UNIVERSITY OF CALIFORNIA

Los Angeles

Multiple-choice Tests as Learning Events:

The Role of Desirably Difficult Alternatives

A dissertation submitted in partial satisfaction of the requirements for the

degree Doctor of Philosophy in Psychology

by

Erin Michelle Sparck

2018

ABSTRACT OF THE DISSERTATION

Multiple-choice Tests as Learning Events:

The Role of Desirably Difficult Alternatives

by

Erin Michelle Sparck

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2018

Dr. Elizabeth Ligon Bjork, Co-Chair

Dr. Robert A. Bjork, Co-Chair

Retrieving information is an effective learning strategy to promote the long-term retention of materials (see, e.g., a meta-analysis by Rowland, 2015), so tests can foster learning as well as assess learning. Multiple-choice testing, however, has often been criticized as not engaging retrieval processes to the same extent as do other test types, such as cued-recall and free-recall (see, e.g., a meta-analysis by Hamaker, 1986). Recent research, however, suggests that as a pedagogical tool, multiple-choice tests can in fact trigger productive retrieval processes, provided the incorrect alternatives are competitive enough to induce the retrieval of why they are incorrect, and can even have benefits over other formats. More specifically, multiple-choice tests can boost the recall of non-tested, related information as learners retrieve and reject information connected to those alternatives (e.g., Little, Bjork, Bjork, & Angello, 2012), especially when the test format requires that learners make confidence judgments about the alternatives (Sparck, Bjork, & Bjork, 2016). Additionally, multiple-choice testing might be an efficient way to study when one is tasked with learning a large amount of information. The experiments reported in this dissertation were designed to explore further the potential of multiple-choice testing as a tool for learning.

The results of Experiments 1 and 2 suggest that taking a confidence-weighted multiple-choice test, which requires that a test taker consider more carefully than does a standard multiple-choice test why a given alternative is correct or incorrect, can lead a learner to transfer that behavior to a subsequent standard-format multiple-choice test. The results of Experiment 3, on the other hand, in which the focus was shifted to the possible benefits of multiple-choice tests as pre-tests presented before a study phase, found no significant benefits of the confidence-weighted format over the standard format.

Experiments 4 and 5 were designed to examine whether multiple-choice testing might be beneficial in the learning of vocabulary words, especially when the to-be-learned words are difficult and confusable. Again, the results demonstrated that multiple-choice testing can have benefits that go beyond the benefits of cued-recall testing, especially for words that are incorrect alternatives on an initial test, but are correct alternatives on a subsequent test. The results have major implications for the design of flashcards.

In summary, the results reported in this dissertation demonstrate that a multiple-choice test, when well designed, can be a powerful pedagogical tool, one that can contribute to optimizing educational practices.

The dissertation of Erin Michelle Sparck is approved.

Alan Dan Castel

Theodore Francisco Robles

Elizabeth Ligon Bjork, Committee Co-Chair

Robert A. Bjork, Committee Co-Chair

University of California, Los Angeles

2018

In loving memory of my grandmothers,

Anna Mae Ryan and Wilma Billings

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Experiments 1a and 4 were presented at Annual Meetings of the Psychonomic Society.

First and foremost, I would like to express my gratitude to my advisors, Dr. Elizabeth Bjork and Dr. Robert Bjork for their mentorship throughout this process. I have learned more than I could have imagined from their guidance. A special thank you to Elizabeth for all of the time she spent giving me feedback on papers, recommendation letters, and everything other request sent her way. I would also like to thank the other members of my committee, Dr. Alan Castel, and Dr. Ted Robles for their service as well as their time and feedback. A huge shout out to all of the CogFoggers, with whom I have overlapped over the years—Carole, Veronica, Doe, Courtney, Mikey, Nick, Saskia, Oliver, Derek, Stefany, Jordan, Catherine, Adam, Mary, Alex, Tara, Tyson, and everyone else. CogFog with all of you has been one of the highlights of my academic career.

Thank you also to members of the Rice community, specifically, Dr. James Pomerantz, for convincing me to do this whole Ph.D. thing and connecting me to the Bjork Lab, and Dr. Anna Cragin, for mentorship and teaching me the fundamentals of research.

I would like to thank Dr. Jeri Little for the creation and sharing of the materials used in Experiments 1a, b, 2, and 3 (as well as for much inspiration in focusing on the use of multiple-choice tests as learning tools), and Gülnaz Kiper for helping create some of the materials used in Experiments 4 and 5 (as well as her support in designing experiments and collecting data throughout her time as a research assistant). I would also like to acknowledge the work of all of my undergraduate research assistants who spent many hours running participants, scoring data, and completing numerous other helpful tasks— Monica, Yvette, Brenda, Simran, Laila, Brittany, Juliana, Rachel, McKenna, Joanna, Girija, Molly, and Bekah.

Next, I would first like to show my immense gratitude to my mom and dad, Virginia and Wayne Sparck, for their love and support and making sure I got the best education possible. I could not have asked for better parents and would not be where I am without everything that they have done for me. Thank you also to my Aunt Kathleen for being more like the sister I never had, my grandmothers, who have now both passed, for their endless love and pride (despite having no clue what it means to be a cognitive psychologist or what graduate school is all about), my in-laws, Carmen and Steven, for their understanding and support of my academic pursuits, the rest of my extended family for always having my back, and finally, to my friends, near and far, for all of your kindnesses.

I would now like to thank my best friend and husband, Dustin, for all of his programming help (none of the research in this dissertation would have been possible without his efforts), but more importantly for everything else he has done for me. He saw the ups and downs of graduate school and was, without a doubt, my biggest cheerleader. I am so grateful that he always believed in me, even when I did not believe in myself. I can honestly say I would not have made it through this journey without him by my side. Finally (and perhaps most importantly), I would like to thank the two most adorable dogs on the planet, Bonnie and Bella, for all of their kisses and snuggles. There is no better stress relief when working on a dissertation than spending time with the two of them!

VITA

2011	B.A., Psychology and Cognitive Science Rice University Houston, Texas
2014	M.A., Psychology University of California, Los Angeles Los Angeles, California
2014-2018	Teaching Assistant, Associate, and Fellow University of California, Los Angeles Los Angeles, California
2017	C. Phil, Psychology University of California, Los Angeles Los Angeles, California

PUBLICATIONS AND PRESENTATIONS

Sparck, E. M., Bjork, E. L., Aryan, L., & Tufenkjian, B. (in prep). The effects of highlighting and mind wandering on learning.

Saravanapandian, V., **Sparck, E. M.**, Cheng, K. Yaeger, C., Hu, T., Ghiani, C. A., Evans, C. J., Carpenter, E. M., & Ge, W. (in prep). Quantitative assessments reveal improved neuroscience engagement and learning through outreach.

Ramachandran, R., **Sparck, E. M.**, & Levis-Fitzgerald, M. (in prep). Using JoVE science education videos in a large introductory chemistry course.

Liu, J., **Sparck, E. M.**, & Bjork, E. L. (2018). Retrieval-based structure-building: The effect of concept organization and concept mapping on text learning. Poster to be presented at the 59th annual meeting of the Psychonomic Society, New Orleans, LA, USA.

Sparck, E. M. & Levis-Fitzgerald, M. (2018). Mobilizing around course evaluation: Improved feedback and integration of custom learning questions. Poster to be presented at the 2018 AAC&U Transforming STEM Higher Education Conference, Atlanta, GA, USA.

Ramachandran, R., & **Sparck, E**. **M**. (2018). Using JoVE science education videos in a large introductory chemistry course. Talk presented at the 25th Biennial Conference on Chemical Education, Notre Dame, IN, USA.

Sparck, E. M., Bjork, E. L., Bjork, R. A. & Kiper, G. (2017). Using multiple-choice tests to improve vocabulary learning via flashcards. Poster presented at the 58th annual meeting of the Psychonomic Society, Vancouver, BC, Canada.

Sparck, E. M., Bjork, E. L., Aryan, L., & Tufenkjian, B. (2017). The effects of highlighting on mind-wandering and memory. Poster presented at the 58th annual meeting of the Psychonomic Society, Vancouver, BC, Canada.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications, 1. doi: 10.1186/s41235-016-0003-x.*

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). Experience with confidence-weighted multiple-choice tests improves later recall of related information. Poster presented at the 57th annual meeting of the Psychonomic Society, Boston, MA, USA.

Soderstrom, N. C., **Sparck, E. M.**, & Bjork, E. L. (2016). Variable practice enhances learning of foreign language vocabulary. Poster presented at the 57th annual meeting of the Psychonomic Society, Boston, MA, USA.

Cheng, K. Y., **Sparck, E. M.**, Yaeger, C., Ghiani, C. A., Ge, W., & Carpenter, E. M. (2016). Neuroscience outreach to the greater Los Angeles K-12 community. Poster presented at the 46th annual meeting of Society for Neuroscience, San Diego, CA, USA.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). Uncovering how multiple-choice testing triggers productive retrieval processes. Poster presented at the 124th annual meeting of the American Psychological Association, Denver, CO, USA.

Bjork, E. L., Soderstrom, N., Little, J. & **Sparck, E.** (2015). Multiple-choice testing as a desirable difficulty: Evidence from the laboratory and the classroom. Talk presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL, USA.

Sparck, E. M., Bjork, E. L., Bjork, R. A. (2015). When and why multiple-choice testing triggers productive retrieval processes. Poster presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL, USA.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2014). Confidence-weighted multiple-choice tests enhance retention of non-tested related information. Poster presented at the 55th annual meeting of the Psychonomic Society, Long Beach, CA, USA.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2014). Can confidence-weighted multiple-choice testing enhance retention of non-tested, but related, information? Poster presented at the 26th annual meeting of the Association for Psychological Science, San Francisco, CA, USA.

Cragin, A., **Sparck, E**. & Pomerantz, J. R. (2011, presenting author Pomerantz). Indicating direction efficiently: A few pointers. Paper presented at the 52nd annual meeting of the Psychonomic Society, Seattle, WA, USA.

Pomerantz, J. R., Stupina, A. I., & **Sparck, E**. (2011). What's the "point"? Assessing the effectiveness of stimuli that indicate direction. Poster presented at the 11th Annual Meeting of the Vision Sciences Society, Naples, FL, USA.
In: Journal of Vision, 11(11), article 1107. doi:10.1167/11.11.110

Chapter 1: Introduction

Testing as a *Desirable Difficulty*

Retrieving information is a potent learning experience, as once a piece of information is retrieved, it becomes more accessible in the future (Bjork, 1975; Carrier & Pashler, 1992). Actively retrieving information, as one might do while taking a practice test (even without feedback), produces superior long-term retention of that information than restudying does, a phenomenon in cognitive psychology known as the testing effect or test-enhanced learning (for reviews see Dempster, 1996; Roediger & Karpicke, 2006; Rowland, 2015). Although from a student's perspective engaging in self-testing requires more effort, feels less fluent, and as a result, can seem less productive than restudying; from a memory standpoint, that additional effort is beneficial for later retention. Testing is one of a number of *desirable difficulties*, or study strategies that may seem initially to slow or even impair learning, but over time, produce more robust, lasting learning (Bjork, 1994).

Much of the past research on testing has focused on what happens to the information that was directly tested. Understanding, however, what happens to the other to-be-learned information—that is, information other than that which was directly tested during retrieval practice— is also of importance, especially with respect for its practical implications for education. When students engage in self-testing or take a practice test as preparation for a future exam, they likely only retrieve a subset of the to-be-learned (and therefore potentially to-be-tested) information. From here, several interesting questions emerge. Specifically, what happens to the information that was not directly tested? Could that non-tested, but related information, inadvertently be harmed by the retrieval practice given to the tested information?

Would nothing of significance happen to that information? Or most interestingly for educational purposes, could memory for that information be benefited in some way?

Inhibition and Facilitation as Consequences of Testing

Research on the memory phenomenon known as retrieval-induced forgetting (RIF) suggests that repeatedly retrieving an answer on a practice test could potentially lead to an impaired ability to retrieve related information on the final test, relative to restudying. In the basic paradigm in which RIF was first observed, individuals studied an assortment of category-exemplar pairs (e.g., *Fruit: Banana*; *Fruit: Orange*; *Drink: Whiskey*; *Drink: Vodka*). After all of the category-exemplar pairs have been studied, individuals practiced retrieving half of the items from half of the studied categories, given the intact category and the first two letters of the exemplar (e.g., *Fruit: Ba*_____). On the final recall test, all of the category-exemplar pairs are tested. Exemplars from the practiced categories that were not retrieved during the retrieval practice phase (e.g., *Fruit: Orange*) are recalled, on average, at a lower rate relative to items from categories whose exemplars were never retrieved (e.g., *Drink: Whiskey*; *Drink: Vodka*). This finding suggests that such related information might have been inhibited during retrieval practice as a way of resolving the competition that would have arisen from these items during the prior retrieval practice (Anderson, Bjork, & Bjork, 1994).

Expanding on this basic finding with highly controlled materials in a laboratory setting, research exploring whether retrieval-induced forgetting also occurs with more educationally realistic materials has produced mixed results. Macrae and MacLeod (1999) had two groups of participants learn 20 geography facts about two fictional locations. Then one of these groups engaged in retrieval practice for a subset of the facts about one of those locations, while none of the facts about the other location underwent retrieval practice. The other group did not engage in retrieval practice for any of the information about either location and thus served as a control group. When both groups were given a later test on all of the information about both locations, the individuals who had repeatedly retrieved 10 of the facts about one location remembered those specific practiced pieces of information well. They, however, remembered significantly fewer of the 10 non-practiced facts compared with individuals in the group who had not engaged in any retrieval practice, exhibiting RIF. Carroll, Campbell-Ratcliffe, Murnane, and Perfect (2007), exploring learning of text passages in the basic RIF paradigm, found that retrieval practice of text-based information led to inhibition of related information from both ordered and disordered text on later free-recall and short answer tests. Related information on later multiple-choice tests, however, showed no such impairment, suggesting that the type of final test may play an important role in determining how related information fares.

Initial test format is also an important consideration when investigating the effect on non-tested information. Hinze and Wiley (2011) found free-recall initial tests where individuals tried to recall sentences following the reading of science materials led to an increase in performance on novel questions, but fill-in-the-blank initial questions did not. LaPorte and Voss (1975) found a similar null result using fill-in-the-blank initial tests. Whereas free-recall initial tests might engage learners in trying to retrieve a wide variety of information, fill-in-the-blank initial tests might restrict the range of information that learners attempt to retrieve.

Similarly, cued-recall initial testing may generally promote a narrow retrieval strategy, focusing the learner only on the answer to the present question, and thus a benefit to related information would not be expected (unless some sort of mediator is present and actively retrieved). Chan, McDermott, and Roediger (2006, Experiment 1), for example, created two sets of short-answer questions (e.g., question and answer from set A: *Where do toucans sleep at* *night? Tree holes*; question and answer from set B: *What other bird species is a toucan related to? Woodpeckers*). One set of questions was used as an initial test after participants read expository text about toucans, while the other served as the final test (questions were counterbalanced across participants). Each pair of questions was broadly related based on conceptual information that appeared in the passage in close temporal proximity. In one sentence, for example, the text conveys that toucans sleep in tree holes, but because they have soft bills, they cannot create the holes. Toucans, therefore, rely on woodpeckers to create their sleeping habitats. In the next sentence, the text further explains that toucans and woodpeckers are in fact from the same family of birds. By retrieving *tree holes* during the initial test, information about woodpeckers might also be recalled, thus generally strengthening access to information about woodpeckers for future questions.

