence test as a function of final test retention interval and age group. Error bars represent standard errors of the means.

Analyzing the raw scores from the vocabulary and matrix reasoning sections of the intelligence test separately revealed slightly different results. For vocabulary, or verbal ability, the ANOVA revealed a significant main effect of age group on intelligence, $F(2, 174) = 31.4, p < .001$, MSE = 59.0. Post-hoc pairwise comparisons (LSD) revealed that the younger adults from Rice ($M = 52.7$, $SEM = .992$) had higher raw vocabulary scores than the older adults from the community ($M = 47.8$, $SEM = .992$) and the older adults from the community had higher raw vocabulary scores than the younger adults from the community did ($M = 41.6$, $SEM = .992$) ($p < .001$ for both comparisons).

Retention interval prior to the final test, $F(1, 174) = .295$, $p = .59$, MSE = 59.0, and the interaction between age group and retention interval, $F(2, 174) = .849$, $p = .43$, MSE = 59.0, did not significantly affect participants' raw vocabulary scores.

For matrix reasoning, the ANOVA revealed a significant main effect of age group on raw matrix reasoning scores, or on nonverbal reasoning ability, $F(2, 174) = 8.20$, $p < .001$, MSE = 10.1. Post-hoc pairwise comparisons (LSD) revealed that the younger adults from Rice ($M = 22.8$, $SEM = .411$) and the younger adults from the community ($M = 22.0$, $SEM = .411$) had higher matrix reasoning scores than the older adults from the community did ($M = 20.4$, $SEM = .411$) ($p < .001$ and $p = .01$, respectively). The former two groups, however, did not differ significantly on these raw matrix reasoning scores ($p = .17$). Retention interval prior to the final test, $F(1, 174) = .079$, $p = .78$, MSE = 10.1, and the interaction between age group and retention interval, $F(2, 174) = .199$, $p = .82$, MSE = 10.1, did not significantly affect participants' intelligence. If intelligence was not used as a covariate, the effect of intelligence may have obfuscated the real effects of testing in younger and older adults.

Initial multiple-choice performance.

Participants' initial performance was examined with a 2 x 3 between-subjects ANCOVA, covarying for intelligence (using the combined raw scores from the vocabulary and matrix reasoning sections of the WASI), with the between-subjects factor of final test retention interval (5 minutes vs. 2 days) and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows how initial recognition test performance differed in the different between-subjects groups of age and how that differed based on final test retention interval.

The ANCOVA revealed that the covariate was significantly related to initial performance, $F(1, 173) = 48.2$, $p < .001$, MSE = 130. After controlling for intelligence, none of the other factors, retention interval, age group, or their interaction significantly affected initial recognition performance; $F(1, 173) = 1.72$, $p = .19$, MSE = 130; $F(2, 173) = .116$, $p = .89$, MSE = 130; and $F(2, 173) = .592$, $p = .55$, MSE = 130; respectively (see Figure 6). Surprisingly, all ages performed similarly on the initial multiple-choice recognition test, even though older adults typically do poorer than younger adults on tests of episodic memory (though this was controlled for intelligence). Unsurprisingly, retention interval (which occurred after this test) did not affect performance. This signifies that the random assignment of participants to the two different retention intervals resulted in similar performing groups prior to the retention interval.

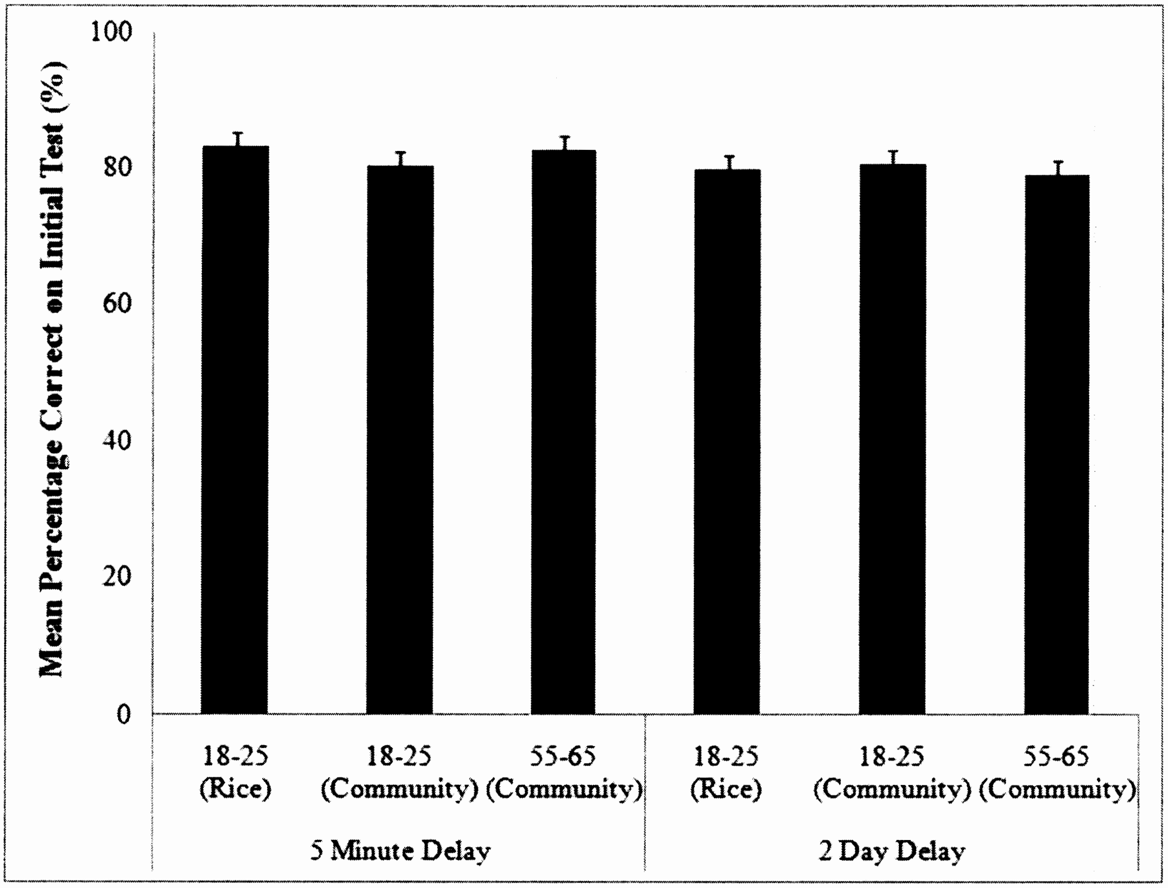

Figure 6. Mean percentage correct on initial multiple-choice recognition test as a function of final test retention interval and age group after being adjusted for intelligence. Error bars represent standard errors of the means.

Final cued-recall test performance on previous knowledge questions.

Final cued-recall test performance on previous knowledge questions for all topics was analyzed using a 2 x 3 between-subjects ANCOVA, covarying for intelligence (using the combined raw scores from the vocabulary and matrix reasoning sections of the WASI), with the between-subjects factor of final test retention interval (5 minutes vs. 2 days) and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows how final cued-recall performance on previous knowledge questions differed by retention interval and age group.

The ANCOVA revealed that the covariate was significantly related to performance on previous knowledge questions, $F(1, 173) = 34.9, p < .001$, MSE = 93.3. After controlling for the covariate, age group was marginally significant, $F(2, 173) = 2.54, p = .08$, MSE = 93.3. Nominally, older adults from the community (*M*= 13.7, *SEM*= 1.26) knew more than younger adults from the community (*M*= 10.9, *SEM*= 1.28) who knew more than younger adults from Rice (*M*= 9.67, *SEM*= 1.32). Neither retention interval nor the interaction between retention interval and age group significantly affected performance on previous knowledge questions, $F(1, 173) = 0.00, p = .98$, MSE = 93.3; and $F(2, 173) = 0.64, p = .53$, MSE = 93.3; respectively (see Figure 7). There was no reason to expect differences of performance based on retention interval, but it may or may not be surprising that older adults performed as well as younger adults did on the previous knowledge questions. Older adults may have come into the experiment with more knowledge about the topics due to life experiences they have had, but they may have been able to retrieve less about these topics at the time of the test.

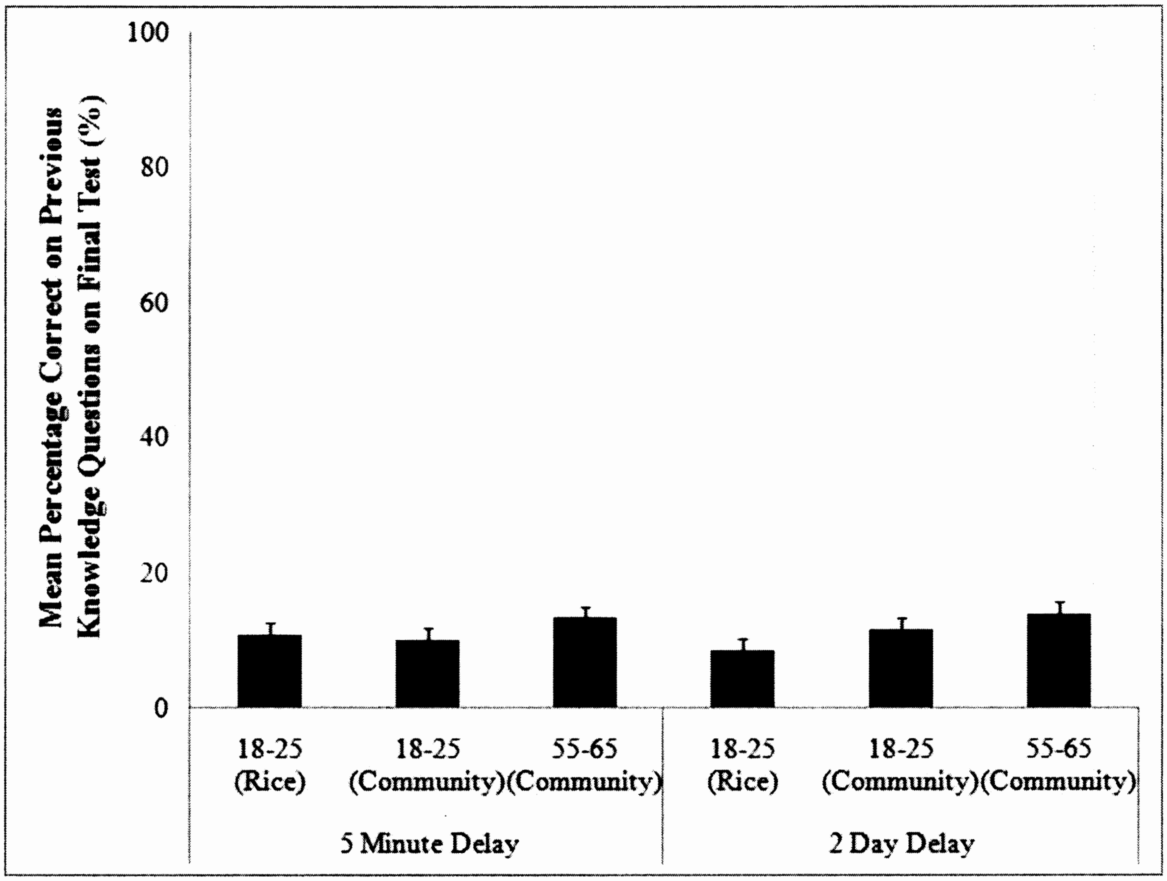

Figure 7. Mean percentage correct on final test for previous knowledge questions as a function of retention interval and age group after being adjusted for intelligence. Error bars represent standard errors of the means.

Source memory performance.

Participants' source memory performance was scrutinized with a 2 x 3 between-subjects ANCOVA, covarying for intelligence (using the combined raw scores from the vocabulary and matrix reasoning sections of the WASI), with the between-subjects factor of final test retention interval (5 minutes vs. 2 days) and the between-subjects grouping factor of age group (younger adults from Rice vs. younger adults from the community vs. older adults from the community). This analysis shows how participants in the different age group and retention interval conditions differed on the source memory task.

The ANCOVA revealed that the covariate was significantly related to source memory performance, $F(1, 173) = 26.3, p < .001$, MSE = 319. After controlling for intelligence, retention interval and age group were also found to be significantly related to source memory performance. Specifically, retention interval affected source memory performance such that participants who took the final test after 5 minutes ($M = 56.7$, $SEM = 1.9$) performed significantly better than the participants who took the final test after a 2 day retention interval ($M = 44.5$, $SEM = 1.9$), $F(1, 173) = 21.0, p < .001$, MSE = 319 (see Figure 8). This fits theories of forgetting that memory declines over time. Source memory follows this pattern. Additionally, age group significantly affected source memory performance, $F(2, 173) = 6.95, p < .001$, MSE = 319. Post-hoc pairwise comparisons (LSD) show that the younger adults from Rice ($M = 54.9$, $SEM = 2.45$) and from the community ($M = 53.4$, $SEM = 2.36$) performed better than the older adults from the community ($M = 43.5$, $SEM = 2.33$), $p = .001$ and $p = .003$, respectively, but did not perform differently from each other, $p = .67$ (see Figure 8). This also fits theories of memory and aging that contend that older adults remember less compared to younger adults. The interaction between retention interval and age group, on the other hand, was not significantly related to source memory performance, $F(2, 173) = 0.07, p = .93$, MSE = 319 (see). Interestingly, older adults did not perform disproportionately worse after a delay of 2 days, which is in opposition to data that suggests that older adults have increased rates of forgetting. This may be due to the fact that older adults in this study are not as old as is typically studied in aging studies and that older adults perform better in situations that include a meaningful context (Castel, 2005). In this study, the topics studied may have some relevance to their lives, making it easier for them to remember.

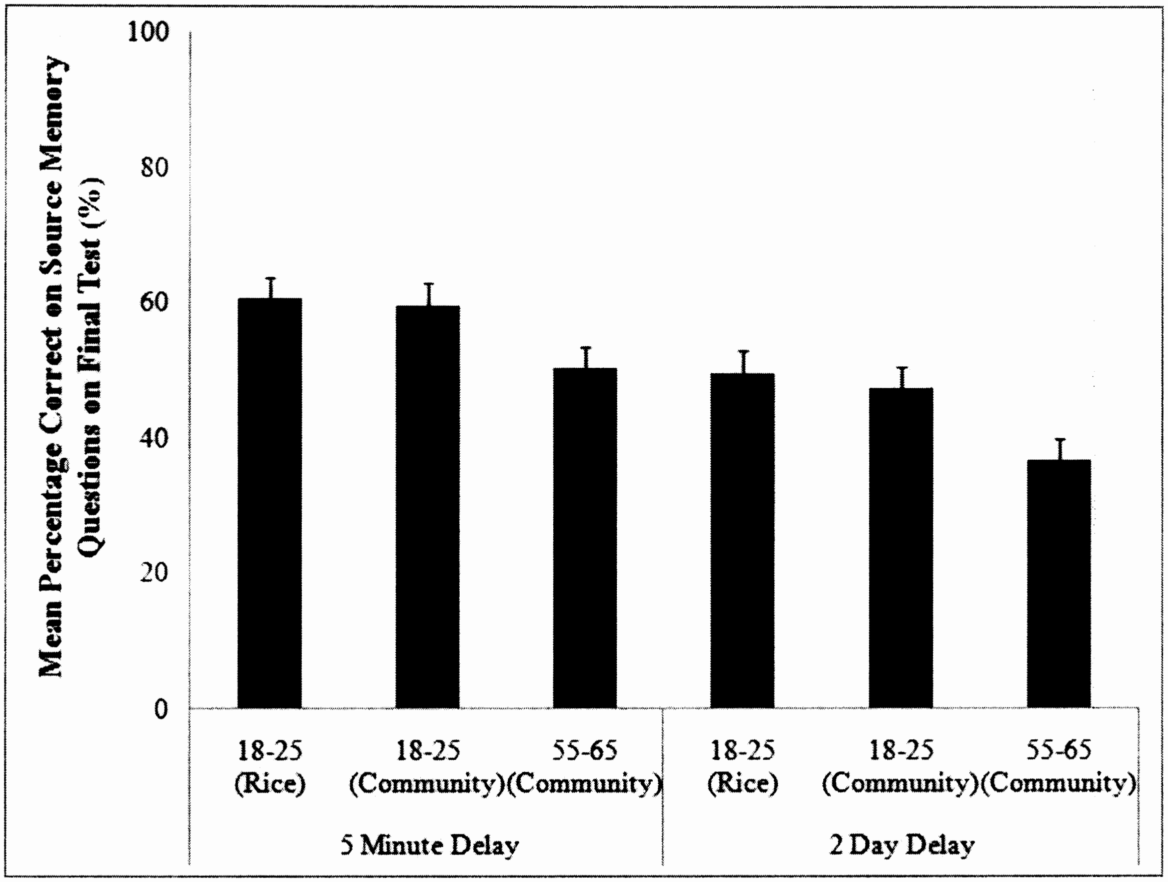

Figure 8. Mean percentage correct on final test source memory questions as a function of retention interval and age group after being adjusted for intelligence. Error bars represent standard errors of the means.

Regression of variables onto positive testing effect.

The positive testing effect was regressed onto age group (coded as 0= younger adults aged 18-25 and 1= older adults aged 55-65), final test retention interval (coded as 0= 5 minute delay and 1= 2 day delay), initial test performance, previous knowledge (tested - restudied), verbal ability (raw score on vocabulary section of the intelligence test), and nonverbal reasoning ability (raw score on matrix reasoning section of the intelligence test). This analysis shows which individual differences are related to the size of a participant's positive testing effect. Since age did not significantly affect the presence of positive testing effects, but positive testing effects vary, looking at related factors gives insight into when testing may or may not be beneficial for particular students.