Chan et al. (2006) assert that individuals who answered these initial questions likely spontaneously engaged in a broad retrieval strategy, actively searching for any information related to the question because recalling that information might also aid in the recall of the current correct response to that question. Thus, when such information was directly relevant to questions on the final test, facilitation for it was seen. In line with this hypothesis, follow-up research showed that when individuals were explicitly given instructions to engage in such a strategy, facilitation occurred, suggesting that this broad strategy might be the type of retrieval strategy in which participants were engaging when facilitation was found in the previous study. Additionally, when individuals were given explicit instructions to use a narrow strategy, specifically to think about the correct answer and only the correct answer, facilitation of related information disappeared. Instead, the recall of the non-tested, related material was comparable to that of the control questions (Chan et al., 2006; Experiment 4).

Benefits of Multiple-choice Testing on Non-Tested Related Information

The idea that multiple-choice tests might be able facilitate the retention of information that was not directly tested has roots dating back to the adjunct question literature from the 1960s. Both multiple-choice and cued-recall tests were shown to facilitate the recall of "incidental" information from the studied passage (Frase, 1968). However, these results are difficult to interpret, as the construction of such questions was not being considered as an important variable in determining how not directly tested information might fare after an initial test.

More recently, Little, Bjork, Bjork, and Angello (2012, Experiment 1) directly compared how initial test format (specifically cued-recall and multiple-choice initial tests) impacted the recall of non-tested, but related information, operationalizing how initial multiple-choice questions and final questions should relate to one another. Similar to Chan et al. (2006), Little et al. constructed two sets of related question pairs (e.g., question and answer from set A: *What is the tallest geyser in Yellowstone National Park? Steamboat Geyser*; question and answer from set B: *What is the oldest geyser in Yellowstone National Park? Castle Geyser*). Rather than just being close in temporal proximity to one another during learning and relying on non-target mediating information as the questions in Chan et al. did, the question pairs constructed by Little et al. related to each other on the basis of being about the same specific topic (e.g., geysers in Yellowstone National Park).

Additionally, all of the alternatives had been studied in the text and were thus potentially confusable with one another. These competitive alternatives were selected to engage the test-takers in processes that encouraged differentiating among them during the initial test, such as recalling all of the information they could about each option during their selection of an answer.

The related questions from each set thus contained the same competitive alternatives (e.g., *Steamboat Geyser*, *Castle Geyser*, and *Old Faithful*) when presented in the multiple-choice format. See Table 1 for a sample question.

Table 1

Example question pair with corresponding correct and incorrect alternatives from Little, Bjork, Bjork, and Angello (2012)

	Alternatives	
Example question pair	**Correct**	**Incorrect**
(A) *What is the tallest geyser in Yellowstone National Park?*	*Steamboat Geyser*	*Castle Geyser* *Old Faithful*
(B) *What is thought to be the oldest geyser in Yellowstone National Park?*	*Castle Geyser*	*Steamboat Geyser* *Old Faithful*

Participants who took an initial multiple-choice test following the reading of a text passage produced a significant improvement in the recall of non-tested, related items on a final cued-recall test given at a 5-minute delay relative to their performance on a control passage that had received no initial testing. In contrast, participants who took an initial cued-recall test suffered a decrease in the recall of non-tested, related items on the 5-min delayed cued-recall test, relative to their performance on this test for the control passage. The same pattern of results emerged both when participants were given feedback after the initial test and when they were not. Little et al. (2012) suggested that the inclusion of plausible incorrect alternatives on the initial multiple-choice test might engage the test-takers in spontaneous retrieval of information about why those competitive alternatives are incorrect (in support of trying to answer the question correctly), while cued-recall initial testing might only focus the test-takers on the correct answer to the present question.

Owing to the activation of information associated to each alternative while trying to discriminate and select among them during the initial multiple-choice test, deeper processing presumably occurs and results in enhanced recall of that information on later tests (e.g., Whitten & Leonard, 1980). The results of Little et al. (2012) are all the more impressive given that typically after retrieval practice in the presence of strong competitors (e.g., Anderson et al., 1994), as could happen with these materials during the initial test, we might expect to see inhibition of that related information. Rather, recalling relevant information during the initial multiple-choice test to select against these competing alternatives appears not only to protect such information from inhibition but to lead to its facilitated recall at a later time.

Similar results showing the benefits of multiple-choice testing have been found outside the laboratory in a classroom setting (Bjork, Little, & Storm, 2014). After multiple-choice retrieval practice with competitive alternatives, students in a large psychology research methods course were better at answering questions on concepts related to those questions (e.g., had better scores on questions related to the directionality problem in correlational research after being initially tested on the third variable problem) on the final exam (M = 90%) as compared to baseline control questions regarding topics that were not initially tested (M = 72%). These results suggest widespread benefits for the use of multiple-choice testing to facilitate the learning of related information.

Multiple-choice tests have been argued to bypass retrieval, the mechanism underlying the testing effect (e.g., Kintsch, 1970), and have in many instances been found to be less effective at promoting the retention of information (Duchastel, 1981; Foos & Fischer, 1988; Hamaker, 1986; McDaniel, Anderson, Derbish, & Morrisette, 2007; although see Kang, McDermott, & Roediger, 2007). The research by Little et al. (2012) and Bjork et al. (2014), however, suggest that multiple-choice initial or practice tests can engage individuals in productive retrieval processes. They may also have the added benefit of strengthening access to related information over their cued-recall counterparts. Multiple-choice initial tests can guide the test-taker to strengthen information, either directly or indirectly, about related concepts (that is, concepts other than the ones explicitly tested) which may then be necessary to retrieve on a later test. That is, the alternatives can potentially activate retrieval of information about many things beyond just the target answer if all answer choices are plausible and competitive with one another. Cued-recall test-takers, on the contrary, do not receive these additional cues as the alternatives are not presented, and thus they may only show facilitation for recall of related information when such information serves as a mediator in the search for the correct target answer (e.g., Chan et al., 2006). Even in that case, however, for individuals predisposed to using only a narrow retrieval strategy, facilitation of the non-tested, related information might not occur.

It is important to note that the activation of such productive retrieval processes regarding incorrect alternatives requires that alternatives are competitive lures (Little & Bjork, 2015). Given, for example, the question, *Which outer planet was discovered by mathematics rather than direct observation?, Uranus* would be a competitive incorrect alternative because Uranus is an outer planet. *Mercury*, on the other hand, would not be considered competitive because it is an inner planet and can be too easily rejected. Consequently, on a later test, while enhanced recall of information about Uranus is likely to be found, enhanced recall of information about Mercury would most likely not be observed.

While the presence of competitive alternatives appears to be a necessary condition for encouraging retrieval about why a given incorrect alternative is incorrect, it does not seem to be a sufficient condition. Little (2011, Experiment 5) found that nearly 70% of test takers did not engage in such a strategy during initial testing without explicit instructions to act in this manner. Even for those that did retrieve information related to the incorrect alternatives during the practice test, it is unknown whether those participants always used that strategy in selecting their answer, or only when they could not immediately come up with the answer and were using such a strategy to try to infer the correct answer. Little also found that the benefit to related information only emerged (compared to a conservative extended restudy control) when participants were given specific instructions to think about why the alternatives they did not choose were incorrect.

Thus, it is possible that the benefit to later recall seen in Little et al. (2012) might be attenuated relative to what it could be if more participants engaged in this optimal strategy when taking a competitive multiple-choice test. Perhaps test-takers generally need additional instruction on effective test-taking strategies, even when the test is multiple-choice, particularly when the purpose of the test is pedagogical—not just for assessment. Test-takers, may otherwise, think too narrowly, or only use a more appropriate broad strategy under certain conditions, such as when they are confused about the question or do not feel that they can immediately select the correct answer. Use of a broad retrieval strategy in which test-takers attempt to retrieve anything that they possibly can remember related to the topic, however, could be beneficial even when the correct answer to the initial question is immediately known. Possibly, alternate formats or presentations of multiple-choice questions might encourage more effective test-taking strategies that support broad retrieval of studied information without direct instruction.

Another potential benefit that has not been directly addressed is whether multiple-choice testing, with careful construction, might offer an efficient way to study. For example, when asked *What is the tallest geyser in Yellowstone National Park?* in the multiple-choice format with three alternatives present, participants potentially are guided to retrieve three pieces of information (the correct answer as well as information about the two incorrect alternatives). The cued-recall version, on the other hand, would encourage retrieval of only the correct answer, consistent with the broad versus narrow strategies previously outlined. If a learner only has a fixed number of practice questions, multiple-choice testing could help increase the sheer amount of information that the learner has access to relative to cued-recall testing. The question of whether multiple-choice testing can lead to more efficient studying will be explored in Chapter 4.

Benefits of Confidence-weighted Multiple-choice Testing (Thus Far)

Using identical materials from Little et al. (2012), Sparck, Bjork, and Bjork (2016, Experiment 1) found further evidence that a multiple-choice testing format, referred to as confidence-weighted multiple-choice testing (adapted from Bruno's Information Reference Testing, Bruno, 1989; Bruno, 1993) in which individuals select their answers based on their relative confidence, can increase the benefit for recalling non-tested, but related information as compared to a standard multiple-choice format. In the confidence-weighted multiple-choice format, three alternatives — one correct and two competitive incorrect (e.g., *Venus*, *Mercury*, and *Saturn*) — are placed at the vertices of a triangle as the plausible answers to the question, *What planet lacks an internal magnetic field?* (see Figure 1.1). The test taker can select one of these alternatives (e.g., *Venus*), indicating complete confidence in their choice. On the other hand, they can select one of the points along either of the lines connecting two vertices of the triangle (e.g., along the line connecting *Venus* and *Mercury*). This selection indicates uncertainty as to which of those alternatives is the correct answer, but certainty that the alternative at the other corner of the triangle is incorrect.

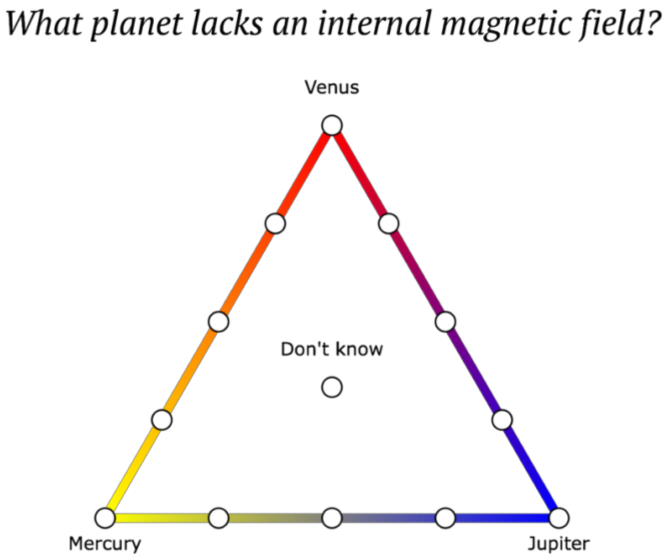

Figure 1.1. Sample confidence-weighted multiple-choice item, as would be seen by the participant in Sparck et al. (2016) with the alternatives appearing at the vertices.

Confidence-weighted multiple-choice testing also allows test takers to indicate relative confidence in the correctness of each of the alternatives being considered. If, for example, the test taker supposes either *Venus* or *Mercury* could be correct, but is more confident in answering *Venus*, he or she can select a point along the line between *Venus* and *Mercury* closer to *Venus* than to *Mercury*. Partial knowledge thus can be demonstrated. Selecting an intermediary point between *Venus* and *Mercury* indicates a confident rejection of *Saturn* as the correct answer, and if that point is closer to *Venus* than *Mercury*, the test taker believes *Venus* is more likely to be the correct answer than is *Mercury*.

The scoring system for the confidence-weighted multiple-choice format differs from that of a standard multiple-choice test in several key ways. First, guessing is strongly discouraged. By choosing an incorrect alternative, or any point on the line between the two incorrect alternatives, which amounts to fully rejecting the correct answer, the test taker will receive a major loss (10 points in Sparck et al., 2016; shown in Figure 1.2). Additionally, an option in the middle of the triangle allows the test taker to indicate that they do not know the answer, and as a result no points would be awarded or lost (although future confidence-weighed multiple-choice experiments discussed in the present research will exclude this option to more closely equate the confidence-weighted and standard multiple-choice conditions). Most notably, as shown in Figure 1.2, the points accumulated for choosing a correct alternative are only marginally greater relative to the points that are close to that alternative on either of the sides that include that alternative.

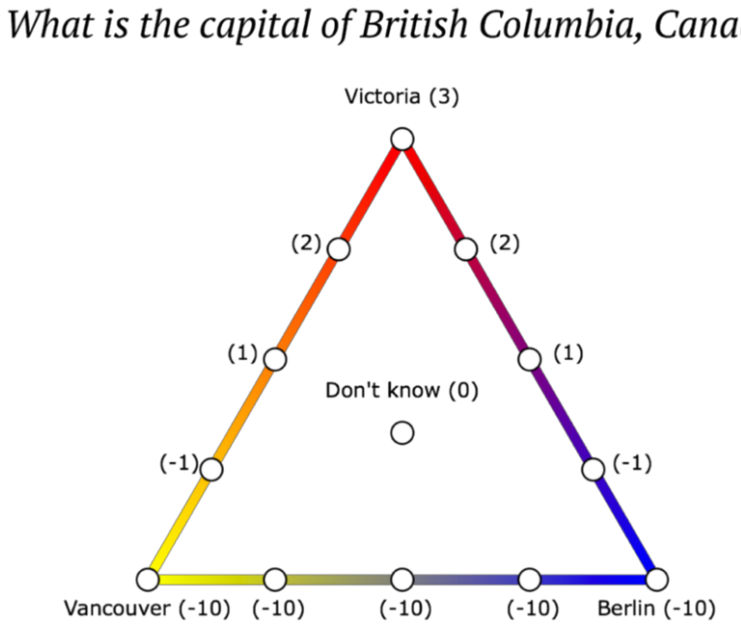

Figure 1.2. Sample confidence-weighted multiple-choice question used as part of the instructions to participants in Sparck et al. (2016). Test takers can select any of the bubbles as their answer. Points that would be gained or lost for each answer are shown in parenthesis next to the corresponding bubble, given that *Victoria* is the correct answer.

The strategies employed by test takers with this multiple-choice test format appear to involve more spontaneous retrieval of information about the incorrect alternatives than standard multiple-choice formats. On a final cued-recall test, the proportion of non-tested, related answers recalled was significantly greater for individuals who took initial confidence-weighted multiple-choice tests compared with those individuals who took standard multiple-choice tests (or took no tests at all, although a robust testing effect emerged for both formats relative to no test).

The increased benefit to initially non-tested, related information does not appear to come from making a confidence judgment alone. Comparing an initial standard multiple-choice test with a standard multiple-choice test in which individuals selected their answer and then indicated how confident they were in their selected answer by indicating a numeric value on a scale of 0-100 showed no differences in the retention of related information on a later cued-recall test (Sparck et al., 2016, Experiment 2). Performance on related questions was again significantly higher for participants in the confidence-weighted multiple-choice test condition compared with conditions that utilized the standard multiple-choice format. The type of processing afforded by the confidence-weighted multiple-choice initial test seems to have led participants to engage in retrieval of information about each of the alternatives to a greater extent than standard multiple-choice testing as they assess their relational confidence in the answers.

It should be noted that the increased benefit to recall of related information after using confidence-weighted multiple-choice testing occurred naturally without having to inform the learners to use a specific strategy, although we know such instructional manipulations work (e.g., Chan et al., 2006; Little, 2011). Great value, however, lies in a testing format that encourages learners to spontaneously engage in more productive retrieval processes without explicit instruction. Potential additional benefits of confidence-weighted multiple-choice testing are further explored in Chapters 2 and 3.

Chapter 2: Further Increasing the Effectiveness of Multiple-choice Post-testing Through Confidence-weighted Testing

Additional research on the benefits of confidence-weighted multiple-choice testing is needed to fully understand under what conditions it may be more useful than standard, traditional multiple-choice testing at improving the recall of non-tested, related information. In this second chapter, I focus on a series of experiments designed to explore this line of research: specifically, whether the confidence-weighted multiple-choice format creates a strategy shift when taking subsequent standard multiple-choice tests (and/or subsequent studying of new information).

Experiments 1a and 1b: Does experiencing a confidence-weighted multiple-choice practice test lead learners to transfer a similar test-taking strategy to standard multiple-choice practice tests?

Previous findings from Sparck et al. (2016) suggest that confidence-weighted multiple-choice tests could be particularly useful as practice tests because they might lead students to become more appreciative of or sensitive to the benefits of using a broad retrieval practice strategy during practice tests without a direct instruction to do so. That is, perhaps experience with this testing format would result in learners becoming more aware of the benefits stemming from actively trying to retrieve information not only about the alternative they considered most likely to be correct, but information related to alternatives they considered most likely to be incorrect as well. If so, experience using the confidence-weighted multiple-choice testing format could improve the degree to which learners actively engage in productive retrieval regarding all of the potential alternatives while taking future standard multiple-choice test, allowing them to maximize the learning benefits of confidence-weighted multiple-choice testing in other testing situations.

A potential criticism of confidence-weighted multiple-choice testing is that it might be more difficult to implement compared to the standard multiple-choice testing format in contexts outside of the laboratory, such as by instructors in a classroom. If, however, the strategy used in taking a confidence-weighted multiple-choice test would generalize to the taking of a standard format multiple-choice test, then perhaps the benefits observed by Sparck et al. (2016) could also be seen in the classroom with use of the more straightforward to construct standard multiple-choice tests following only a brief exposure of students to the confidence-weighted testing format.

The present experiment was designed to assess whether individuals who first experience the strategy invoked during the answering of questions on a confidence-weighted multiple-choice test then transfer use of that strategy to a subsequent multiple-choice test on completely new material. If so, then it is predicted that participants who take confidence-weighted multiple-choice initial tests after study of a first passage will correctly answer a greater proportion of related questions on the final test of the second passage.

After data collection for Experiment 1a had been completed, it was discovered that a group of participants had experienced a technical glitch, such that these individuals read each passage twice before being tested. The data from participants known to have experienced the glitch were excluded. It is possible, however, that other participants could have also experienced the glitch, so Experiment 1b is a direct replication without the technical error.

Method

Participants

In Experiment 1a, 96 undergraduates from the University of California, Los Angeles psychology subject pool were recruited to participate for partial course credit. Eight participants were excluded due to the technical glitch previously mentioned, as well as two others due to the computer failing to record the responses or for the participant failing to answer all questions on the final test. After those exclusions, 86 participants remained (22 male, 64 female; $M_{\text{age}} = 21.2$ years).

Using the effect size ($d = .43$) from Experiment 1a, it was estimated that approximately 130 participants would be necessary for sufficient power for the replication. As such, 130 undergraduates from the University of California, Los Angeles psychology subject pool were recruited to participate for partial course credit in Experiment 1b. Seven participants were excluded for either their final test answers failing to be recorded by the computer or leaving all answers to both the initial and final tests completely blank, leaving 123 participants (30 male, 93 female; $M_{\text{age}} = 20.8$ years) in the sample.

Design

There were two between-subjects conditions. The format of the initial test (either standard multiple-choice or confidence-weighted multiple-choice) following the reading of the first passage varied between conditions.

Materials

Two passages (each approximately 1200 words), one on Saturn and one on Yellowstone National Park, as well as two 10-question sets of related question pairs for each passage (set A and set B) with competitive incorrect alternatives (when presented in either of the multiple-choice formats) as determined by Little et al. (2012) were used. The materials used are presented in Appendix A.

Procedure

All participants were randomly assigned to either the standard multiple-choice or the confidence-weighted multiple-choice test condition. If assigned to the confidence-weighted multiple-choice condition, participants were briefed on how to appropriately answer questions in the unfamiliar format and were not allowed to move on to the rest of the experiment until they had demonstrated understanding of the scoring system being used. The format for the confidence-weighted multiple-choice tests used in the present experiment differed from that of the one previously described in Sparck et al. (2016) by not presenting the option to select *Don't Know*, but rather, participants were forced to select an along the sides or the vertices of the triangle. Otherwise, presentation of the confidence-weighted multiple-choice format was presented and scored identically to Sparck et al. (2016). The updated scoring system without the *Don't Know* option is shown in Figure 2.1. Participants in the standard multiple-choice condition were awarded one point for correct answers and no points for incorrect answers.

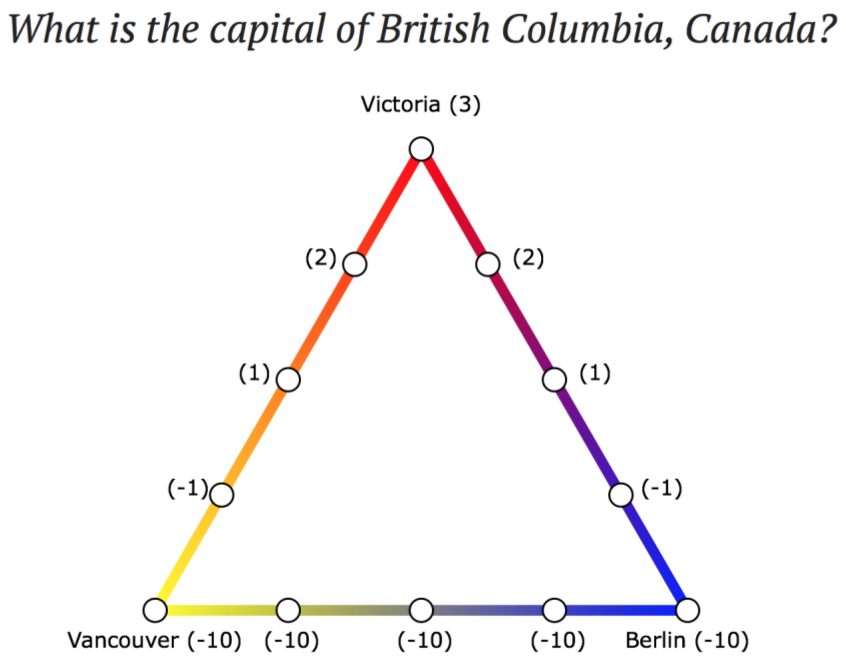

Figure 2.1. Updated confidence-weighted multiple-choice scoring system for Experiments 1a, 1b, 2, and 5. Test takers can select any of the bubbles as their answer. Points that would be gained or lost for each answer are shown in parenthesis next to the corresponding bubble, given that *Victoria* is the correct answer.

All materials were presented by means of a computer, and as can be seen in the procedure depicted in Figure 2.2, all participants began by reading one of the two passages (order counterbalanced across participants) for 9 minutes. After reading the first passage, everyone took an initial test, either using the standard multiple-choice format or the confidence-weighted multiple-choice format as outlined. The initial test consisted of 10 questions from either question set A or question set B (counterbalanced across participants), and each question remained on the screen for 25 seconds (with a 10 second warning) before the participant automatically was moved on to the next question. After the initial test was complete, all participants read the other passage for 9 minutes. Everyone took a standard multiple-choice test on the second passage. After answering all of the questions on each initial test, participants were told a summary of their score for those 10 questions. Participants, however, were not given any specific item-by-item feedback as to which questions were answered correctly or incorrectly.

After a 10-minute Tetris distractor task, everyone took a self-paced, cued-recall final test on the related question set for the second passage only. That is, if participants answered questions from set A on the initial test for the second passage, on the final test they answered questions from set B and no alternatives were provided. Given the construction of the materials as previously described, the correct answer to each question on the final cued-recall test had always appeared as an incorrect alternative to a question that was answered on the initial standard multiple-choice test for the second passage. Participants were encouraged to attempt to answer all questions even if they were unsure of their answers. Following the completion of the experiment, participants answered survey questions regarding their strategy use during the experiment.

.

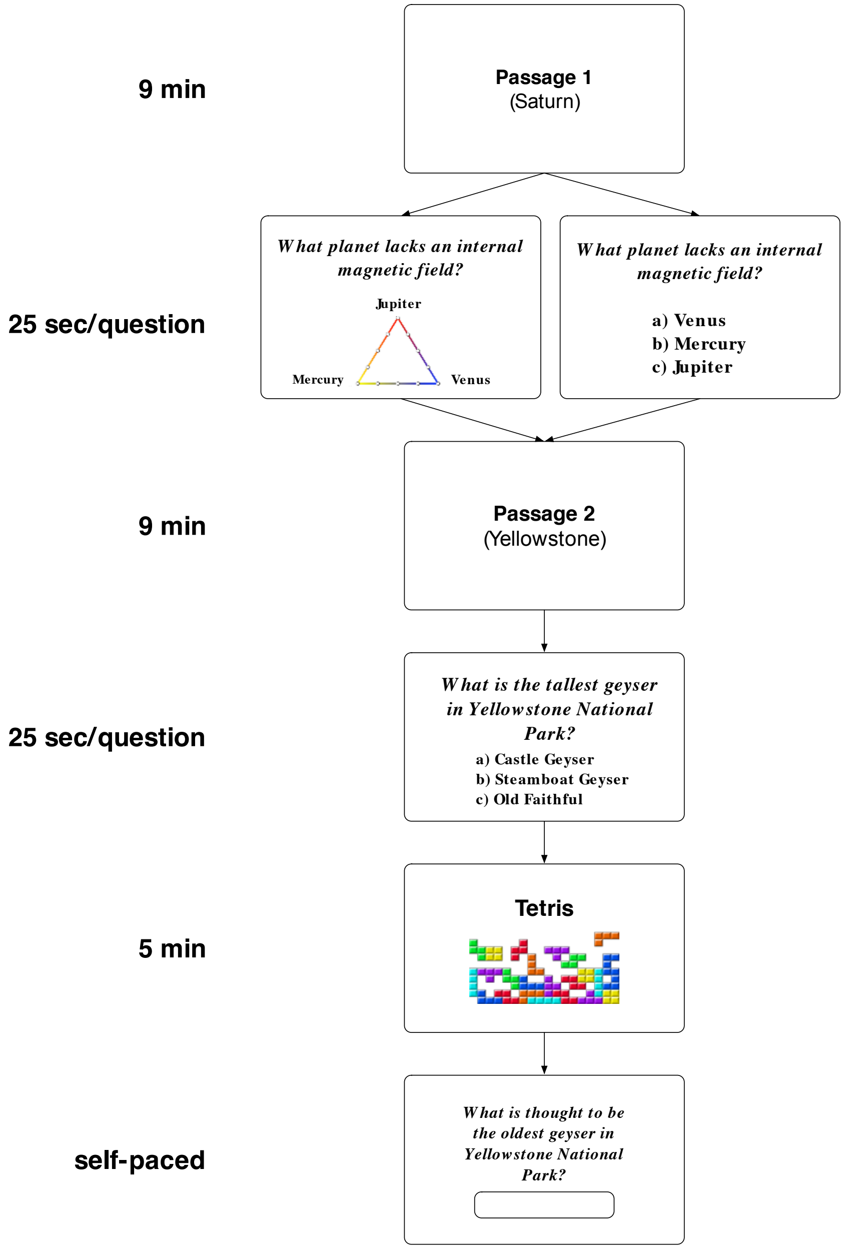
9 min
Passage 1
(Saturn)
25 sec/question
What planet lacks an internal magnetic field?
Jupiter
Mercury
Venus
What planet lacks an internal magnetic field?
a) Venus
b) Mercury
c) Jupiter
9 min
Passage 2
(Yellowstone)
25 sec/question
What is the tallest geyser in Yellowstone National Park?
a) Castle Geyser
b) Steamboat Geyser
c) Old Faithful
5 min
Tetris
self-paced
What is thought to be the oldest geyser in Yellowstone National Park?

Figure 2.2. Diagram of the procedure for Experiments 1a and b. Participants began by reading a passage followed by the experimental manipulation of taking either an initial standard multiple-choice or an initial confidence-weighted multiple-choice test on information from the first passage. All participants then read a different passage followed by a standard multiple-choice test for it. Finally, participants engaged in a short Tetris distractor and then took a final cued-recall test on non-tested, related information from the second passage only.

Results

Responses on the final cued-recall test were scored by an independent rater, who was blind to the participants' conditions, and according to a lenient scoring guide to allow for slight misspellings and typos. No penalties for incorrect answers were assessed on the final test. The proportion of correct responses on the final cued-recall test was calculated on the basis of 10 items (from the second passage). Independent sample *t*-tests were used in the analyses to compare the means of the two groups on both the initial and final tests. Results for Experiments 1a and b are shown in Figure 2.3.

For Experiment 1a, participants in the confidence-weighted multiple-choice condition (*M* = .47, *SD* = .19) significantly outperformed participants in the standard multiple-choice condition on the final test of related questions (*M* = .38, *SD* = .2), *t*(84) = 2, *p* = .049, *d* = .43. Performance on the second, and always, standard initial multiple-choice test was numerically different depending on the type of test taken after the first passage. Participants who took a confidence-weighted multiple-choice test after the first passage answered .7 (*SD* = .18) of the questions on the second (now standard multiple-choice) test correctly, while participants who took a standard standard multiple-choice after the first passage answered .62 (*SD* = .22) of the questions correctly on their second multiple-choice test, *t*(84) = 1.69, *p* = .095.

The participants in Experiment 1b showed a similar pattern to those in Experiment 1a. Participants in the confidence-weighted multiple-choice condition (*M* = .43, *SD* = .17) significantly outperformed participants in the standard multiple-choice condition on the final test of related questions (*M* = .36 *SD* = .18), *t*(121) = 2.15 , *p* = .034, *d* = .39. Participants who took a confidence-weighted multiple-choice after the first passage answered .74 (*SD* = .19) of questions on the second (now standard multiple-choice) test correctly, while participants who took a standard multiple-choice after the first passage answered .68 (*SD* = .18) of the questions correctly on their second multiple-choice test, *t*(121) = 1.8, *p* = .075. See Table 2.1 for a comparison of on scores on the initial standard multiple-choice test for the second passage.