The model accounted for about 16% of the variance ($R^2 = .166$) in the size of the positive testing effect, which was significant, $F(6, 173) = 5.75, p < .001$. Final test retention interval ($\beta = .154, p = .03$), initial multiple-choice test performance ($\beta = .277, p < .001$), previous knowledge ($\beta = .259, p < .001$), and verbal ability ($\beta = -.180, p = .024$) demonstrated a significant relationship to the positive testing effect. Also, age was marginally related to positive testing effects ($\beta = -.140, p = .059$). Positive testing effects increased with increasing retention intervals, was higher for participants who scored higher on the initial recognition test (i.e., had more successful retrieval during learning), was higher with increased previous knowledge of tested relative to restudied topics, was higher for people with lower verbal ability scores, and was marginally higher for younger learners. Verbal ability, however, acted as a suppressor. It had a low correlation with positive testing effects ($r = -.04, p = .54$), but high correlations with initial test performance ($r = .39, p < .001$) and previous knowledge ($r = .188, p = .01$) at the bivariate level. When verbal ability was removed from the regression, the total variance accounted for and the regression weights of final test delay, initial test performance, and previous knowledge decreased (R^2 with verbal ability included = .166, R^2 without verbal ability included = 141; standardized regression weights with verbal ability in regression: $\beta = .154$, $\beta = .277$, and $\beta = .259$, respectively; standardized regression weights without verbal ability in regression: $\beta = .146$, $\beta = .224$, and $\beta = .234$, respectively). When verbal ability is included in the model, it is suppressing the parts of the aforementioned factors that are not related to positive testing effects, resulting in a better model with increased standardized regression weights.

Regression of variables onto negative testing effect.

The negative testing effect was regressed onto age group (coded as 0= younger adults aged 18-25 and 1= older adults aged 55-65), final test retention interval (coded as 0= 5 minute delay and 1= 2 day delay), initial test performance, previous knowledge (tested - restudied), source memory performance, verbal ability (raw score on vocabulary section of the intelligence test), and nonverbal reasoning ability (raw score on matrix reasoning section of the intelligence test). This analysis shows which individual differences were related to the size of a participant's negative testing effect, which will uncover which individual differences lead to differing lure production as a result of testing.

The model accounted for about 26% of the variance ($R^2 = .262$) in the size of the negative testing effect, which was significant, $F(7, 172) = 8.72, p < .001$. Initial test performance ($\beta = -.535, p < .001$) and previous knowledge ($\beta = -.142, p = .04$) demonstrated a significant relationship to the negative testing effect. Additionally, final test retention interval ($\beta = -.127, p = .07$), verbal ability ($\beta = .136, p = .07$), and nonverbal reasoning ability ($\beta = .137, p = .08$) were marginally related to negative testing effects. Source memory performance ($\beta = -.074, p = .35$), on the other hand, was not significantly related to negative testing effects. Relatively poor initial performance increased the size of participants' negative testing effects. This agrees with previous research on the negative testing effect in which poorer performers in the learning phase showed increased rates of lure production after testing (Marsh, Agarwal, & Roediger, 2009). The previous knowledge predictor was also negatively related to negative testing effects, signifying that knowing more about topics going into a test may reduce later lure production. Final test retention interval was marginally negatively related to negative testing effects, meaning that learners actually showed fewer negative testing effects with an increasing retention interval (this is in opposition to the ANCOVA findings). This may be because the lures (along with their source) are forgotten more over a longer retention interval (mimicking normal forgetting), meaning their effect may be short-lived. Verbal ability was marginally related to negative testing effects such that those with higher verbal ability incurred higher negative testing effects. It is not clear why this would occur, but, perhaps, people with higher verbal abilities are better able to process the answers and the lures, making it easier for them to incorporate lures into their knowledge. Nonverbal reasoning ability was also marginally positively related to negative testing effects, such that those with higher nonverbal reasoning scores incurred higher negative testing effects. Previous research shows that the lures that tend to persist are those that are reasoned about (Marsh, Roediger, Bjork, & Bjork, 2007). As such, perhaps people with better reasoning abilities are reasoning too much or too effectively about the wrong answers, making them more susceptible to the negative effects of testing.

Correlation of all variables.

To see how all of the variables were related, a Pearson correlation matrix was crafted (see Table 2). Notable significant correlations existed between positive testing effects and retention interval, initial recognition test performance, and previous knowledge of tested relative to restudied topics and between negative testing effects and initial recognition test performance and previous knowledge of tested relative to restudied topics. Interestingly, age was not related to either testing effects at the bivariate level, but was related to the previously mentioned factors. So, though age was not significantly related to positive and negative testing effects at the bivariate level, it was related to other factors that were related to the positive and negative effects of testing, meaning age may be indirectly related to positive and negative effects of testing via individual differences. This supports the findings in the previous regressions, that show that age is related to positive and negative testing effects when multiple variables are taken into consideration.

	Positive Testing Effects	Negative Testing Effects	#Age	+Final Test Retention Interval	Initial Test	Source Memory	Previous Knowledge (Tested-Restudied)	Verbal ability (Raw Vocabulary)	Nonverbal Reasoning Ability (Raw Matrix Reasoning)	Combined WASI
Positive Testing Effects	—									
Negative Testing Effects	-.510**	—								
#Age	-.121	.036	—							
+Final Test Retention Interval	.159*	-.078	.000	—						
Initial Test	.194**	-.449**	-.078	-.081	—					
Source Memory	-.091	-.125	-.292**	-.288*	.319**	—				
Previous Knowledge (Tested-Restudied)	.251**	-.171*	.100	.125	.084	.008	—			
Verbal ability (Raw Vocabulary)	-.046	-.080	.035	.035	.393**	.284**	.188*	—		
Nonverbal Reasoning Ability (Raw Matrix Reasoning)	.030	-.084	-.276**	-.020	.433**	.372**	.035	.332**	—	
Combined WASI	-.010	-.101	-.147*	.009	.506**	.402**	.137	.816**	.816**	—

Table 2. Pearson correlation matrix of all variables.
**p < .01, *p < .05, #Age: Coded as 0 = younger 18-25 year olds; 1 = older 55-65 year olds, +Final Test Retention Interval: Coded as 0 = 5 minute delay, 1 = 2 day delay.

Summary of Results

Correct memory performance and the positive testing effect.

Looking at the final test performance, positive testing effects were found for both retention intervals, meaning that testing in the experimental learning phase resulted in better final test performance than restudying did for both 5 minute and 2 day delays. Additionally, this benefit increased with an increasing retention interval (i.e., the longer the interval, the more of a benefit was seen for testing compared to restudying). Interestingly, overall performance on this final test was not affected by age when controlling for IQ using the ANCOVA and age did not interact with any of the other variables. This means that older adults benefited from testing just like younger adults did.

Lure production and the negative testing effect.

Lure production increased with increasing retention intervals, but no negative testing effect was found, meaning learners produced as many final test lures after restudying (so, by chance) as they did after testing in the experimental learning phase. Surprisingly, older adults and younger adults from both Rice and from the community all produced a similar number of lures in all conditions (retention intervals and experimental learning conditions) when comparing group means.

Regression of variables onto positive testing effect.

The size of learners' positive testing effects was positively related to retention interval, initial performance, and previous knowledge and was marginally and negatively related to age, meaning older adults benefitted less from testing than younger adults did (though they still benefited).

Regression of variables onto negative testing effect.

The size of learners' negative testing effects was negatively related to initial performance and previous knowledge, meaning that higher initial performance and more previous knowledge is related to decreased negative effects of testing. Also, negative testing effects were marginally related to final test retention interval in the negative direction, meaning that longer retention intervals were associated with fewer negative testing effects. Lastly, verbal ability and nonverbal reasoning ability were marginally related to negative testing effects in the positive direction, meaning that increased verbal and nonverbal reasoning ability led to higher negative testing effects.

Overall findings.

Age was not found to be significantly related to the presence of either positive or negative testing effects when controlled for intelligence using ANCOVAs, but age was related to the size of the positive testing effects when multiple individual differences were taken into account using multiple regressions. Furthermore, initial performance and previous knowledge were positively related to positive testing effects, while initial performance and previous knowledge were negatively related to negative testing effects, and verbal ability and nonverbal reasoning ability were positively related to negative testing effects.

Discussion

Given the prevalence of adult learning (Creighton & Hudson, 2002), this research sought to examine the benefits and costs associated with a highly touted learning technique--testing--and its use in both younger and older learners alike. Up until now it had only been "tested" on younger learners and had not been examined as a learning tool in older adults. As such, this study examined both positive and negative effects of testing in both younger and older adults to fill this important gap in the literature and to inform and affect both educators and learners.