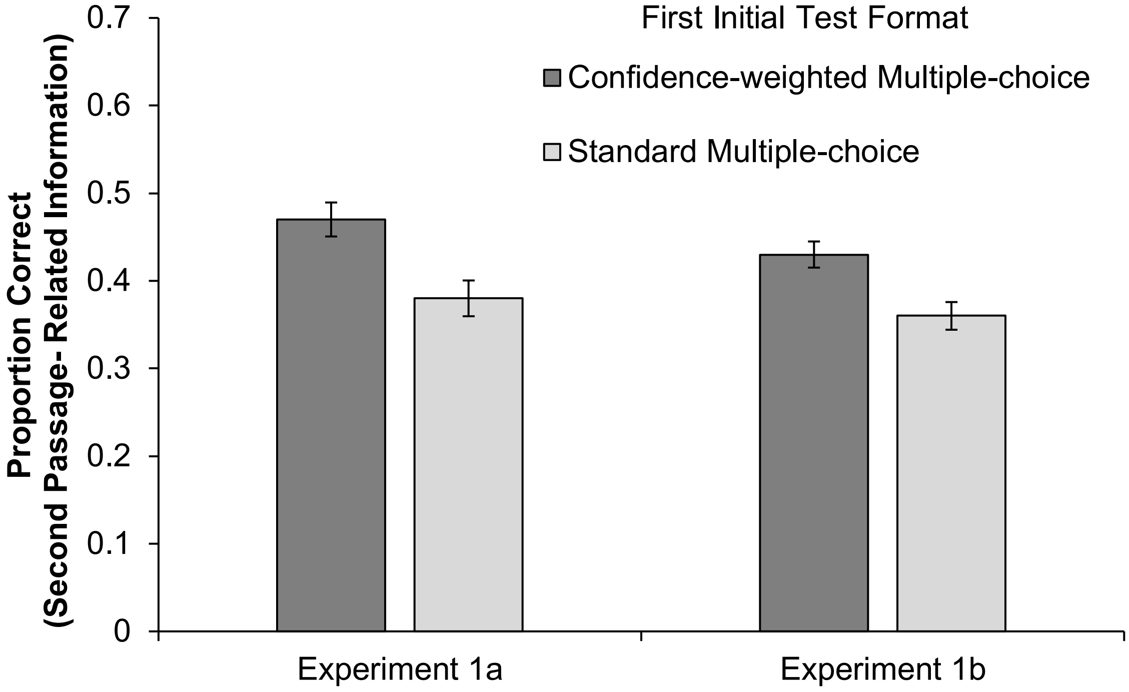

Figure 2.3. The results of Experiments 1a (left bars) and 1b (right bars). Proportion of correct answers to related questions from the second passage as a function of the initial test format taken after reading the first passage. The darker bars represent the proportion correct for participants in the confidence-weighted multiple-choice condition while the lighter bars represent the proportion correct for participant in the standard multiple-choice condition. Error bars represent ± 1 standard error of the mean.

Table 2.1

Proportion of items correct on the first initial test taken for Experiments 1a, 1b, and 2

	Experiment		
First Initial Test Format	**1a***	**1b***	**2**
Confidence-weighted Multiple-choice	0.70	0.74	0.61
Standard Multiple-choice	0.62	0.68	0.61

*indicates trend towards significant difference (*p* < .1)

Discussion

The pattern of results from Experiments 1a and 1b suggest that something about first taking a confidence-weighted multiple-choice test might have impacted strategy-use during a subsequent standard multiple-choice test. There is, however, also the possibility given the design of the experiment that participants may have encoded information during the reading of the second passage differently after having experienced an initial confidence-weighed multiple-choice test.

Prior research suggests that testing can potentiate subsequent learning (e.g., Arnold & McDermott, 2013). Specifically, retrieval practice can make the subsequent study of new, unrelated information on a completely different topic more effective (Yue, Soderstrom, & Bjork, 2015; Experiment 2). Test format could thus also play a role in how effective that subsequent study may be, with confidence-weighted multiple-choice tests providing a greater benefit than standard multiple-choice tests to future studying.

Assessing performance on the second, and always, standard multiple-choice test in Experiments 1a and 1b tells us this might be a possibility. Performance on this initial test for

Experiments 1a and 1b was numerically different depending on the type of test taken after the first passage. Across both experiments participants who took a confidence-weighted multiple-choice test after the first passage answered 70% and 74%, respectively, of questions on the second (now standard multiple-choice) test correctly while participants who answered standard multiple-choice after the first passage answered 62% and 68%, respectively, of the questions correctly on their second multiple-choice test (see Table 2.1).

Although only approaching significance in both Experiments 1a and 1b, the results from performance on the second initial test suggest a trend that perhaps the study of the second passage (on a completely new topic) was more effective for those individuals who had previously taken a confidence-weighted multiple-choice test. It is unclear in Experiments 1a and 1b whether any benefits occurred due to strategy change while answering questions on the second multiple-choice practice test, or more effective encoding during the reading of the second passage itself, or alternatively, some combination of the two options.

Experiment 2: Does experiencing a confidence-weighted multiple-choice practice test also potentiate the effectiveness of subsequent study?

If all of the interesting cognitive processes leading to the related question benefit are happening during the initial testing, the pattern of results for Experiment 2 should look similar to Experiments 1a and 1b. If all of the interesting processes leading to the benefit are occurring during the encoding of the second passage, the effect will no longer be present. This result would be inconsistent with the original hypotheses. It is also possible, and predicted based on Little et al. (2012), Sparck et al. (2016), and Yue et al. (2015), that both a strategy change and potentiation during reading occurred, leading to better performance on the final related questions, which would mean an effect, albeit a smaller one.   The majority of the effect, however, is hypothesized to come from an increase in productive retrieval on the second multiple-choice test, making the expected effect size for Experiment 2 only slightly smaller.

Experiment 2 allows us to separate out the potential benefits of each by blocking together the reading of both passages followed by the blocking together of taking both initial tests, so that any benefits to related information on the final test can be attributed to what occurred during the taking of the second initial test. Whereas the basic procedure of Experiment 1 might be briefly described as *study*, *test*, *study*, *test*, the procedure of Experiment 2 would be described as *study*, *study*, *test*, *test*.

Method

Participants

Using an average of the effect sizes from Experiments 1a and 1b (*d* = .44 and .39), it was estimated in G*Power using an alpha of 0.05 and a power of 0.80, that approximately 160 participants would be needed to detect an effect of the magnitude seen in Experiments 1a and 1b. To account for a possible smaller effect size, given the possibility that a more effective reading of the second passage is occurring, 180 undergraduate students from the University of California, Los Angeles were recruited from the psychology department's subject pool for participation. Eleven participants were excluded for either their final test answers failing to be recorded by the computer or leaving all answers to the both the initial and final tests completely blank, resulting in a total of 169 participants remaining (44 male, 125 female; M_{age} = 20.5 years).

Design

There were two between-subjects conditions: Namely, the format of the initial test of the first passage, which occurred after both passages had been read, was either standard multiple-choice or confidence-weighted multiple-choice, as it has been in Experiments 1a and 1b.

Materials

The same two passages on Saturn and Yellowstone National Park (approximately 1200-word each) from Experiments 1a and 1b, as well as the same two 10-question sets of related question pairs for each passage (set A and set B) with competitive incorrect alternatives (when presented in multiple-choice format) as determined by Little et al. (2012) were used. These materials are presented in Appendix A.

Procedure

The procedure, diagramed in Figure 2.4, was similar to that of Experiments 1a and 1b with a reordering of the events. All participants were randomly assigned to either the standard multiple-choice or the confidence-weighted multiple-choice condition. If assigned to the confidence-weighted condition, participants were briefed on how to appropriately answer questions in the unfamiliar format and were not allowed to move on to the rest of the experiment until they had demonstrated understanding of the scoring system being used. As in Experiments 1a and 1b, participants were not presented with a *Don't Know* option; otherwise, presentation and scoring of the confidence-weighted multiple-choice format was identical to Sparck et al. (2016). (See Figure 2.1 for scoring system.) Participants in the standard multiple-choice condition were awarded one point for correct answers and no points for incorrect answers.

All materials were presented by means of a computer, and all participants began by reading two passages (order of reading counterbalanced across participants) for 9 minutes each.

After reading the second passage, everyone took an initial test, either using the standard multiple-choice format or the confidence-weighted multiple-choice format on items about the first passage. The initial test consisted of 10 questions from either question set A or question set B (counterbalanced across participants), and each question remained on the screen for 25 seconds (with a 10 second warning) before the participant was automatically moved on to the next question. After the initial test on the first passage was complete, all participants took a standard multiple-choice test on the second passage. After answering all of the questions on each initial test, participants were told their summary score for that initial test. Participants were not given any specific item-by-item feedback as to which questions were answered correctly or incorrectly.

After a 10-minute Tetris distractor task, everyone took a self-paced, cued-recall final test on the related question set from the second passage only. That is, if participants answered questions from set A on the initial test from the second passage, on the final test they answered questions from set B and no alternatives were provided. Given the construction of the materials as previously described, the correct answer to each question on the final cued-recall test always had appeared as an incorrect alternative to a question that was answered on the initial standard multiple-choice test for the second passage. Participants were encouraged to attempt to answer all questions even if they were unsure of their answers.

Following the completion of the experiment, participants answered survey questions regarding their strategy use during the experiment.

.

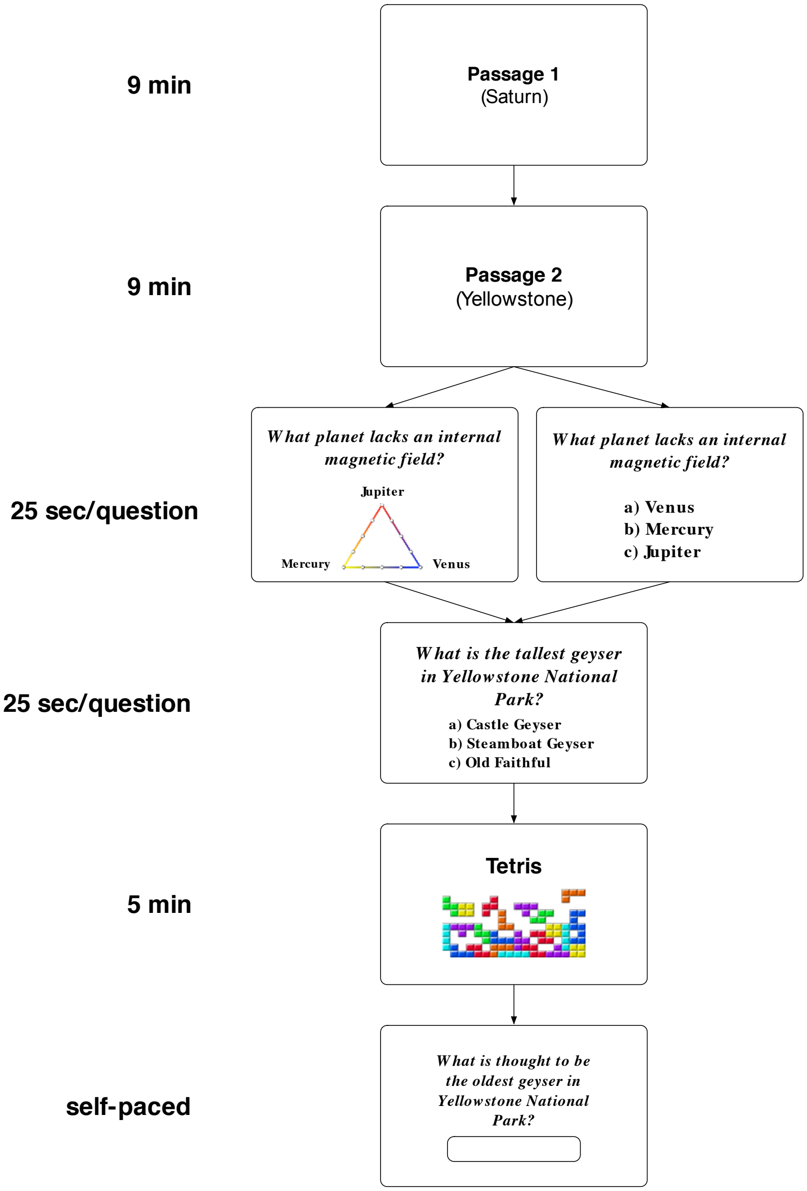
9 min
Passage 1
(Saturn)
9 min
Passage 2
(Yellowstone)
25 sec/question
What planet lacks an internal magnetic field?
Jupiter
Mercury
Venus
What planet lacks an internal magnetic field?
a) Venus
b) Mercury
c) Jupiter
25 sec/question
What is the tallest geyser in Yellowstone National Park?
a) Castle Geyser
b) Steamboat Geyser
c) Old Faithful
5 min
Tetris
self-paced
What is thought to be the oldest geyser in Yellowstone National Park?

Figure 2.4. Diagram of the procedure for Experiment 2. Participants began by reading both passages followed by the experimental manipulation of taking of either a standard multiple-choice or confidence-weighted multiple-choice initial test on information from the first passage. All participants then took a standard multiple-choice test on information from the second passage. Finally, participants engaged in a short Tetris distractor and then take a final cued-recall test on non-tested, related information from the second passage only.

Results

The cued recall responses on the final test were scored by an independent rater, who was blind to the participants' conditions, according to a lenient scoring guide to allow for slight misspellings and typos. No penalties for incorrect answers were assessed on the final test. The proportion of correct responses on the final cued-recall test, calculated on the basis of 10 items (related questions from the second passage). Independent sample *t*-tests were used in the analyses to compare the means of the two groups on both the initial and final tests. Results of Experiment 2 are shown in Figure 2.5.

As expected, an independent samples *t*-test showed that performance on the second, and always multiple-choice practice test, did not differ between the confidence-weighted (*M* = .61; *SD* = .21) and standard multiple-choice conditions (*M* = .61; *SD* = .20); *t*(167) = .15, *p* = .89. Participants in the confidence-weighted multiple-choice condition (*M* = .34, *SD* = .17) showed numerically higher performance on the cued-recall final test than participants in the standard multiple-choice condition (*M* = .30; *SD* = .19); however, the results of this independent samples *t*-test indicated this difference was non-significant; *t*(167) = 1.48, *p* = .14.

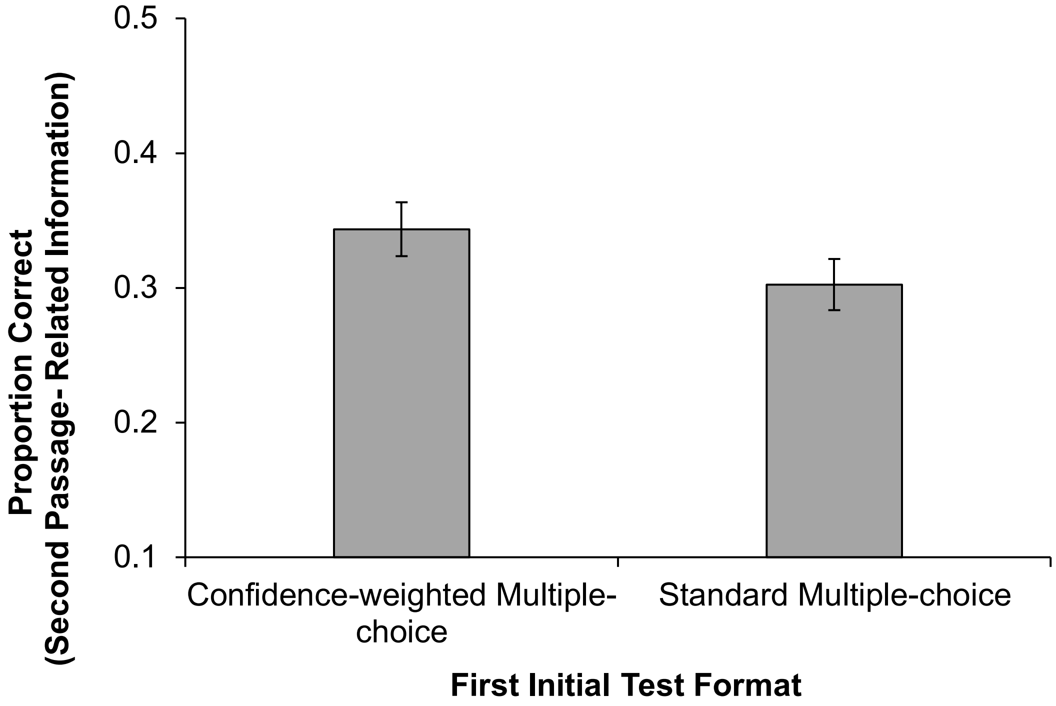

Figure 2.5. The results of Experiment 2. Proportion of correct answers to related questions from the second passage as a function of the initial test format (confidence-weighted multiple-choice or standard multiple-choice) taken after reading the first passage. Error bars represent ± 1 standard error of the mean.