Previous research indicated that younger adults accrue both benefits and costs as a result of testing (Butler & Roediger, 2008; Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008; Kuo & Hirshman, 1996; Roediger, Agarwal, Kang & Marsh, 2010; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978). Testing improves memory for the to-be-remembered information, but it also increases memory for wrong information that was only presented as an incorrect alternative on a test. The prevailing theory about why testing is beneficial centered around variable processing that occurs with successful retrieval during testing and the notion that successful retrieval in a testing situation increases the number of cues available for later retrieval (since they differ from those at study) (Carpenter & DeLosh, 2006; Glover, 1989; Toppino & Cohen, 2009). If this is true, older learners may not benefit as much from testing. The reasons for this include: poorer long-term episodic memory (Park, 2001) and increased rates of forgetting (Wheeler, 2000). Additionally, older adults might acquire more negative testing effects due to associative and source memory problems and long-term episodic memory impairments (Dywan & Jacoby, 1990; Hashtroudi, Johnson, & Lindsay, 1989; Johnson, Hashtroudi, & Lindsay, 1993; McIntyre & Craik, 1987; Mitchell, Johnson, & Mather, 2003; Naveh-Benjamin, 2000; Park, 2001; Rabinowitz, 1989). Given the deficits that are often seen in older adults, it was reasonable to hypothesize that older adults would not benefit as much (or even at all) from testing as a learning technique.

The Positive Testing Effect

In opposition to the aforementioned hypotheses, this current study shows that older adults *do* benefit from testing compared to restudying material (i.e., show positive testing effects). Specifically, when younger and older learners restudied some topics and took a practice test on other topics during learning, their performance on a final test was much better on the previously tested topics than on the restudied topics. The increased performance of tested over restudied items (or the positive testing effect) was found for all age groups at both retention intervals. So, the main inquiry of this study, which was "do older adults benefit from testing similarly to younger adults," is a clear yes (at least for this sample of educated, older adults). To emphasize this point even more, when looking at individuals, 148 of the 180 people tested (82.2%) incurred a positive testing effect. Looking at each age group tested, 48 of the 60 older adults from the community (80.0%), 48 of the 60 younger adults from the community (80.0%), and 52 of the 60 younger adults from Rice University (86.7%) showed positive testing effects and the average positive testing effect was almost a 17% difference (almost two letter grades in education!) ($M = 16.6$, $SEM = 1.22$). This is great news for older adults in education; as the extant literature suggests for younger adults (Butler & Roediger, 2008; Carrier & Pashler, 1992; Glover, 1989; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008; Kuo & Hirshman, 1996; Roediger, Agarwal, Kang & Marsh, 2010; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978; Toppino & Cohen, 2009), older adults should also be given or should take the opportunity to test themselves on material during learning. This is also good news to educators: educators can continue to tout the benefits of testing for all ages of students.

An additional finding concerning the positive testing effect in this study is the observation that positive testing effects were found after both a 5 minute delay and a 2 day delay. Previous reports show that testing effects are only found after a longer delay (Roediger & Karpicke, 2006a). Roediger and Karpicke found positive testing effects only after 1 or 2 days, but not after 5 minutes. The differences may arise from different distractor periods inserted after initial studying or encoding of the studied material. Roediger and Karpicke had learners solve multiplication problems for 2 minutes in between the initial reading of the material and the subsequent learning phase where learners restudied some material and took a test on other material. The current study, however, had learners solve multiplication problems for 5 minutes between the initial reading of the articles and the subsequent experimental learning phase. Although the difference between 2 and 5 minutes might seem inconsequential, the extra time inserted into the distractor task in the current study may have altered the results such that the benefits of testing already appeared after the short retention interval of 5 minutes. This may be because testing is supposed to engage retrieval processes, and after 2 minutes, perhaps the participants in Roediger's study did not have to retrieve the information from long-term memory, but still had the information in their working memory (assuming they were able to rehearse the information during the 2 minutes and assuming that participants in the current study would be less able to do so after 5 minutes of distraction). If this is the case, the implications are that testing can be beneficial after a very short amount of time and the benefits of restudying are even more cursory than what was previously thought.

Perhaps more interesting than the finding that positive testing effects were found at both retention intervals is the finding that older adults showed similar positive testing effects as younger adults at both intervals, meaning they did not show an increased amount of forgetting over the longer delay after being tested in the learning phase. Testing seemed to help both younger and older adults thwart forgetting by inducing retrieval processes. The fact that older adults benefit from testing similarly to younger adults in this experiment may be that older adults often perform well when the material is relevant to them (unlike in many studies that employ the use of random, context-less word lists) and also when effective retrieval operations are induced by the task. This study used study materials that older adults may be somewhat familiar with or at least interested in (e.g., the human heart, tsunamis, etc.), increasing the chances that they were able to relate the material to things they already knew and that they engaged appropriate encoding processes during study. Previous studies have also shown the benefit of meaningful context to memory with older adults (Castel, 2005). For example, Castel (2005) found that older adults remembered less of the prices associated with grocery items when the prices were arbitrary when compared to younger adults, but they remembered just as much as younger adults did when prices were not arbitrary allowing them to rely on their previous knowledge. Additionally, the use of a recognition test during the learning phase instead of a cued-recall or free-recall test may have helped lead them toward successful retrieval operations (and simultaneously, more variable processing) as asserted by Rabinowitz and Craik (1986). If they had been given a cued-recall or free-recall test during the learning phase, they might not have had as effective retrieval operations, resulting in smaller positive testing effects.

The main aim of this research was to see if older learners benefit from testing. Though the answer to this is yes, subsequent analyses also showed how various individual differences, *including age*, can be related to how much a person shows these positive testing effects. The regression analyses showed that in addition to retention interval (as discussed above), age, learners' initial performance, and previous knowledge about the topics were significantly related to the size of their positive testing effects.

First off, age was found to be significantly related to positive testing effects in the regression. Specifically, when one accounted for multiple individual differences, age *was* related to positive testing effects, such that increased age was associated with decreased positive testing effects. Though the regression included initial performance and source memory performance, these factors could not account for all of the differences between younger and older adults. As such, additional age-related factors are likely accounting for this difference. Possibilities include slowed processing seen in older adults, decreased inhibitory abilities, or decreased myelination and general atrophy, resulting in weaker connections between ideas and memories (Balota, Dolan, & Duchek, 2000; Park, Polk, Mikels, Taylor, & Marshuetz, 2001). If older adults have slowed processing, perhaps they need more time to do the experimental tasks. On the other hand, only one significant part of the experiment was timed- the initial reading of the articles. The rest of the experiment (not including distractor tasks) was un-timed, so this seems unlikely to be a major factor. Older adults also show inhibitory abilities and perhaps they have trouble inhibiting incorrect answers. If this were true, however, one would expect them to show increased negative testing effects (which they do not). A factor that may explain the discrepancy may be decreased neural myelination and general atrophy. This may decrease the benefit seen from testing. If testing is supposed to enhance the memory, presumably through the strengthening of neural connections, perhaps older adults need more testing to see the same benefits that younger adults do or perhaps the strengthening is not as strong in older adults. Future studies could address this question. However, despite this significant finding, the news is still good, since most older adults still benefited from testing compared to restudying (and in fact, a similar percentage of older adults and younger adults benefitted from testing compared to restudying). So, though the benefit may be slightly lower than that incurred by younger adults, testing can still be a beneficial study technique for older adults.

Besides age, initial performance was also related to the size of one's positive testing effects, though in a positive direction. This is consistent with theories about the source of the positive testing effect. The benefit of testing is thought to arise from variable processing via successful retrieval during test-taking. The more success learners have in retrieving during initial testing, the more they should show a benefit over merely restudying material. Indeed, the better learners in the current study did on the multiple-choice test, the more they showed a positive effect of testing. Some may argue that multiple-choice testing is not retrieval (since it is a recognition test), but some retrieval may still occur after a participant reads the question and before they read all of the answer choices. Additionally, recall may occur for material when reading questions or answers to related material (in fact, researchers have found positive testing-like effects for material that was not tested, when related material was tested, e.g., Chan, McDermott, & Roediger, 2006). Nevertheless, larger positive testing effects would likely be seen in cases where subjects are given recall rather than recognition tests during learning (Roediger, Agarwal, Kang, & Marsh, 2010). Such materials would not enable one to look at negative effects of testing, however, which was a major goal of the current study. Regardless, this finding does lead to a recommendation to students about the educational use of testing: study enough before a test, such that you can successfully retrieve information, thereby increasing the availability of cues and increasing your ultimate retention of the learned information. Again, testing will lead to longer-term retention than repetitive studying will, which is the ultimate goal of education, but testing will not lead to these gains in long-term retention on its own. Testing will induce one to process the information in a way different than studying would, but only if the information is successfully retrieved. In order to have this successful retrieval, initial studying is important. In other words, without studying beforehand, the benefits of testing will be less as successful retrieval is limited.