Discussion

First, it should be noted that the final cued-recall test scores for both the standard and confidence-weighted multiple-choice conditions were lower compared to Experiments 1a and 1b. By nature of the design, blocking the reading of the passages together at the beginning of the experiment before taking any practice tests, creates a longer retention interval from the reading of the second passage to the final cued-recall test for Experiment 2 than for Experiments 1a and 1b. Once the second passage has been read there is approximately a 9-minute delay until the final cued-recall test in Experiments 1a and 1b compared with a delay of over 13 minutes for Experiment 2. This approximate 30% increase in the retention interval from reading the passage to taking the final test that occurred in Experiment 2 might explain the slightly lower final test performance that was observed.

There is also an additional gap of time between reading and being tested on the second passage (250 seconds, or just over 4 minutes) in Experiment 2 that is not present in Experiments 1a and 1b also due to the nature of the design. Although the delay difference for retrieving information on the practice test is relatively short between these two experiments, the ability to retrieve recently learned information drops rapidly right after learning; and thus, the delay between initial learning and a first retrieval attempt of that information may have effects on later retention, as can be seen in studies of expanding retrieval practice (e.g., Landauer & Bjork, 1978, Storm, Bjork, & Storm, 2010). Indeed a post-hoc independent samples *t*-test comparing the second initial standard multiple-choice test performance from Experiments 1a and 1b with Experiment 2 (see Table 2.1), showed performance on the second initial multiple-choice tests taken immediately after reading the passage to be significantly higher (*M* = .69, *SD* = .20) than performance on initial tests taken at the 250 second delay (*M* = .61, *SD* = .20) [*t*(376) = 3.74, *p* < .001, *d* = .39]. Such timing differences might contribute to the final test performance differences observed in the present experiments.

Performance on the second initial (and always standard) multiple-choice test did not differ in Experiment 2, as was intended to separate the effects of productive retrieval about incorrect alternatives during the second multiple-choice test and changes in encoding strategies during reading. Differences between the two conditions were not expected to emerge until the final test on related information. Any differences on the final test are intended to measure how much productive retrieval occurred about alternatives that were initially incorrect, but were then correct answers on the final test. Any differences between the groups seen in Experiment 2 can thus confidently be attributed to a strategy change occurring while answering questions on the second initial test.

Although performance on the final test between the two groups did not reach statistical significance, we can still observe the would-be effect size (*d* = .23) as a general measure of how large the potential difference between the two formats in the case of a Type II error. If this difference is the true benefit of related information due to retrieval during the practice test phase portion of the experiment only, the present experiment would be underpowered to detect such an effect. A post-hoc power analysis suggests that with an effect size of *d* = .23 and the present sample size, power to detect an effect is only around .3. The present sample size was determined by expecting an effect size slightly smaller than that of the average of Experiments 1a and b (*d* = .44 and .39, respectively; see Table 2.2).

These results suggest that some of the benefit for related information seen in Experiments 1a and 1b for the confidence-weighted multiple-choice group comes from a more effective reading of the subsequent passage after experience with the confidence-weighted multiple-choice format. It appears, however, that the effect size of the benefit stemming from more productive retrieval during the second multiple-choice test was overestimated and the effect size of the benefit related to a more effective study session was underestimated.

There are several potential explanations as to why the confidence-weighted multiple-choice test may have led to better encoding of the second passage. There might be some sort of novelty bump that happens after experiencing the confidence-weighted multiple-choice format that captures the participant’s attention. Additionally, perhaps the saliency of either the potential for a negative score or actually receiving a negative score after taking the first confidence-weighted multiple-choice test contributed to a closer reading of the next passage. While participants did not receive item-by-item feedback, they were told their overall score as a means of encouraging engagement during the experiment. Unlike in the confidence-weighted multiple-choice condition, in the standard multiple-choice test condition, there were no penalties for answering questions incorrectly. Whether the scoring system plays a role and how large it might be in seeing the benefits of confidence-weighted multiple-choice testing remains another crucial area of exploration. Given the present scoring system, participants in the confidence-weighted condition may have felt an additional pressure not to do poorly on a future test. Thus, the students in that condition may have been more motivated to pay attention to information in the second passage because of that testing experience.

Testing has also been shown to reduce the impact of certain cognitive biases during subsequent learning, leading to more effective self-regulated studying (Soderstrom & Bjork, 2014). Perhaps experiencing the confidence-weighted multiple-choice format reduces some of the illusions of competence (Koriat & Bjork, 2005) that the learners may have experienced while learning about the first passage to a greater extent than the standard multiple-choice format, leading to greater metacognitive sophistication for the confidence-weighted multiple-choice participants. More reflection during the confidence-weighted practice test regarding (the lack of) what was learned from their first reading efforts as they engage in relational processing of the alternatives could have encouraged participants to engage more deeply with material in the second passage.

Table 2.2

Effect size comparison for Experiments 1a, 1b, and 2

	Experiment		
Effect Size	**1a**	**1b**	**2***
Cohen's *d*	0.44	0.39	0.23

* *p > .05*

General Discussion

Taken together the results of Experiments 1a, 1b, and 2 suggest that experiencing a confidence-weighted multiple-choice test might lead to a change in strategy during the reading of subsequent texts and the taking of subsequent standard multiple-choice tests. While the result that taking a confidence-weighted multiple-choice test may have future benefits during forthcoming learning is promising, how long these benefits might last is unknown and should be explored in future research.

Presumably after one 10-question study session with the confidence-weighted multiple-choice format, college-age students will not permanently overhaul the multiple-choice test-taking strategy they have developed over many years of experience with standard multiple-choice tests, but perhaps its effects will only be seen in the immediate future. In Experiments 1a and 1b the second initial test comes only 9 minutes after the first initial test; and in Experiment 2, the second initial test immediately follows the first initial test. Both of the initial tests were also taken within the same episodic context (i.e., the participant is sitting in front the of the computer screen continuously without taking any sort of physical break between the tests), and thus may be more likely to be connected in a way that changes how the participant approaches the second passage and/or the initial test for that passage. If a substantial delay were to be implemented between the first initial test and the second reading and initial test, would the increased benefit to non-tested related information from the second passage persist for the confidence-weighted multiple-choice group? Additionally, if it were to persist, how long can the delay be such that the increased benefit for future learning continues?

Another variable that may impact the duration of the benefit is how much time is spent with the confidence-weighted multiple-choice format. Would multiple sessions answering questions using the confidence-weighted multiple-choice format have longer lasting impacts on the encoding of future information while taking standard multiple-choice tests? And what is the optimal length of each session? In the present studies, participants answered 10 confidence-weighted multiple-choice questions, and a benefit was seen. It remains to be seen whether just showing a few "practice" questions in the format, such as *What is the capital of British Columbia, Canada?* and walking through the proposed strategy that underlies the benefit leads to any changes.

Furthermore, would the benefit afforded to those initially trained in the confidence-weighted multiple-choice format be domain-specific or would it generalize to other fields of study? Say that a student develops a more effective strategy to use on standard multiple-choice tests after taking tests with a confidence-weighted format in one subject (e.g., biology). Would that student also apply that strategy to another distinct area of study (e.g., history), or instead, would that student revert back to using a previous strategy? Prior research suggests that even within a narrow domain, specificity of learning can occur and prevent transfer even under relatively similar conditions (e.g., Pan, Gopal, & Rickard, 2015; Rickard, Bourne, & Healy, 1994). Perhaps something similar might apply to strategy use, such that use of the more effective strategy only emerges under conditions where the type of knowledge expected to be learned is similar to the type of knowledge learned when the strategy was developed. If students do not recognize the broader benefits of using the effective strategy, they might only employ them under specific circumstances.

Understanding the answers to these outlined questions is an important step in translating laboratory research on nuanced testing effects, particularly those pertaining to initially non-tested, related information, to educationally realistic situations. Before broad recommendations can be made to instructors and students about the value, utility, and how to successfully implement confidence-weighted multiple-choice testing, further research on topics discussed is needed. These experiments, however, provide the foundational work for being able to make such recommendations.

Chapter 3: Investigating the Confidence-weighted Multiple-choice Format as a More Effective Pretest

While the benefits of testing individuals after original study has been well documented, less is known about the benefits of testing people before they have studied. Understanding the role that pretests, often involving the making of errors, might play in learning is an active area of investigation within the field of cognitive psychology, and particulrly, as it applies to educational settings. A general concern in this endeavor is whether the generation of errors or incorrect answers before study of the material to be learned, which is often incurred in the giving of some form of pretest, could lead to interference, thereby making the correct information even more difficult to learn than it would have been without such a pretest activity.

It appears, however, that pretesting information before it has been studied does not harm subsequent learning; and rather, it can actually potentiate the learning of that information during ensuing study across a variety of pretest formats (e.g., Kornell, Hays, & Bjork, 2009; Grimaldi & Karpicke, 2012; Potts & Shanks, 2014). Participants, for example, are better able to remember the weakly related paired associates *whale – mammal* after first seeing *whale –* _______ and making an (almost always) incorrect guess about what goes in the blank, relative to reading or studying the intact pair of *whale – mammal* for an extended period of time (Kornell et al., 2009). How such findings relate to more complex, educationally relevant materials where only a subset of to-be-learned information can realistically be pretested is of great interest. Another concern, specifically related to pretests given before the reading of text passages containing the directly pretested information, but also other important information to be learned, is that they may focus attention too narrowly on just the learning of the correct answers to the questions asked on the pretest to the detriment of learning the other information (e.g., Hamaker, 1986).

To investigate the role that the type of pretest plays in affecting information other than what was directly pretested, Little and Bjork (Experiment 1, 2016) compared the effects on later learning of giving multiple-choice pretest questions with competitive alternatives to cued-recall pretest questions and pre-study fact reading (both with and without alternatives present) and found that competitive, related information was better recalled on a final cued-recall test in the case when a multiple-choice pretest was given compared to when with a cued-recall pretest was given. Simple exposure to the alternatives by reading them before studying does not benefit learners in the same way. Additionally, differences could not be attributed to increased time spent studying that information (Experiment 2). Similar results have been reported outside of the laboratory using multiple-choice pretests as effective learning tools in the classroom (Bjork, Soderstrom, & Little, 2015).

Despite both standard multiple-choice pretests and posttests functioning to improve learning (Little & Bjork, 2016; Little et al., 2012), the mechanism by which the benefit occurs is thought to be different. During a pretest, learners have not yet learned any information, thus retrieval of relevant information, the proposed reason in Little et al. (2012) will be unsuccessful (unless the learner has sufficient prior knowledge, in which case, the pretest would be functioning more akin to a delayed posttest). Instead, prior research on the benefits of pretesting have suggested a more effective processing of the information during the subsequent reading of the passage to account for the benefits of a pretest. Little and Bjork (2016) propose that when learners do not have enough relevant information to know or deduce the correct answer to a multiple-choice pretest question, they may employ a variety of strategies in considering all the alternatives, such as choosing the most pleasant or the most personally relevant one, which have known benefits for memory (e.g., Craik & Tulving, 1975; Packman & Battig, 1978). Engaging in these sorts of deep processing of all the alternatives for a pretest multiple-choice question during the multiple-choice pretest may then make it easier or more efficient for the learner to acquire information regarding all the alternatives during study relative to answering cued-recall pretest questions or reading facts. It is also possible that pretesting might activate general preexisting knowledge regarding the pretested topic, and while that knowledge might not be directly relevant in helping the test taker to answer the pretest questions correctly, its activation might allow the test taker to learn new information about that topic more efficiently and effectively during the subsequent reading.

Experiment 3: Can the benefits of multiple-choice pretests be expanded through confidence-weighted testing?

The present research examines how benefits of pretesting might be affected by the format of the pretest used. Specifically, it compares potential pretesting benefits of confidence-weighted multiple-choice pretests with those of standard multiple-choice pretests. Sparck et al. (2016) posited that the confidence-weighted multiple-choice test format, operating as a posttest, encourages learners to retrieve information they have learned after reading a text (beyond what occurs during a standard multiple-choice test) in order, in part, to ensure they do not make a highly confident error and lose a large number of points. It is thought that such a format might encourage more retrieval, even when learners are relatively confident in their answer selection, in comparison to a standard multiple-choice format (without explicit instruction to engage in retrieval of information about the incorrect alternatives), which seems not to encourage that strategy spontaneously (Little 2011, Experiment 5).

In the case of pretesting, where no information has yet been learned and where the benefit might come from more efficient encoding of information in the yet-to-be-read materials, it is currently unknown whether the same increased benefits to non-directly tested information observed with confidence-weighted multiple-choice post-testing compared to those observed with standard multiple-choice post-testing, will also be seen with pretesting. Thus, the present study addresses whether confidence-weighted multiple-choice pretest questions might engage learners in processing new to-be-learned information better than standard multiple-choice pretest questions, analogous to how they seemed to encourage more productive retrieval processes in the post-test situation, as reflected in better final test performance after confidence-weighted multiple-choice initial posttests than after standard multiple-choice initial posttests as observed in Sparck et al. (2016).

Thinking about the possible relationships between the alternatives as directed by the confidence-weighted multiple-choice format might lead to deeper processing than just being presented with a list of potential alternatives after the question, even if relatively little to no information is known on the topic, thus acting as a more effective pretest. Further, it is expected that results from the present study will further clarify underlying causes of the differences obtained in the use of confidence-weighted multiple-choice questions versus standard multiple-choice questions. It was hypothesized that confidence-weighted multiple-choice testing should increase test-takers engagement with alternatives as compared to standard multiple-choice testing, ultimately allowing for even better encoding of the to-be-learned information.

As in Sparck et al. (2016; Experiment 2), the present experiment also includes the standard multiple-choice format with the addition of requiring numeric confidence judgments, as well as the standard multiple-choice format and the confidence-weighed multiple-choice format.

Including all three of these formats ensures being able to assess whether any greater benefits observed for the confidence-weighted multiple-choice format are coming from the type of processing being done about alternatives during the confidence-weighted multiple-choice pretest, rather than simply making a confidence judgment in isolation. Given the results of Sparck et al. (2016), no difference between the standard multiple-choice and the standard multiple-choice plus numeric confidence judgment formats were expected.

All pretesting conditions were compared with a study-only baseline control condition (similar to Sparck et al., 2016; Experiment 1). Taking some form of multiple-choice pretest is expected to lead to better final test performance (on related, but not directly pretested information) than is taking no form of pretest at all. Results will further contribute to the limited body of research suggesting that multiple-choice pretests can function as effective learning tools.

Method

Participants

The participants, all of whom served for partial course credit, were recruited from the University of California, Los Angeles psychology department's subject pool. Of the 215 participants who were recruited, 12 were excluded from data analysis due to technical error or leaving all answers to both the initial and final test questions blank, resulting in a total of 203 participants (43 male, 160 female; M_{age} = 20.6 years) remaining. It was estimated that a minimum of 180 participants would be needed to find an effect according to a power analysis by G*power assuming a medium effect size of f = .25, an alpha of 0.05, and a power of 0.80.

Design

Four between-subject conditions were used: study-only, standard multiple-choice, standard multiple-choice with numeric confidence judgment, and confidence-weighted multiple-choice (analogous to Experiment 2 in Sparck et al., 2016, with the addition of a study-only baseline control as in Sparck et al. Experiment 1).

Materials

The same two passages (one on Saturn and one on Yellowstone National Park, approximately 1200-words each) from Experiments 1a, 1b, and 2, as well as the same two 10-question sets of related question pairs for each passage (set A and set B) with competitive incorrect alternatives (when presented in multiple-choice format) as determined by Little et al. (2012) were used as the materials.

Procedure

Participants were randomly assigned to one of the four experimental conditions. Part way through data collection, a questionnaire to assess participants' prior knowledge of Saturn and Yellowstone National Park was introduced. Participants (*n* = 63) were asked to assess their prior knowledge on a Likert scale from 1 to 5 to ensure that the pretests were indeed functioning as pretests. A response of 1 on the scale indicated that participant strongly disagreed that they were knowledgeable about the topic, while a response of 5 indicated strong agreement that they were knowledgeable about the topic.

The procedure, diagramed in Figure 3.1, was similar to that of Experiment 2 in Sparck et al. (2016), except that the tests serving as initial posttests in the previous study were now presented before the reading of each passage. If assigned to the confidence-weighted condition, participants were briefed on how to answer questions appropriately in the unfamiliar format and were not allowed to move on to the rest of the experiment until they demonstrated understanding of the scoring system being used. The confidence-weighted multiple-choice condition was scored as previously outlined with the no option to select *Don't Know* version, but with one change.

Unlike in the previous studies, answers at the middle points were scored differently than answers at the vertices. Only three points were subtracted from each participant's score for incorrect answers marked along the line between the two incorrect alternatives, so that low confidence incorrect answers would not be punished as severely as confident ones given that the information had yet to be learned. This change was motivated by trying to ensure that participants treated the middle points differently than the vertices; otherwise, participants might have felt compelled to use the vertices more if the seeming risk of using some of the middle points was just as high. Highly confident incorrect answers at the triangle vertices still resulted in a 10-point deduction, as this answer would indicate participants incorrectly believed they had some prior knowledge about the answer question with a high degree of certainty.

The strategy described to confidence-weighted multiple-choice participants in both Sparck et al. (2016) experiments encouraged an attempt to eliminate one incorrect answer that they were sure was incorrect; and thus, by doing so, participants would be guaranteed nothing more than a 1-point loss (and only if they were ultimately unsure of the answer, but more confident in the remaining incorrect alternative relative to the correct answer). Effectively implementing such a strategy to get to the correct answer, however, would be impossible to implement on a pretest without some levels of relevant prior knowledge. In the present experiment, participants were walked through the same strategy of trying to eliminate an answer, but the wording of the instructions for using the confidence-weighted format was changed slightly to encourage the participant's best guesses based on their reasoning skills and any general prior knowledge of the topic they might already have.

Participants in the standard multiple-choice and standard multiple-choice plus numeric confidence conditions were awarded one point for correct answers and no points for incorrect answers. Participants in all pre-test conditions were told in the instructions that it was understood that they might not know the answers, but that they should do their best to get the highest score they possibly could. Aggregate scores were shown at the end of each pretest, but no specific item-by-item feedback was provided. Participants in the study-only condition simply began reading the passage, and then played Tetris for the same amount of time that the pretesting conditions answered questions about the second passage before it was read to equate the spacing between conditions.

After the instructions, participants answered initial pretest questions using the testing format they were assigned. The initial test consisted of 10 questions from either question set A or question set B (counterbalanced across participants) from the first passage. Each question remained on the screen for 25 seconds (with a 10 second warning) before the participant was automatically moved on to the next question. Participants then read the first passage. Afterwards, participants took the initial pretest (same format as before) for the second passage, and then read the second passage. Passage ordering was counterbalanced across participants.

After a 10-minute Tetris distractor task, everyone took a self-paced, cued-recall final test. For participants in one of the three pretesting conditions, questions were related to the pretest questions from each passage. Thus, for example, if participants answered questions from set A on the initial pretest for the second passage, then on the final test for the second passage they answered questions from set B but with no alternatives provided. As previously describe, given the construction of the materials, the correct answer to each question on the final cued-recall test had always appeared as an incorrect alternative to a question that was answered on the initial multiple-choice pretest for the second passage. No questions from the pretests were repeated on the final test. One set of questions for each passage was randomly selected for the study-only participants to answer.

The passages were tested in the same order as they had been read. Questions were blocked by passage, but their order was randomized within that block. Participants were encouraged to attempt to answer all questions even if they were unsure of their answers. Following the completion of the experiment, participants answered survey questions regarding their strategy use during the experiment.

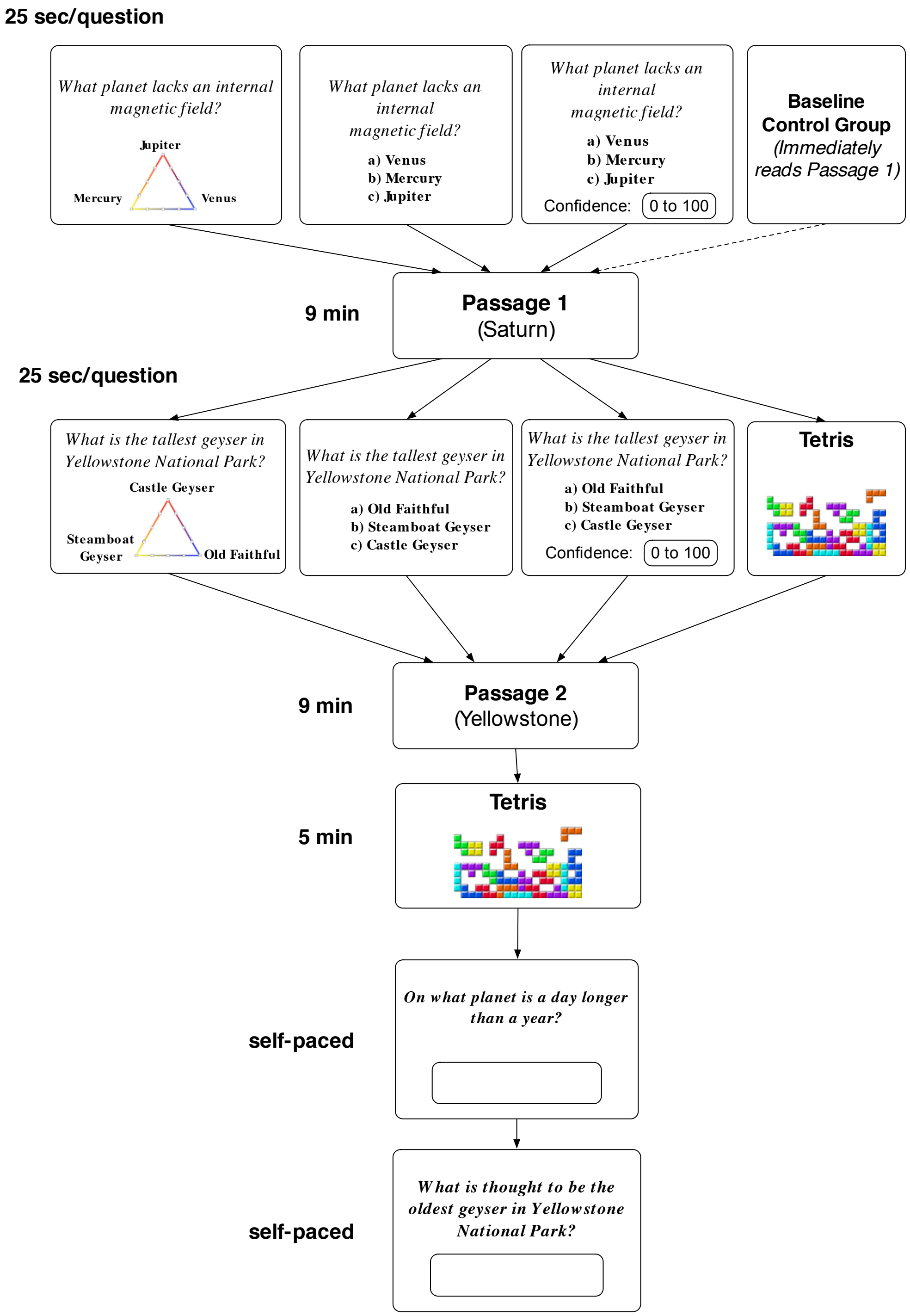
25 sec/question
What planet lacks an internal magnetic field?
Jupiter
Mercury
Venus
What planet lacks an internal magnetic field?
a) Venus
b) Mercury
c) Jupiter
What planet lacks an internal magnetic field?
a) Venus
b) Mercury
c) Jupiter
Confidence: 0 to 100
Baseline Control Group
(Immediately reads Passage 1)
9 min
Passage 1
(Saturn)
25 sec/question
What is the tallest geyser in Yellowstone National Park?
Castle Geyser
Steamboat Geyser
Old Faithful
What is the tallest geyser in Yellowstone National Park?
a) Old Faithful
b) Steamboat Geyser
c) Castle Geyser
What is the tallest geyser in Yellowstone National Park?
a) Old Faithful
b) Steamboat Geyser
c) Castle Geyser
Confidence: 0 to 100
Tetris
9 min
Passage 2
(Yellowstone)
5 min
Tetris
self-paced
On what planet is a day longer than a year?
self-paced
What is thought to be the oldest geyser in Yellowstone National Park?

Figure 3.1. Diagram of the procedure for Experiment 3. Participants in any of the pretest conditions attempted to answers questions about the passage before being allowed to read it, while participants in the baseline control condition immediately began reading the first passage. After reading the first passage, participants in the pretest conditions then attempted to answers

questions about the next passage before being allowed to read it, while participants in the baseline control condition played Tetris during this interval. All participants then read the second passage, after which they engaged in a short Tetris distractor task, and then took final cued-recall tests on each of the passages, which consisted of initially non-tested, but related information from both passages.

Results

Overall, participants reported low levels of prior knowledge. The average knowledge self-assessment score on the Likert scale for the topic of Saturn was 1.62 (median = 1; mode = 1), while the average self-assessment score on the Likert scale for the topic of Yellowstone National Park was 2.01 (median = 2; mode = 1). No participants rated their knowledge above 3; and only one participant answered 3 (for the Yellowstone passage), suggesting the possibility of a moderate level of knowledge. The final test performance for this participant, however, suggests that any self-assessed knowledge about the park was untrue or was not relevant, as performance was near, but slightly below, the average for all participants.

The results from this subsample suggest that most participants were not previously familiar with either of the topics that they were to learn about in the present experiment, other than perhaps very basic knowledge. There was no evidence suggesting a need to separate the data based on the participants' prior knowledge with the present topics, as that variable is likely playing little to no role in the current experiment. We can thus assume that pretests were indeed functioning as pretests.

The cued recall responses on the final test were scored by an independent rater, who was blind to the participants' conditions, and according to a lenient scoring guide to allow for slight misspellings and typos. No penalties for incorrect answers were assessed on the final test. The proportion of correct responses on the final cued-recall test, calculated on the basis of 20 items (10 items related to the pretested questions from each of the two passages), was .27 (*SD* = .14) for participants in the study-only group; .33 (*SD* = .17) for those in the standard-multiple choice group; .34 (*SD* = .17) for those in the standard multiple-choice plus numeric confidence judgment; and .34 (*SD* = .14) and for those in the confidence- weighted multiple-choice condition. Results from Experiment 3 are shown in Figure 3.2.

The results of the one-way omnibus analysis of variance (ANOVA) were significant, $F(3,199) = 2.87$, $p = .038$, $\eta^2 = .04$, indicating that there was a significant difference between at least two of the groups on the cued-recall final test. As predicted, a planned comparison, independent samples *t*-test test showed that participants in the pretest groups significantly outperformed participants in the study-only group, $t(201) = 2.92$, $p = .004$. Also as expected, a planned comparison, independent samples *t*-test showed no final test performance differences between the performance of participants in the standard multiple-choice and standard multiple-choice plus numeric confidence judgment conditions, $t(103) = -.263$, $p = .79$]. Critically, however, a planned comparison, independent samples *t*-test between the two standard multiple-choice conditions and the confidence-weighted multiple-choice condition also showed no final test performance differences, $t(150) = -.233$, $p = .82$.

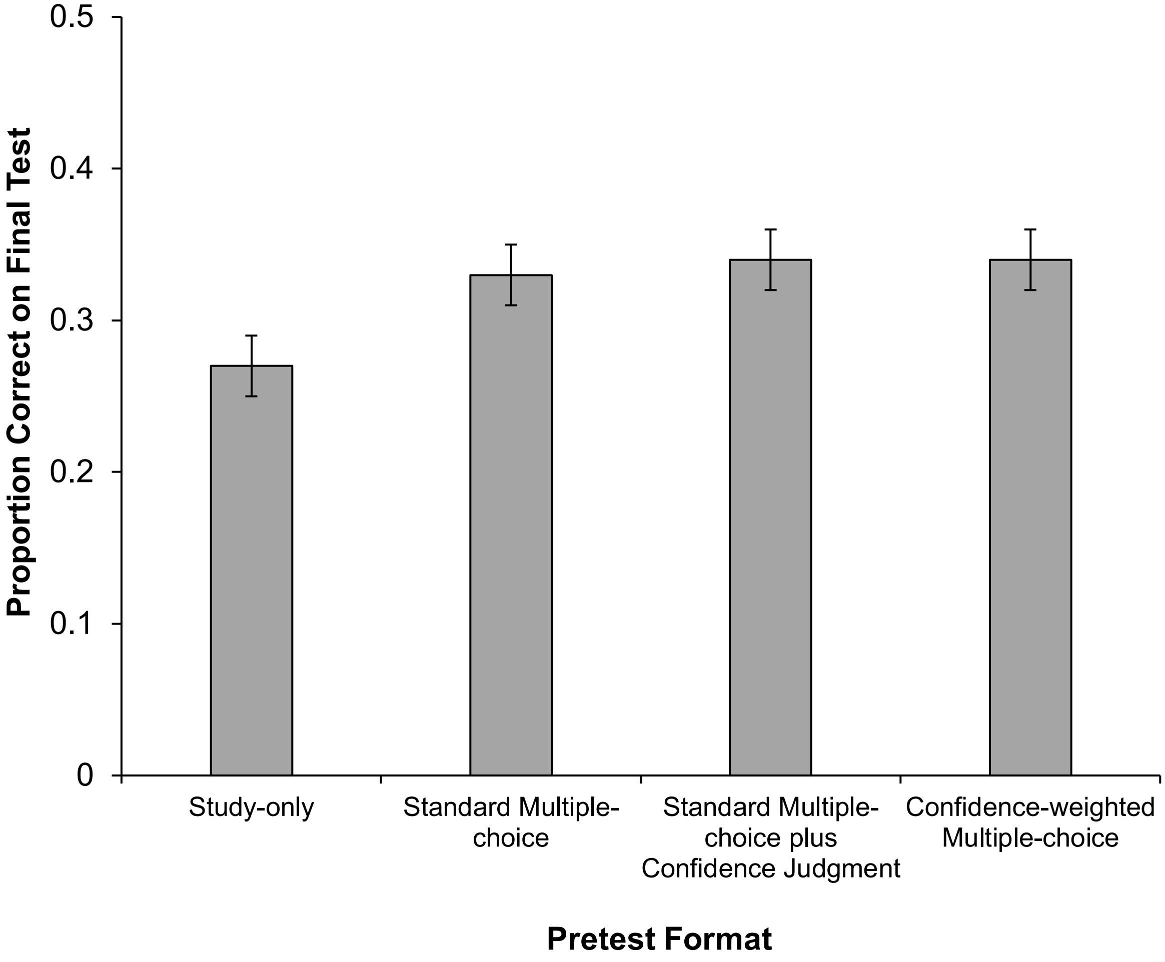

Figure 3.2. Results from Experiment 3. Proportion of questions correctly answered on the final, cued-recall test as a function of pretest format (study-only baseline control, standard multiple-choice, standard multiple-choice plus numeric confidence-judgment, confidence-weighted multiple-choice). Error bars represent ± 1 standard error of the mean.