In addition to age and initial performance, another variable that was related to the size of one's positive testing effect was previous knowledge. Specifically, the more a person knew about the tested topics relative to the restudied topics, the more of a positive testing effect they showed. Perhaps this is because previous knowledge about a topic allows one to more accurately retrieve during the initial test, which leads to a higher positive testing effect. A few other explanations also exist. First, if one already knows much about a topic, learning becomes easier because one can more easily make associations between what one knows and what one is learning. Secondly, it could be that if a person happens to come into the study knowing a lot about the topics that he or she was randomly assigned to take a test on during the learning phase, but little to nothing about the topics that he or she was randomly assigned to restudy during the learning phase, it would appear that testing helped him or her more than restudying did on the final test. Although testing may have still helped him or her relative to restudying, the effect would be exaggerated in this case. The good news about this last explanation is that articles were randomly assigned to conditions for each participant, and it is equally likely that a participant was tested on topics they knew little about and restudied topics they knew a lot about or even that participants knew similar amounts about the topics in each condition. In fact, when looking at group means, one can see that participants knew similar amounts about the topics from each experimental learning condition coming into the experiment: participants answered 7.44% of the previous knowledge questions for tested topics correctly and 7.94% of the previous knowledge questions for the restudied topics correctly ($t(179) = -.52$, $p = .61$). As such, it is unlikely that all 148 of the 180 subjects who showed a positive testing effect were randomly assigned articles for each condition in such a way that this would account for all of the positive testing effects (and, in fact, previous knowledge was not the only significant factor related to positive testing effects in the regression).

The Negative Testing Effect

Now, turning the discussion towards the negative effects of testing, this study shows that in this paradigm, though lure production increases with increasing retention intervals, neither younger nor older adults reliably accrued negative effects of testing (at least, when looking at the overall group means), which again goes in opposition to aforementioned hypotheses. This means that people are no more likely to produce previously seen wrong answers after taking a test during learning than they would produce those same wrong answers randomly after restudying (i.e., when they are not exposed to the wrong answers). In fact, only 75 out of the 180 participants (41.7%) showed a negative testing effect and the average negative testing effect was practically zero with a very low standard error ($M = 0.88$, $SEM = 0.68$). This includes 29 out of the 60 older adults, 23 out of the 60 younger adults from the community, and 23 out of the 60 younger adults from Rice University. The other half of the participants showed the opposite pattern: they produced fewer lures after being tested on material in the learning phase than they did randomly after restudying.

Although participants did not show reliable negative testing effects in terms of group means, there were individual factors that related to these negative effects of testing. The regression shows that initial performance and previous knowledge were reliably related to negative effects of testing in the negative direction. Additionally, verbal ability and nonverbal reasoning ability were marginally related to negative testing effects in the positive direction.

Initial performance was negatively related to negative testing effects, such that people who did worse on the initial multiple-choice test produced more lures seen on that multiple-choice test than they did randomly after studying. This makes sense and is consistent with previous reports (Marsh, Agarwal, & Roediger, 2009). Although the participants received feedback about their incorrect answers (and their correct answers), they likely spent more time considering the incorrect alternatives on questions that they missed than on questions that they did not miss. When people got the answer correct (as a mere observation), they seemed to not even reread the question but skip ahead to the questions that they answered incorrectly. Furthermore, the more questions a person missed on the initial multiple-choice test, the more questions they seemed to go back to, exposing themselves once again to the incorrect answers. If participants were required to reread all questions and view feedback, whether they answered correctly or not, one might see increased negative testing effects and this may be why some studies show negative testing effects and some (like this one) do not reliably show them for all participants. However, this finding is important because it points out that though negative testing effects may not be seen for all students, they may be seen for those who do worse in the first place, placing them at an additional disadvantage. Furthermore, this finding relates to the potential hazard of using pre-tests in education. Though they provide diagnostic information about what one ought to study in the future, it also exposes students to wrong information that they may have trouble overriding with the correct information. Future research could be done to scrutinize what role exposure to wrong answers plays in attaining negative testing effects apart from answering incorrectly.

Previous knowledge also appears to be related to the size of one's negative testing effect. This relationship is also negative, such that higher previous knowledge of material tested in the learning phase relative to material restudied in the learning phase leads to smaller negative testing effects. Knowing more about a topic before taking a test seems to help one avoid taking in the wrong information presented on the multiple-choice test and using it later. This finding leads to a recommendation made earlier in the discussion section that learners should try to be prepared before taking a test, because doing well on tests leads to both increased positive testing effects and decreased negative testing effects.

Additionally, verbal ability and nonverbal reasoning ability were positively, yet marginally related to negative testing effects. It is not clear why people with higher verbal and nonverbal abilities would incur more negative testing effects. Perhaps those with higher verbal abilities are better able to comprehend and understand all of the lures presented, making it easier for them to take them in and incorporate them into their knowledge. Also, perhaps those with higher reasoning abilities reason too much about the incorrect answers, increasing memory for lures. Previous researchers have shown that lures that are reproduced are usually the ones people reasoned about. As such, it may be that people with higher reasoning abilities are actually hurting themselves by forcing themselves to reason about incorrect answers.

Notably missing from the variables that are related to the negative testing effect are both source memory and age. Faulty source memory was postulated to be a reason why people may erroneously reproduce lures after seeing them on a multiple-choice test. In this study, however, source memory performance was not related to negative testing effects. This may have occurred for several reasons. For one, perhaps source memory does not affect lure production. Memory for items does not necessitate memory for an item's source. Also, without knowing the learner's reasons for answering with lures, it is impossible to know whether learners simply put down lures when they could not think of anything else or if they indeed misattributed the source of the information. Learners might have responded with lures- not because they thought it was the right answer, but because it was the only thing they could think of and they did not want to leave an answer blank. On the other hand, though learners in this experiment were simply told to do their best with no instructions to guess or leave answers blank, previous research shows that explicit instructions warning against guessing still leads to lure reproduction (Roediger & Marsh, 2005). Of course, people may have difficulty overriding this ingrained test-taking strategy, since most academic testing situations do not penalize for guessing (unlike the SAT or Scholastic Aptitude Test).

Another reason that source memory may not be related to negative testing effects in this study is that if participants thought about which two articles they reread and which two articles they were tested on in the learning phase, they could easily answer the source memory questions even if they did not remember where they saw certain pieces of information. It seems unlikely that many subjects realized this, because the average performance on the source memory questions was 50.6% (*SEM* = 1.57), which is far above chance (which would have been 25%). Ideally, a study would be set up so that participants could reread parts of topics and test over other parts of it, but this would make the logistics of the experiment very difficult, so that participants cannot simply recall which topics they were previously tested on and which ones they restudied. To scrutinize the relationship between source memory and negative testing effects in the future, it would be ideal to either test and restudy material within topics or to include a separate source memory task in the future that is not associated with the rest of the study.

Besides source memory, age was not significantly related to the size of negative testing effects. This was very surprising, but is great news for older learners. However, age did correlate to the other factors that related to negative testing effects, including initial performance and previous knowledge. As such, older adults are not totally in the clear. They may still be more susceptible to negative testing effects. Negative testing effects in this particular study, however, were scarce, so one should definitely take that in to consideration before hesitating to use testing as a learning tool in fear of incurring negative testing effects.

Implications

Testing has been researched and utilized as a learning tool, but prior to now it had only been examined in the population of younger, college students and faintly in high school students. This research adds to that literature by considering cases where older students are the recipients of the method and has found that testing is also effective for learning in this sample. As such, teachers of older students and older students themselves can more justifiably utilize it as a learning technique. Importantly, this research has the potential to affect the large population of older students and their lifelong education. Furthermore, the results show how individual differences, like age, intelligence, and previous knowledge, alter testing's effectiveness and can be beneficial when planning study strategies for students on an individual basis.

As a result of the research above, various specific recommendations can be made in regards to the use of testing in education. To begin with, educators should use tests as learning tools, not just as diagnostic tools, in students both young and old in both early education and lifelong learning. Furthermore, they should take steps to ensure that their students (especially those likely to have poorer initial performance) are adequately prepared for the learning tests (i.e., make sure that they have thoroughly read and understand the materials) before administering them perhaps by making practice tests available to students at their leisure and stressing the importance of them taking it only when prepared. This will decrease the chances that the students will be overly exposed to wrong information and that they take the information and incorporate it into their knowledge. Additionally, if an educator feels the need to use a pre-test for diagnostic purposes (i.e., to determine what areas to focus their teaching on or to determine what areas to tell their students to focus their studying on), the diagnostic test should not provide incorrect information to the students. As such, it should be a cued-recall or free-recall test and not a recognition test. Using free-recall or cued-recall tests will decrease the negative testing effects acquired by exposing students to wrong information during learning. This particular point is also relevant to how students study in groups where quizzing may occur. Students who attend group study sessions, but have not prepared in advance, may inadvertently obtain incorrect knowledge from their peers if their peers misspeak or are simply incorrect as has been seen in previous research (Meade & Roediger, 2009; Roediger, Meade, & Bergman, 2001). This case would be especially hazardous if no one is around to provide corrective feedback. As such, educators should teach students about the positive and negative effects of taking practice tests, so that they use it to their utmost advantage.