Discussion

The present study examined whether confidence-weighted multiple-choice pretest questions would increase performance on a final cued-recall test relative to standard multiple-choice test pretest questions, and to ensure, such a benefit comes from the type of relational processing that the confidence-weighted multiple-choice condition affords, not just making a confidence judgment in isolation. Unlike the increased benefit seen in Sparck et al. (2016) for confidence-weighted initial test taken after learning, confidence-weighted multiple-choice pretests did not improve learning of information related to the incorrect alternatives relative to either standard multiple-choice pretests or to standard multiple-choice pretests where a numeric confidence judgment was provided after selection.

The pretest conditions in general, however, did show improved final test performance relative to the baseline study-only condition, replicating the finding that multiple-choice pretests can facilitate the learning of related information and do not just focus learners on the directly pretested information or create interference from the endorsement of incorrect answers. This experiment adds to our body of knowledge regarding multiple-choice pretests. The results suggest that the specific format of the multiple-choice pretest may not be a factor in the size of the benefit to non-tested information, whereas it can a play a significant role in post-testing.

One reason that learners might not always engage in effective strategies for competitive standard multiple-choice post-testing is that they may often believe they know the answer following their reading of the question and, thus, treat it more like a cued-recall question than a competitive multiple-choice question, attempting to retrieve the answer nd then simply matching what they produce with an alternative without fully considering and retrieving information about all of the options, which would amount to aprocess analogous to the narrow strategies outlined previously. Test takers may only resort to the broader strategy of trying to retrieve what they know about each of the alternatives when they do not immediately think they know the correct answeror when presented with the confidence-weighted multiple-choice format. The confidence-weighted multiple-choice format may lead learners to think in a broader, more nuanced way drawing upon any additional knowledge accessible to them as they consider their relative confidences in the various alternatives. Or, perhaps because the penalty for being highly confident and incorrect is so severe, learners might be more cautious before selecting an answer, ruling out the other possibilities by productively retrieving related information.

Perhaps when confronted with a multiple-choice pretest on a topic for which it is not expected they have already learned the relevant knowledge, and on which they themselves might fully expect to receive poor scores, then—regardless of the format of the pretest—they are led to adopt an answering strategy involving processing of all alternatives and arriving at their answer selections in the same (effective) way in both the standard multiple-choice condition and the confidence-weighted multiple-choice condition. As a result, they may then encode the information contained in the passage during their subsequent reading of it in the same manner. In such a case, no differences in final cued-recall test performance would emerge between standard multiple-choice and confidence-weighted multiple-choice questions, as was seen in the present experiment.

Chapter 4: Applying Benefits of Multiple-choice Testing to Vocabulary Learning

Between July 2013 and June 2014 over half a million individuals took the Graduate Record Exam (GRE) according to the most recent statistics provided by the creators of the exam (Educational Testing Service, 2014). The test consists of three major parts: analytical writing, quantitative reasoning, and verbal reasoning. One of the principal topics tested within the verbal reasoning section is vocabulary knowledge. Given the emphasis placed on vocabulary by the GRE, knowing the definitions associated with those difficult vocabulary words is important for achieving a high score.

Many college-age students report self-testing as a method for studying, although more so as a way of monitoring their memory, than as a tool for learning (Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007). It is thus likely that many GRE vocabulary learners use testing with flashcards as a way of preparing for the exam. Many GRE vocabulary flashcard systems exist on the market, traditionally listing the vocabulary word on one side and the definition on the other, analogous to cued-recall testing (if used appropriately).

The research on the benefits of multiple-choice testing as a pedagogical tool (e.g., Bjork et al., 2014; Little et al., 2012; Sparck et al., 2016) suggest that there may be a better and more efficient way to learn a large set of vocabulary words when you have a limited amount of time, and thus a limited number of test trials that you can administer to yourself, in your pursuit of this goal. Perhaps some sort of flashcard system that incorporates different forms of multiple-choice questions—where, for example, in attempting to select the correct word during testing, learners also select against competitive words, and as a result, retrieve and strengthen access to those associated definitions as well—could increase overall learning while not requiring additional (and often unavailable) extra time. Using the same number of practice trials (or, in this example, individual flashcards), a learner could strengthen access to not only the definitions of directly tested words, but also similar words that act as competitive multiple-choice alternatives (i.e., a broad retrieval strategy), compared with the traditional cued-recall style flashcard where one would only expect improved memory for the directly tested word and definition (i.e., a narrow retrieval strategy).

A multiple-choice flashcard system that encourages such processing might well be more efficient, and thus more effective, than a traditional flashcard system that simply lists the word on one side and the definition on the other side. Specifically, a flashcard system taking advantage of this added benefit of multiple-choice testing could help learners achieve access to a large (and constantly growing) pool of vocabulary words and their definitions under circumstances when time for studying is a limited resource. Research that underlies the creation of a more effective system for learning difficult vocabulary is the focus of Chapter 4.

Experiment 4: Can the benefits of multiple-choice testing be applied to learning difficult, confusable vocabulary words?

In Experiment 4, three different types of activities (test or study) are compared to assess their impact on vocabulary learning, both on words that are directly tested (or studied) and not directly tested (or studied). The cued-recall condition approximates traditional flashcards that are actively engaging the learner in retrieval practice before checking the correct answer, typically displayed on the back of the card. The study-only condition approximates either reading a list of vocabulary words and their associated definitions or traditional flashcards, but in a passive manner where the learner just flips over the flashcard and reads the answer without first attempting to retrieve the definition. The multiple-choice condition, which is expected to show the best performance for words that are not directly tested, is the first step in researching what might be a flashcard system that could maximize learning using the benefits afforded by multiple-choice testing, as seen in Little et al. (2012).

Method

Participants

Eighty-five undergraduate students recruited from the psychology department's subject pool at the University of California, Los Angeles participated in this experiment for partial course credit. All participants were fluent in English and reported never having prepared for the GRE. No other demographic information was recorded for this experiment.

Design

The experiment involved a 3 (initial activity: multiple-choice, cued-recall, and study-only) x 2 (item type: directly tested/studied during initial activity phase and not directly tested/studied during initial activity phase) within-subjects design.

Materials

The materials, all of which were presented by means of a computer, consisted of 36 GRE vocabulary words and their definitions, divided into nine groups of four words each. All of the words in each word group began with the same letter to increase confusability and competitiveness among the words within a group (e.g., *abnegate*, *aver*, *allay*, and *abet*) and to make relying on a superficial recognition strategy more difficult. Each word group was represented by a unique first letter, and all of the words within a word group were selected to be the same part of speech. The words were pilot tested in a previous experiment and determined to be sufficiently competitive with one another.

For each participant, two of the words from each word group were randomly selected to be directly tested/studied during the initial activity phase, and the other two words were assigned to be not directly tested/studied. For each participant, three words groups were randomly assigned to each of the three initial activity types: multiple-choice test, cued-recall test, and study-only, but all trial types (and therefore all of the words) were randomly mixed together, so the participant was not tested on or studying all the words from a word group in a single block, in order to encourage more retrieval.

Two sentences were constructed using each of the words. One of these sentences was used during the initial activity phase for the directly tested/studied words. The other sentence was used during the final test for the directly tested/studied words. The ordering of the two sentences was counterbalanced across participants. One of the sentences was randomly selected during the final test for the not directly tested/studied words.

The vocabulary words and sentences used are listed in Appendix B.

Procedure

The experiment consisted of three phases: the study phase, the initial activity phase, and the final test phase, and the overall procedure used is diagramed in Figure 3.2.

For the study phase, participants began by seeing the vocabulary words and their definitions together on the screen (e.g., *Abnegate: refuse, reject*), one at a time for 8 seconds each in a random order. Once all 36 words had been studied one time, this process was repeated using a new random order, so that in total, each word was studied by all participants twice. Participants played Tetris for 5 minutes as a short distractor in between the two study sessions.

Next, as part of the initial activity phase, participants engaged in one of the three initial activities for the directly tested words, either studying or retrieving the words in the context of a sentence. For words in the word groups assigned to the study-only condition, an intact sentence appeared using one of the studied vocabulary words with the correct vocabulary word underlined (e.g., *Hopefully, the company's soaring stock price will allay the concerns of nervous stockholders.*). Participants were told to study the word and sentence for a total of 25 seconds. For words in the word groups assigned to the cued-recall and multiple-choice conditions, a sentence appeared with a missing word (e.g., *Hopefully, the company's soaring stock price will _____________ the concerns of nervous stockholders.*), and participants were asked to fill in the blank with one of the previously studied words (e.g., *allay*) based on the context of the sentence. For words in the cued-recall condiiton, only the sentence was visible. For words in the multiple-choice condition four alternatives, all belonging to the same word group and thus beginning with the same letter, appeared beneath the sentence. After 20 seconds, the correct answer appeared in the blank, and participants had 5 seconds to study the intact sentence. Time on task during the initial activity was thus kept constant across all conditions.

After a 10-minute Tetris distractor, participants were tested on all 36 words (in a random order) by being asked to fill in the blank of a new sentence with one of the studied words based on context (identical to what would be seen on the actual GRE for increased ecological validity). The final test looked identical in format to the multiple-choice condition with the four alternatives (one correct, three incorrect but from the same word group) appearing beneath a new sentence missing the GRE vocabulary word. The final test was self-paced.

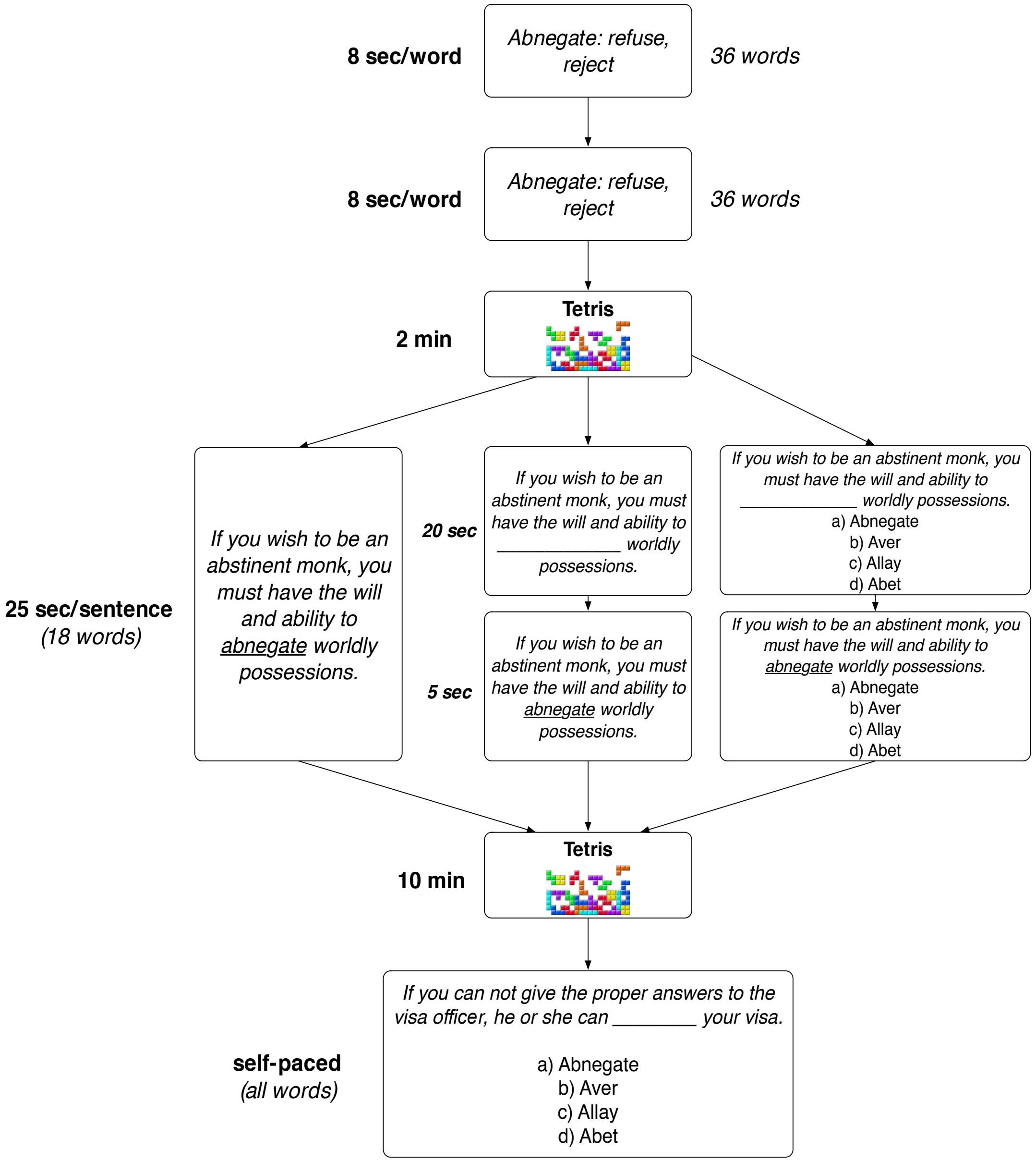
8 sec/word
Abnegate: refuse, reject
36 words
8 sec/word
Abnegate: refuse, reject
36 words
Tetris
2 min
25 sec/sentence
(18 words)
If you wish to be an abstinent monk, you must have the will and ability to abnegate worldly possessions.
20 sec
If you wish to be an abstinent monk, you must have the will and ability to ____________ worldly possessions.
5 sec
If you wish to be an abstinent monk, you must have the will and ability to abnegate worldly possessions.
If you wish to be an abstinent monk, you must have the will and ability to ____________ worldly possessions.
a) Abnegate
b) Aver
c) Allay
d) Abet
If you wish to be an abstinent monk, you must have the will and ability to abnegate worldly possessions.
a) Abnegate
b) Aver
c) Allay
d) Abet
Tetris
10 min
self-paced
(all words)
If you can not give the proper answers to the visa officer, he or she can ________ your visa.
a) Abnegate
b) Aver
c) Allay
d) Abet

Figure 4.1. Diagram of the procedure for Experiment 4. All participants studied all of the vocabulary words two times, followed by a brief Tetris distractor. Half of the words were selected to be initially tested/studied by each participant. Each of those words was assigned to be studied in the context of an intact sentence, tested in the cued-recall format in the context of a sentence, or tested in the multiple-choice format in the context of a sentence. Test trials were given feedback as to the correct answers. After a short Tetris distractor, participants were given

a final multiple-choice test where they were to select the word that correctly completed the sentence.

Results

The results for Experiment 4 are diagramed in Figure 4.2, where correct performance on the final test multiple-choice test is shown in relation to the three initial types of activity.

A 3 (initial activity: multiple-choice, cued-recall, and study-only) x 2 (item type: directly tested/studied during initial activity phase and not directly tested/studied during initial activity phase) repeated measures ANOVA was used to analyze the data. As expected due to more exposure, there was a main effect of item type, such that directly tested/studied words during the initial activity phase were answered correctly more often on the final test (*M* = .56; *SD* = .26) than words that were not directly tested/studied (*M* = .46; *SD* = .23); *F*(1,84) = 41.02, $p < .001$, $\eta_p^2 = .33$. No significant main effect for initial activity emerged; *F*(2,168) = 2.52, $p = .084$; but a significant item type X initial activity interaction was observed; *F*(2,168) = 10.37, $p < .001$, $\eta_p^2 =$ .11.