Appendices

Appendix A

Study Materials

Tsunamis: Killer Waves

A tsunami is a series of ocean waves that sends surges of water, sometimes reaching heights of over 100 feet (30.5 meters), onto land. These walls of water can cause widespread destruction when they crash ashore.

These awe-inspiring waves are typically caused by large, undersea earthquakes at tectonic plate boundaries. When the ocean floor at a plate boundary rises or falls suddenly it displaces the water above it and launches the rolling waves that will become a tsunami.

Most tsunamis, about 80 percent, happen within the Pacific Ocean's "Ring of Fire," a geologically active area where tectonic shifts make volcanoes and earthquakes common.

Tsunamis may also be caused by underwater landslides or volcanic eruptions. They may even be launched, as they frequently were in Earth's ancient past, by the impact of a large meteorite plunging into an ocean.

Tsunamis race across the sea at up to 500 miles (805 kilometers) an hour—about as fast as a jet airplane. At that pace they can cross the entire expanse of the Pacific Ocean in less than a day. And their long wavelengths mean they lose very little energy along the way.

In deep ocean, tsunami waves may appear only a foot or so high. But as they approach shoreline and enter shallower water they slow down and begin to grow in energy and height. The tops of the waves move faster than their bottoms do, which causes them to rise precipitously.

A tsunami's trough, the low point beneath the wave's crest, often reaches shore first. When it does, it produces a vacuum effect that sucks coastal water seaward and exposes harbor and sea floors. This retreating of sea water is an important warning sign of a tsunami, because the wave's crest and its enormous volume of water typically hit shore five minutes or so later. Recognizing this phenomenon can save lives.

A tsunami is usually composed of a series of waves, called a wave train, so its destructive force may be compounded as successive waves reach shore. People experiencing a tsunami should remember that the danger may not have passed with the first wave and should await official word that it is safe to return to vulnerable locations.

Some tsunamis do not appear on shore as massive breaking waves but instead resemble a quickly surging tide that inundates coastal areas.

The best defense against any tsunami is early warning that allows people to seek higher ground. The Pacific Tsunami Warning System, a coalition of 26 nations headquartered in Hawaii, maintains a web of seismic equipment and water level gauges to identify tsunamis at sea. Similar systems are proposed to protect coastal areas worldwide.

Black Holes: A Mighty Void

Black holes are the cold remnants of former stars, so dense that no matter—not even light—is able to escape their powerful gravitational pull.

While most stars end up as white dwarfs or neutron stars, black holes are the last evolutionary stage in the lifetimes of enormous stars that had been at least 10 or 15 times as massive as our own sun.

When giant stars reach the final stages of their lives they often detonate in cataclysms known as supernovae. Such an explosion scatters most of a star into the void of space but leaves behind a large "cold" remnant on which fusion no longer takes place.

In younger stars, nuclear fusion creates energy and a constant outward pressure that exists in balance with the inward pull of gravity caused by the star's own mass. But in the dead remnants of a massive supernova, no force opposes gravity—so the star begins to collapse in upon itself.

With no force to check gravity, a budding black hole shrinks to zero volume—at which point it is infinitely dense. Even the light from such a star is unable to escape its immense gravitational pull. The star's own light becomes trapped in orbit, and the dark star becomes known as a black hole.

Black holes pull matter and even energy into themselves—but no more so than other stars or cosmic objects of similar mass. That means that a black hole with the mass of our own sun would not "suck" objects into it any more than our own sun does with its own gravitational pull.

Planets, light, and other matter must pass close to a black hole in order to be pulled into its grasp. When they reach a point of no return they are said to have entered the event horizon—the point from which any escape is impossible because it requires moving faster than the speed of light.

Small But Powerful: Black holes are small in size. A million-solar-mass hole, like that believed to be at the center of some galaxies, would have a radius of just about two million miles (three million kilometers)—only about four times the size of the sun. A black hole with a mass equal to that of the sun would have a two-mile (three-kilometer) radius.

Because they are so small, distant, and dark, black holes cannot be directly observed. Yet scientists have confirmed their long-held suspicions that they exist. This is typically done by measuring mass in a region of the sky and looking for areas of large, dark mass.

Many black holes exist in binary star systems. These holes may continually pull mass from their neighboring star, growing the black hole and shrinking the other star, until the black hole is large and the companion star has completely vanished.

Extremely large black holes may exist at the center of some galaxies—including our own Milky Way. These massive features may have the mass of 10 to 100 billion suns. They are similar to smaller black holes but grow to enormous size because there is so much matter in the center of the galaxy for them to add. Black holes can accrue limitless amounts of matter; they simply become even denser as their mass increases.

Black holes capture the public's imagination and feature prominently in extremely theoretical concepts like wormholes. These "tunnels" could allow rapid travel through space and time—but there is no evidence that they exist.

The Human Heart

The heart is the body's engine room, responsible for pumping life-sustaining blood via a 60,000-mile-long (97,000-kilometer-long) network of vessels. The organ works ceaselessly, beating 100,000 times a day, 40 million times a year—in total clocking up three billion heartbeats over an average lifetime. It keeps the body freshly supplied with oxygen and nutrients, while clearing away harmful waste matter.

The fetal heart evolves through several different stages inside the womb, first resembling a fish's heart, then a frog's, which has two chambers, then a snake's, with three, before finally adopting the four-chambered structure of the human heart.

About the size of its owner's clenched fist, the organ sits in the middle of the chest, behind the breastbone and between the lungs, in a moistened chamber that is protected all round by the rib cage. It's made up of a special kind of muscle (cardiac muscle) that works involuntarily, so we don't have to think about it. The heart speeds up or slows down automatically in response to nerve signals from the brain that tell it how much the body is being exerted. Normally the heart contracts and relaxes between 70 and 80 times per minute, each heartbeat filling the four chambers inside with a fresh round of blood.

These cavities form two separate pumps on each side of the heart, which are divided by a wall of muscle called the septum. The upper chamber on each side is called the atrium. This is connected via a sealing valve to the larger, more powerful lower chamber, or ventricle. The left ventricle pumps most forcefully, which is why a person's heartbeat is felt more on the left side of the chest.

When the heart contracts, the chambers become smaller, forcing blood first out of the atria into the ventricles, then from each ventricle into a large blood vessel connected to the top of the heart. These vessels are the two main arteries. One of them, the pulmonary artery, takes blood to the lungs to receive oxygen. The other, the aorta, transports freshly oxygenated blood to the rest of the body. The vessels that bring blood to the heart are the veins. The two main veins that connect to the heart are called the vena cava.

Blood Delivery: Since the heart lies at the center of the blood delivery system, it is also central to life. Blood both supplies oxygen from the lungs to the other organs and tissues and removes carbon dioxide to the lungs, where the gas is breathed out. Blood also distributes nourishment from the digestive system and hormones from glands. Likewise our immune system cells travel in the bloodstream, seeking out infection, and blood takes the body's waste products to the kidneys and liver to be sorted out and trashed.

Given the heart's many essential functions, it seems wise to take care of it. Yet heart disease has risen steadily over the last century, especially in industrialized countries, due largely to changes in diet and lifestyle. It has become the leading cause of death for both men and women in the United States, claiming almost 700,000 lives a year, or 29 percent of the annual total. Worldwide, 7.2 million people die from heart disease every year.

Armadillo: Dasypodidae

Of the 20 varieties of armadillo, all but one live in Latin America. The familiar nine-banded armadillo is the only species that includes the United States in its range.

Armadillo is a Spanish word meaning "little armored one" and refers to the bony plates that cover the back, head, legs, and tail of most of these odd looking creatures. Armadillos are the only living mammals that wear such shells.

Closely related to anteaters and sloths, armadillos generally have a pointy or shovel-shaped snout and small eyes. They vary widely in size and color, from the 6-inch-long (15-centimeter-long), salmon-colored pink fairy armadillo to the 5-foot-long (1.5-meter-long), dark-brown giant armadillos. Others have black, red, gray, or yellowish coloring.

Contrary to popular belief, not all armadillos are able to encase themselves in their shells. In fact, only the three-banded armadillo can, curling its head and back feet and contorting its shell into a hard ball that confounds would-be predators.

Armadillos live in temperate and warm habitats, including rain forests, grasslands, and semi-deserts. Because of their low metabolic rate and lack of fat stores, cold is their enemy, and spates of intemperate weather can wipe out whole populations.

Most species dig burrows and sleep prolifically, up to 16 hours per day, foraging in the early morning and evening for beetles, ants, termites, and other insects. They have very poor eyesight, and utilize their keen sense of smell to hunt. Strong legs and huge front claws are used for digging, and long, sticky tongues for extracting ants and termites from their tunnels. In addition to bugs, armadillos eat small vertebrates, plants, and some fruit, as well as the occasional carrion meal.

Population numbers of nearly all species are threatened by habitat loss and over-hunting. Many cultures in the Americas consume armadillo flesh, which is said to resemble pork in its flavor and texture. Currently, only the nine-band population is expanding, and some species, including the pink fairy, are threatened.

Appendix B

Participant ___________

Math Problems 1

Work on the following math problems until the experimenter tells you to stop.

33 x 24 ____	35 x 27 ____	25 x 11 ____	31 x 13 ____	44 x 12 ____
12 x 11 ____	12 x 11 ____	16 x 13 ____	40 x 34 ____	24 x 10 ____
13 x 12 ____	36 x 21 ____	14 x 12 ____	39 x 18 ____	35 x 24 ____
24 x 14 ____	32 x 19 ____	26 x 23 ____	13 x 12 ____	41 x 36 ____
33 x 19 ____	40 x 24 ____	43 x 29 ____	32 x 17 ____	31 x 18 ____

45	47	40	22	18
x 31	x 34	x 28	x 16	x 13
46	28	25	20	18
x 31	x 24	x 18	x 18	x 17
30	17	23	49	36
x 27	x 12	x 12	x 13	x 11
50	44	30	37	19
x 12	x 24	x 29	x 34	x 16
45	30	39	17	11
x 23	x 22	x 31	x 31	x 10

Participant ___________

Math Problems 2

Work on the following math problems until the experimenter tells you to stop.

20 x 13	23 x 10	41 x 19	11 x 10	36 x 31
22 x 16	22 x 11	28 x 19	36 x 35	34 x 32
14 x 13	44 x 34	13 x 11	48 x 21	13 x 11
23 x 20	40 x 21	18 x 14	18 x 16	30 x 28
17 x 13	15 x 12	26 x 10	37 x 24	50 x 20

46 x 28	43 x 36	48 x 27	34 x 19	44 x 15
24 x 17	18 x 15	20 x 19	37 x 17	41 x 22
46 x 24	23 x 19	43 x 37	47 x 27	42 x 18
28 x 20	34 x 30	40 x 19	34 x 15	24 x 11
21 x 12	12 x 11	24 x 11	38 x 32	43 x 13

Appendix C

Participant ___________

Version _MASTER_

Multiple-choice Questions

For this section, choose the answer choice that best answers each question according to the readings you just did.

1. How high can tsunami waves reach on land?
   a. 100 feet
   b. 75 feet
   c. 200 feet
   d. 150 feet

2. What typically causes a tsunami?
   a. ocean tides
   b. island hurricanes
   c. tornadoes
   d. undersea earthquakes

3. What percentage of tsunamis occurs near the "Ring of Fire"?
   a. 25%
   b. 40%
   c. 80%
   d. 60%

4. What else can cause a tsunami?
   a. thunderstorms
   b. volcanic eruptions
   c. hurricanes
   d. underwater warfare

5. How fast can tsunami's move across the ocean?
   a. 500 miles an hour
   b. 300 miles an hour
   c. 200 miles an hour
   d. 100 miles an hour

6. What happens to a tsunami's energy level as it moves?
   a. it stays consistent
   b. it increases
   c. it decreases a little
   d. it is all lost

7. In a tsunami, which part of the wave moves faster?
   a. both parts move at the same speed
   b. the bottoms move faster than the tops
   c. the tops moves faster than the bottoms
   d. it varies

8. What is an important warning sign of a coming tsunami?
   a. stranded sea life
   b. disintegrating sea floors
   c. thunderstorm
   d. retreating sea water

9. Tsunamis are usually made up of a series of waves called a
   a. wave train
   b. string of waves
   c. wave runner
   d. wave surge

10. Where is the headquarters for the Pacific Tsunami Warning System?
    a. the Philippines
    b. New Zealand
    c. Japan
    d. Hawaii

11. What did black holes used to be?
    a. moons
    b. planets
    c. asteroids
    d. stars

12. Before becoming black holes, the matter black holes come from are approximately how much bigger than our sun?
    a. 5 times
    b. 10 times
    c. 20 times
    d. 50 times

13. What can escape the gravitational pull of a black hole?
    a. other black holes
    b. light
    c. nothing
    d. whole galaxies

14. The last stage before a black hole is formed, a detonation occurs, known as a(n)
    a. starburst
    b. explosion
    c. blastula
    d. supernovae

15. In the dead remnants of the detonation before a black hole is formed, the matter begins to
    a. collapse in upon itself
    b. create energy
    c. create both outward and inward pressure
    d. increase in size

16. What is the volume of a black hole?
    a. it varies
    b. a positive value
    c. a negative value
    d. zero

17. Compared to other cosmic objects of a similar mass, the gravitational force of a black hole is
    a. the same
    b. larger than
    c. smaller than
    d. it varies

18. What is it called when a planet or other object reaches the point at which they can no longer avoid being pulled into a black hole?
    a. returning point
    b. event horizon
    c. interior space
    d. yielding proximity

19. Large black holes exists at the center of
    a. solar systems
    b. planets
    c. the universe
    d. galaxies

20. What is the name of the fictional tunnel that allows one to time-travel through a black hole?
    a. continuum
    b. portal
    c. wormhole
    d. quasars

21. How long is the "network" of vessels that the heart pumps blood through?
    a. 6,000 miles
    b. 60,000 miles
    c. 20,000 miles
    d. 2,000 miles

22. Approximately, how many times a day does a human heart beat?
    a. 24,000 times
    b. 10,000 times
    c. 100,000 times
    d. 240,000 times

23. How many chambers does the developed human heart have?
    a. four
    b. three
    c. two
    d. one

24. The heart is divided by a wall of muscle called the
    a. cardiac muscle
    b. ventricle
    c. atrium
    d. septum

25. Which part of the heart is the most powerful?
    a. the left atrium
    b. the left ventricle
    c. the right ventricle
    d. the septum

26. What takes blood to the lungs to receive oxygen?
    a. the vena cava
    b. the aorta
    c. the pulmonary artery
    d. the veins

27. The heart supplies oxygen and removes ____ from the lungs.
    a. waste products
    b. carbon dioxide
    c. hormones
    d. infection

28. In addition to the liver, what organ helps sort and get rid of the body's waste?
    a. pancreas
    b. kidneys
    c. colon
    d. lungs

29. Heart disease has risen over the last century, especially in
    a. Urban centers
    b. Industrialized countries
    c. Developing Nations
    d. Rural areas

30. In the United States, approximately what percent of people die every year from heart disease?
    a. 29%
    b. 21%
    c. 13%
    d. 37%

31. How many varieties of armadillos are there?
    a. 5
    b. 30
    c. 10
    d. 20

32. Where do the majority of armadillo varieties live?
    a. Australia
    b. Latin America
    c. North America
    d. The Sahara Dessert

33. What does "armadillo" translate to?
    a. little arms
    b. arms of steel
    c. little armored one
    d. steel shell

34. What is the armadillo's nose shaped like?
    a. beak
    b. shovel
    c. pencil
    d. tube

35. Armadillos can be as big as
    a. 5 feet
    b. 2 feet
    c. 9 feet
    d. 7 feet

36. Which type of armadillo can contort itself into a hard ball to thwart predators?
    a. Dark-brown giant
    b. Sloth
    c. Three-banded
    d. Pink fairy

37. What kind of habitats can wipe out whole armadillo populations?
    a. cold
    b. rainy
    c. warm
    d. windy

38. Armadillos mostly eat
    a. fruit
    b. carrion
    c. plants
    d. insects

39. Which type of armadillo is currently threatened?
    a. Nine-banded
    b. Three-banded
    c. Pink Fairy
    d. Dark-Brown Giant

40. Armadillo meat resembles what kind of commonly eaten meat?
    a. **pork**
    b. beef
    c. ham
    d. pastrami

Appendix D

Participant ___________

Version _MASTER_

Fill-In-The-Blank Questions + M/C

For this section, fill in the blank with the correct answer. Then, answer the multiple-choice question that corresponds to the fill in the blank.

1. How high can tsunami waves reach on land? ____________. **(100 feet)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

2. What typically causes a tsunami? ____________. **(undersea earthquakes)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

3. What percentage of tsunamis occurs near the "Ring of Fire"? ____________. **(80%)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

4. What else can cause a tsunami? ____________. **(volcanic eruptions)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

5. How fast can tsunami's move across the ocean? ____________. **(500 mph)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

6. What happens to a tsunami's energy level as it moves? ____________. **(it decreases a little)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

7. In a tsunami, which part of the wave moves faster? ____________. **(the tops move faster than the bottoms)**
   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

8. What is an important warning sign of a coming tsunami? ___________. **(retreating sea water)**

   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

9. Tsunamis are usually made up of a series of waves called a ___________. **(wave train)**

   Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

   a. readings
   b. previous test
   c. readings AND previous test
   d. previous knowledge outside of the experiment
   e. I'm just guessing

10. Where is the headquarters for the Pacific Tsunami Warning System? ___________. **(Hawaii)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

11. *"Tsunami" literally translates to ___________. **(harbor wave)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

12. *Approximately how many tsunamis have been recorded in human history? ____________. **(195)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

13. *A type of storm surge that closely resembles (but is not) a tsunami is known as ____________. **(meteotsunami)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

14. *What natural factor prevented a tsunami from destroying a small Indian village? ____________. **(trees)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

15. *Which ocean produced the deadliest tsunami in known history? ____________. **(Indian Ocean)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

16. What did black holes used to be? ____________. **(stars)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

17. Before becoming black holes, the matter black holes come from are approximately how much bigger than our sun? ____________. **(10 times)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

18. What can escape the gravitational pull of a black hole? ____________. **(nothing)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

19. The last stage before a black hole is formed, a detonation occurs, known as a(n) ____________. **(supernovae)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

20. In the dead remnants of the detonation before a black hole is formed, the matter begins to ____________. **(collapse in upon itself)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

21. What is the volume of a black hole? ____________. **(zero)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

22. Compared to other cosmic objects of a similar mass, the gravitational force of a black hole is ____________. **(the same)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

23. What is it called when a planet or other object reaches the point at which they can no longer avoid being pulled into a black hole? ____________. **(event horizon)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

24. Large black holes exist at the center of ____________. **(galaxies)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

25. What is the name of the fictional tunnel that allows one to time-travel through a black hole? ____________. **(wormhole)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

26. *The term "black hole" was first used publicly by ____________. **(John Wheeler)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

27. *Black holes are surrounded by regions where it is impossible to stand still. These are known as ____________. **(ergospheres)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

28. *A thermal radiation predicted to be emitted by black holes is known as ____________. **(Hawking radiation)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

29. *What theory explains the physics behind a black hole? ____________. **(Theory of Relativity)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

30. *The type of black holes with the largest mass is known as ____________. **(supermassive black holes)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

31. How long is the "network" of vessels that the heart pumps blood through? ____________. **(60,000 miles)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

32. Approximately, how many times a day does a human heart beat? ____________. **(100,000 times)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

33. How many chambers does the developed human heart have? ____________. **(four)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

34. The heart is divided by a wall of muscle called the ____________. **(septum)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

35. Which part of the heart is the most powerful? ____________. **(the left ventricle)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

36. What takes blood to the lungs to receive oxygen? ____________. **(the pulmonary artery)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

37. The heart supplies oxygen and removes ____ from the lungs. ____________. **(carbon dioxide)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

38. In addition to the liver, what organ helps sort and get rid of the body's waste? ____________. **(kidneys)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

39. Heart disease has risen over the last century, especially in ____________. **(industrialized countries)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

40. In the United States, approximately what percent of people die every year from heart disease? ____________. **(29%)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

41. *The human heart weighs about ____________. **(300 grams)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

42. *The human heart is enclosed in a double-walled sac known as a ____________. **(pericardium)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

43. * In which process does the heart drop off carbon dioxide and pick up oxygen? ____________. **(diffusion)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

44. *Heart disease is also known as ___________. **(cardiopathy)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

45. *On average, someone dies every from heart disease in the U.S. ___________. **(34 seconds)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

46. How many varieties of armadillos are there? ___________. **(20)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

47. Where do the majority of armadillo varieties live? ___________. **(Latin America)**

Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

48. What does "armadillo" translate to? ____________. **(little armored one)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

49. What is the armadillo's nose shaped like? ____________. **(shovel)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

50. Armadillos can be as big as ____________. **(5 feet)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

51. Which type of armadillo can contort itself into a hard ball to thwart predators? ____________. **(Three-banded)**

    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:

    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

52. What kind of habitats can wipe out whole armadillo populations? ____________.
(cold)
    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

53. Armadillos mostly eat ____________. **(insects)**
    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

54. Which type of armadillo is currently threatened? ____________. **(Pink-fairy)**
    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

55. Armadillo meat resembles what kind of commonly eaten meat? ____________.
(pork)
    Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
    a. readings
    b. previous test
    c. readings AND previous test
    d. previous knowledge outside of the experiment
    e. I'm just guessing

56. *Armadillos can remain underwater for up to ____________. **(six minutes)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

57. *Armadillo skin can be described as ____________. **(leathery)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

58. *How many armadillos typically live together? ____________. **(one)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

59. *Armadillo shells have been used for what leisurely human purpose? ____________. **(musical instrument)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

60. *What disease are armadillos most studied for? ____________. **(leprosy)**
Indicate where you've seen the answer to the previous question (what the source of your memory is) by picking ONE of the following:
a. readings
b. previous test
c. readings AND previous test
d. previous knowledge outside of the experiment
e. I'm just guessing

*Question covering material not studied.

Appendix E

Participant ___________

Follow-Up Questions

1. Circle the one that best describes your level of education.
   a. GED
   b. High school
   c. Some college
   d. Completed college
   e. Vocational school
   f. Graduate school
   g. PhD/ MD/ JD

2. What is your sex? __________

3. Do you know anything about the "testing effect"? __________
   What?

   Do you know anything about the "negative testing effect"? __________
   What?

4. Did you try to look up these topics between experimental sessions? (Omit if your participation occurred in 1 session). _____________ Explain.

5. Do you have any comments/concerns about the current study?

References

Balota, D. A., Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy young and older adults. In F. I. M. Craik (Ed.), *The Oxford Handbook of Memory* (395–409). London: Oxford University Press.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55,* 485-540.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition, 36, 604-616.

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

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

Castel, A. D. (2005). Memory for grocery prices in younger and older adults: The role of schematic support. *Psychology and Aging, 20,* 718-721.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553-571.

craigslist (2010). Online community. Retrieved February 19, 2011, from houston.craigslist.org/

Creighton, S., & Hudson, L. (2002). Participation trends and patterns in adult education. 1991-1999. Statistical Analysis Report #NCES-2002-119. National Center for Education Statistics, Washington, D. C.

Dywan, J., & Jacoby, L. (1990). Effects of aging on source monitoring: Differences in susceptibility to false fame. *Psychology and Aging, 5,* 379-387.

Erngrund, K., Mäntylä, T., & Nilsson, L.-G. (1996). Adult age differences in source recall: A population-based study. *The Journal of Gerontology, 51B,* 335-345.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392-399.

Hashtroudi, S., Johnson, M. K., & Chrosniak, L. D. (1989). Aging and source monitoring. *Psychology and Aging, 4,* 106-112.

Henkel, L. A. (2007).The benefits and costs of repeated memory tests for young and old adults. *Psychology and Aging, 22,* 580-595.

Henkel, L. A. (2008). Maximizing the benefits and minimizing the costs of repeated memory tests for older adults. *Psychology and Aging, 23,* 250-262.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562-567.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

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. *European Journal of Cognitive Psychology, 19,* 528-558.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966-968.

Kuo, T-M, & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451-464.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L., III. (2009). Memorial consequences of answering SAT II questions. Journal of Experimental Psychology: Applied, 15, 1-11.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194-199.

Meade, M. L., & Roediger, H. L., III. (2009). Age differences in collaborative memory: The role of retrieval manipulations. *Memory & Cognition, 37,* 962-975.

McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology, 41,* 175-192.

Mitchell, K. J., Johnson, M. K., & Mather, M. (2003). Source monitoring and suggestibility to misinformation: Adult age-related differences. *Applied Cognitive Psychology, 17,* 107-119.

Myers, G. C. (1914). Recall in relation to retention. *Journal of Educational Psychology, 5,* 119-130.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: tests of an associative deficit hypothesis. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 26,* 1170-1187.

Park, D. C., Polk, T. A., Mikels, J. A., Taylor, S. F., & Marshuetz, C. (2001). Cerebral aging: Integration of brain and behavioral models of cognitive function. *Dialogues in Clinical Neuroscience, 3,* 151-165.

Rabinowitz, J. C. (1989). Judgments of origin and generation effects: Comparisons between young and elderly adults. *Psychology and Aging, 4,* 259-268.

Rabinowitz, J. C., & Craik, F. I. M. (1986). Prior retrieval effects in young and old adults. *Journal of Gerontology, 41,* 368-375.

Roediger, H. L., III, Agarwal, P. K., Kang, S. H. K, & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), New frontiers in applied memory (pp. 13-49). Brighton, U.K.: Psychology Press.

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

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. Journal of Experimental Psychology: Learning, Memory and Cognition, 31, 1155-1159.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology, 9*, 371-377

Sona Systems, Ltd. (2000-2006). Momentum™ Experiment Scheduling System: Web-based experiment scheduling and tracking. Retrieved February 19, 2011, from experimetrix2.com/Rice/

Spitzer, H. F. (1939). Studies in retention. *The Journal of Educational Psychology, 30,* 641-656.

Sproule, C. E. (1934). Suggestion effects of the true-false test. *Journal of Educational Psychology, 25,* 281-285.

SuperKids (1998-2010). Education for the future. Retrieved July 1, 2010, from http://www.superkids.com/

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210-221.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true-false examinations. Journal of Educational Research, 83, 119-124.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56,* 252-257.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. Journal of Educational Psychology, 86, 357-362.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio, TX: Harcourt Assessment.