Of most interest for the current line of research is how the information that was competitive, but not directly tested fared. Planned comparison *t*-tests were used to compare condition means. Questions about the non-tested multiple-choice words, which had appeared as incorrect alternatives during the initial activity phase, were answered correctly (*M* = .51; *SD* = .22) more often than were questions about the non-tested words assigned to the cued-recall (*M* = .43; *SD* = .24), [*t*(84) = 2.94, $p = .004$] or study-only conditions (*M* = 43.3%; *SD* = 23.3%) [*t*(84) = 3, $p = .004$]. No difference emerged for the recall of non-tested/studied words in the cued-recall and study-only conditions [*t*(84) = -.3, $p = .77$].

Questions about the directly tested words in the cued-recall condition (*M* = .62; *SD* = .28) were answered correctly on the final test significantly more than directly tested words in the multiple-choice (*M* = .54; *SD* = .24) [*t*(84) = 3.16, *p* = .002] and study-only conditions (*M* = .53; *SD* = .25) [*t*(84) = 2.78, *p* = .007]. There was no difference between directly tested words in the multiple-choice condition and the study-only condition *t*(84) = .2, *p* = .84, although our delay was short (10 minutes).

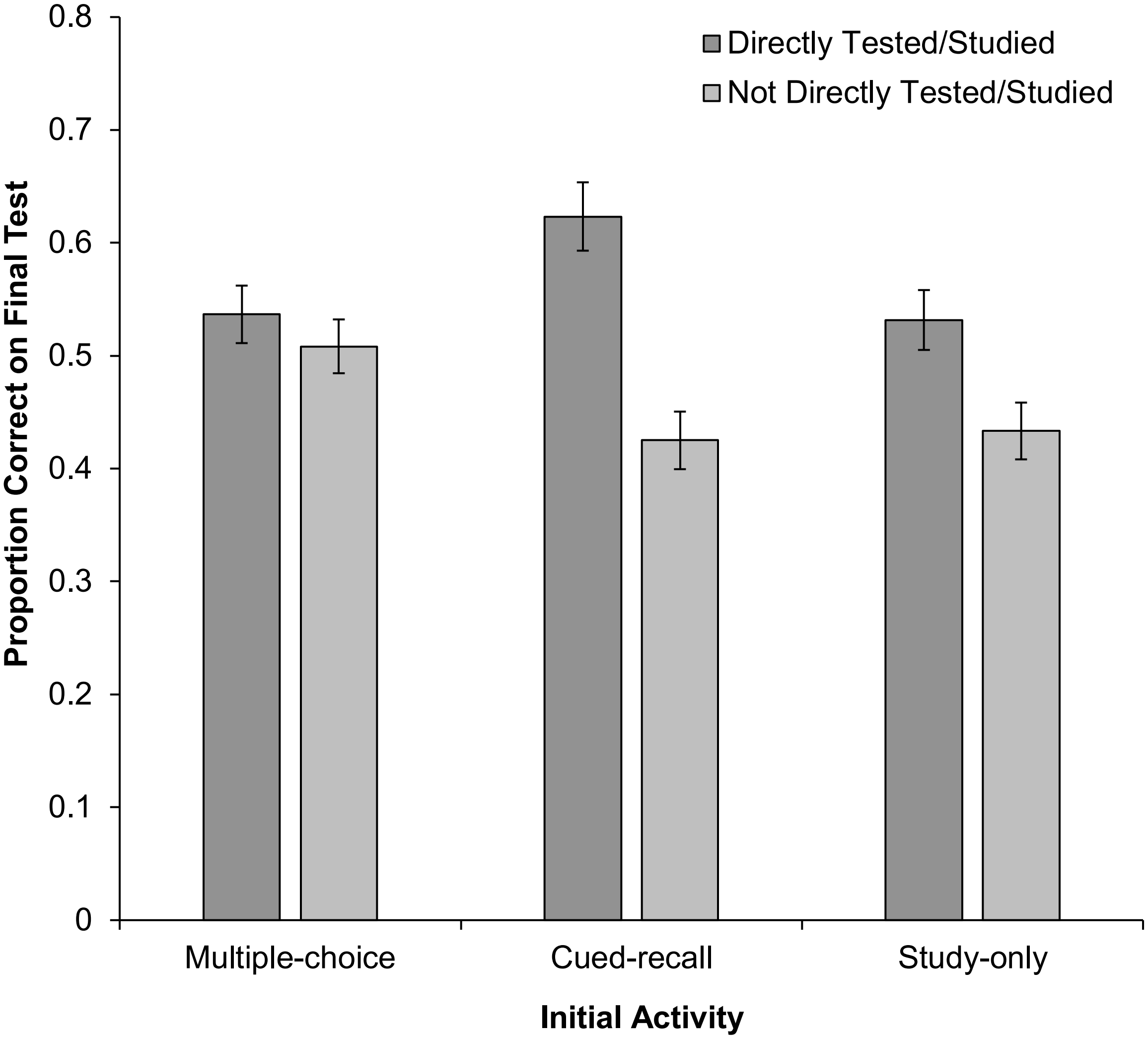

Figure 4.2. Results from Experiment 4. Proportion of words correctly used on the final multiple-choice/sentence completion test as a function of the type of initial activity (multiple-choice, cued-recall, or study-only). Darker bars represent the correct proportion of directly tested/studied words. Lighter bars represent the correct proportion of not directly tested/studied words. Error bars represent ± 1 standard error of the mean.

Discussion

The results from Experiment 4 suggest that when presented with difficult and confusable vocabulary words that begin with the same letter in the multiple-choice test format, learners do retrieve the definitions to the competitive incorrect alternatives, as is seen with fact-based information learned from coherent text (e.g., Little et al., 2012). More research is needed to determine if confidence-weighted testing can further increase the benefits to non-tested, competitive vocabulary words in the same way it did for the factual information learned from text (Sparck et al., 2016).

Although multiple-choice testing did not produce better overall performance in the present study, the not directly tested words in the multiple-choice condition outperformed those in the cued-recall and study-only conditions. Multiple-choice testing may be an effective way to study vocabulary words when the learner has a limited number of test trials relative to the total number of words that could potentially be tested (as would typically be the case with the GRE). In the current study in which it was found that directly tested words fared significantly better in the cued-recall condition whereas not-directly tested words fared better in the multiple-choice condition, an equal number of directly tested and not directly tested words were used. Thus, it is possible that were more words to be in the not-directly-tested pool, multiple-choice testing might reveal an advantage.

To summarize, cued-recall retrieval practice appears to make the learner better at recalling definitions of the directly tested words on the final test than does multiple-choice retrieval practice—perhaps because it constitutes a more difficult retrieval attempt with fewer cues, which may support better later retention (e.g., Carpenter & DeLosh, 2006). But it is also a narrow strategy; and as a result, it would only benefit those few cued-recall words unless participants were spontaneously recalling definitions to similar, studied words, but that appears from the pilot study not to be the case. There could thus be competing dynamics that lead to overall similarities on final test performance after taking cued-recall and multiple-choice initial practice tests when the number of words that are directly tested equals the number of words that are not directly tested.

Multiple-choice testing, however, may encourage a broad retrieval strategy and strengthen access to a greater number of words: not only the directly tested word (although perhaps not to the extent of cued-recall), but also to the competitive incorrect alternatives that could all potentially be tested in the future. Multiple-choice testing may thus be a more efficient use of a learner's time when the set of information that may potentially be tested is far larger than what is capable of being learned in a relatively short time. Confidence-weighted multiple-choice testing may even further increase this benefit. Both of these issues will be explored as part of Experiment 5.

Experiment 5: Can confidence-weighted multiple-choice testing aid in the learning of vocabulary words?

Previously, confidence-weighted multiple-choice testing has primarily been applied to initial tests presented after the reading of text passages where factual information is heavily integrated (e.g., Sparck et al., 2016)—the exception being the previously presented study exploring the use of confidence-weighted multiple-choice questions as pretest. Integration of information into a coherent passage has been shown to protect against retrieval-induced forgetting (e.g., Anderson & McCulloch, 1999; Chan, 2009; Little, Storm & Bjork, 2011), and—as shown in Chan et al. (2006) and Little et al. (2012)—can facilitate the retrieval of information that is related, but not directly tested. Experiment 5 focuses on whether the benefits of confidence-weighted multiple-choice testing can be applied to other types of information, specifically a type of paired-associates learning—vocabulary learning. In Experiment 5, the confidence-weighted multiple-choice format is compared to the standard multiple-choice format, as well as the cued-recall format and a study-only condition.

As in Experiment 4, some words were directly tested or studied. In the present study, however, more words are not directly tested during the initial activity phase, so as to allow evaluation of the efficiency question more directly. Another goal of Experiment 5 is to test whether multiple-choice tests are generally more effective than either cued-recall or study-only trials, which both promote narrow study strategies, when trying to learn a pool of words with a very limited number of trials.

Another difference between Experiments 4 and 5 is that participants in Experiment 5 were given no feedback after they engaged in the initial activity. This change in procedure was motivated by trying to better isolate the effects of testing on learning from the effects of feedback after testing. The role of feedback could not be separated out by the design of Experiment 4. Memory benefits may have appeared during both the retrieval practice and feedback stages of Experiment 4.

Additionally, the final test occurred at a longer delay (between 48 and 72 hours), so that an overall main effect of testing over studying would be more likely to emerge. The benefits of testing and other *desirable difficulties* do not always emerge with short delays (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006). Shorter delays measure immediate performance, whereas longer delays are a better measure of more durable learning. Soderstrom and Bjork

(2015) offer a comprehensive review explaining the dynamics for why the retention interval is an important consideration when studying the testing effect.

Overall, it is expected that a main effect of item type would be observed, such that directly tested/studied words will be recalled correctly on average more often than words that are not directly tested/studied, as was seen in Experiment 4. Given that the present experiment has a greater number of not directly tested/studied words (unlike in Experiment 4), a main effect of initial activity is expected, with confidence-weighted multiple-choice participants performing the best, followed by standard multiple-choice, cued-recall, and finally, study-only participants.

A significant interaction is also predicted to emerge. For the directly tested words, the cued-recall condition participants are expected to outperform participants in all of the other conditions (both multiple-choice conditions by a small margin and the study-only condition by a large margin). The standard-multiple choice and confidence-weighted multiple-choice participants are expected to outperform the study-only condition participants (but both multiple-choice condition participants will not differ significantly from one another) at a multiple day delay.

For not-directly-tested words, the standard multiple-choice and confidence-weighted multiple-choice condition participants are expected to outperform participants in both the cued-recall and the study-only conditions, with no difference between the cued-recall and study-only condition participants. Participants in the confidence-weighted multiple-choice condition are expected to outperform the standard multiple-choice condition participants if the strategies used during the initial vocabulary test are similar to those used by participants in Sparck et al. (2016) when tested on factual information from a text passage.

Method

Participants

One hundred and eighty undergraduate students at the University of California, Los Angeles were recruited from the psychology department's subject pool. Sample size was determined by a power analysis in G*power assuming a medium effect size of $f = .25$, an alpha of 0.05, and a power of 0.80. The data from 27 of these participants were excluded due to their not completing both sessions of the experiment, not attempting to answer all the questions on the final test, or not following final test instructions (e.g., generating their own non-studied synonyms for all of the presented definitions), resulting in data from a total of 153 participants (33 male, 118 female, 2 other; $M_{\text{age}} = 20.1$ years) remaining for analyses. All participants reported being fluent in English and never having taken or prepared for the GRE.

Design

This experiment employed a 4 (initial activity: study-only, cued-recall, standard multiple-choice, and confidence-weighted multiple-choice conditions) x 2 (item type: directly tested/studied and not directly tested/studied) mixed-subjects design. Initial activity was manipulated as a between-subjects variable while item type was manipulated as a within-subjects variable.

Materials

The materials, all of which were presented by means of a computer, consisted of 36 GRE vocabulary words and their definitions, divided into 12 groups of three words each since the confidence-weighted multiple-choice format in its current design only supports three alternatives. Twenty-seven of the words were borrowed from Experiment 4. The word with highest final test performance (i.e., the easiest word) on the final test of Experiment 4 from eight out of the nine word groups was dropped for the current experiment. From the remaining group, one word was selected for removal because of its similar spelling to another word in the group.

Nine new words (forming three new word groups) were added. All of the words in each word group began with the same letter to increase confusability and competitiveness among the words within a group (e.g., *abnegate*, *aver*, and *allay*) and to make relying on a superficial recognition strategy more difficult. Each word group was represented by a unique first letter, and all of the words within a group were selected to be the same part of speech (as in Experiment 4). For each participant, one of the words from each word group was randomly selected for direct testing/studying during the initial activity phase, and the other two words were assigned to be not directly tested/studied. The vocabulary words are listed in Appendix B.

Procedure

The overall procedure is diagramed in Figure 4.3. Participants were randomly assigned to one of the four initial activity conditions. If assigned to the confidence-weighted condition, participants were briefed on how to appropriately answer questions in the unfamiliar format and were not allowed to move on to the rest of the experiment until they demonstrated understanding of the scoring system being used. As in Experiments 1-3, participants were not allowed to select the *Don't Know* option; otherwise presentation and scoring of the confidence-weighted multiple-choice format was identical to Sparck at al. (2016), and the updated scoring guide is shown in Figure 2.1. Participants in the standard multiple-choice and cued-recall conditions were awarded one point for correct answers and no points for incorrect answers.

Similar to Experiment 4, the experiment consisted of three phases: the study phase, the initial activity phase, and the final test phase.

For the study phase, participants began by seeing the vocabulary words and their definitions together on the screen (e.g., *Abnegate: refuse, reject*) in a random order, one at a time for 5 seconds each. Once all 36 words were studied one time, this process was repeated in a new random order, so that each word was studied twice by all participants. Participants played Tetris for 5 minutes as a distractor task after the study phase.

Next, as part of the initial activity phase, participants engaged in one of the four initial activities for the directly tested/studied words, retrieving or studying the definition of those words. For participants assigned to study-only condition, the vocabulary word and definition was presented intact, just as it was during the study phase, but this time for 20 seconds. For participants assigned to one of the multiple-choice testing conditions, either a standard multiple-choice test or a confidence-weighted multiple-choice test appeared showing the definition to one of the vocabulary words along with three alternatives (all from the same word group) beneath the definition. Participants were given 20 seconds and asked to select the answer they believed was correct. For participants in the cued-recall testing condition, a definition was shown without any alternatives. Participants were given 20 seconds to type in the correct word. A countdown timer appeared, letting participants know when they had 10 seconds remaining to type an answer. Aggregate feedback consisting of the final score was shown at the end of this phase, but no item-by-item feedback was given.

For the final test phase, participants were tested on the definitions of all 36 words using a self-paced, cued-recall test. Participants were shown a definition and asked to type in the studied GRE word that matched. The 24 words that were not directly tested or studied in the previous phase were tested first as a way to control for output interference. The order of those 24 words was randomized for each participant. The 12 directly tested words were tested last, and also in a random order for each participant. Participants were encouraged to attempt to answer all questions even if they were unsure of their answers.

Participants completed the study and initial activity phases of the experiment, back-to-back, in the lab; however, they did not have to return to the lab for the final test phase. Rather, they were emailed a link and were able to complete the final test from their own computer. This link was sent 48 hours after completion of the initial activity phase. A longer delay was chosen to maximize the benefits of testing. Participants had 24 hours from the time they were sent the link to complete the final cued-recall test, making the delay somewhere between 48 and 72 hours for all participants.

Following the completion of the experiment, participants answered survey questions regarding their strategy use during the experiment and their own studying, as well as metacognitive questions regarding their beliefs about testing and learning vocabulary.

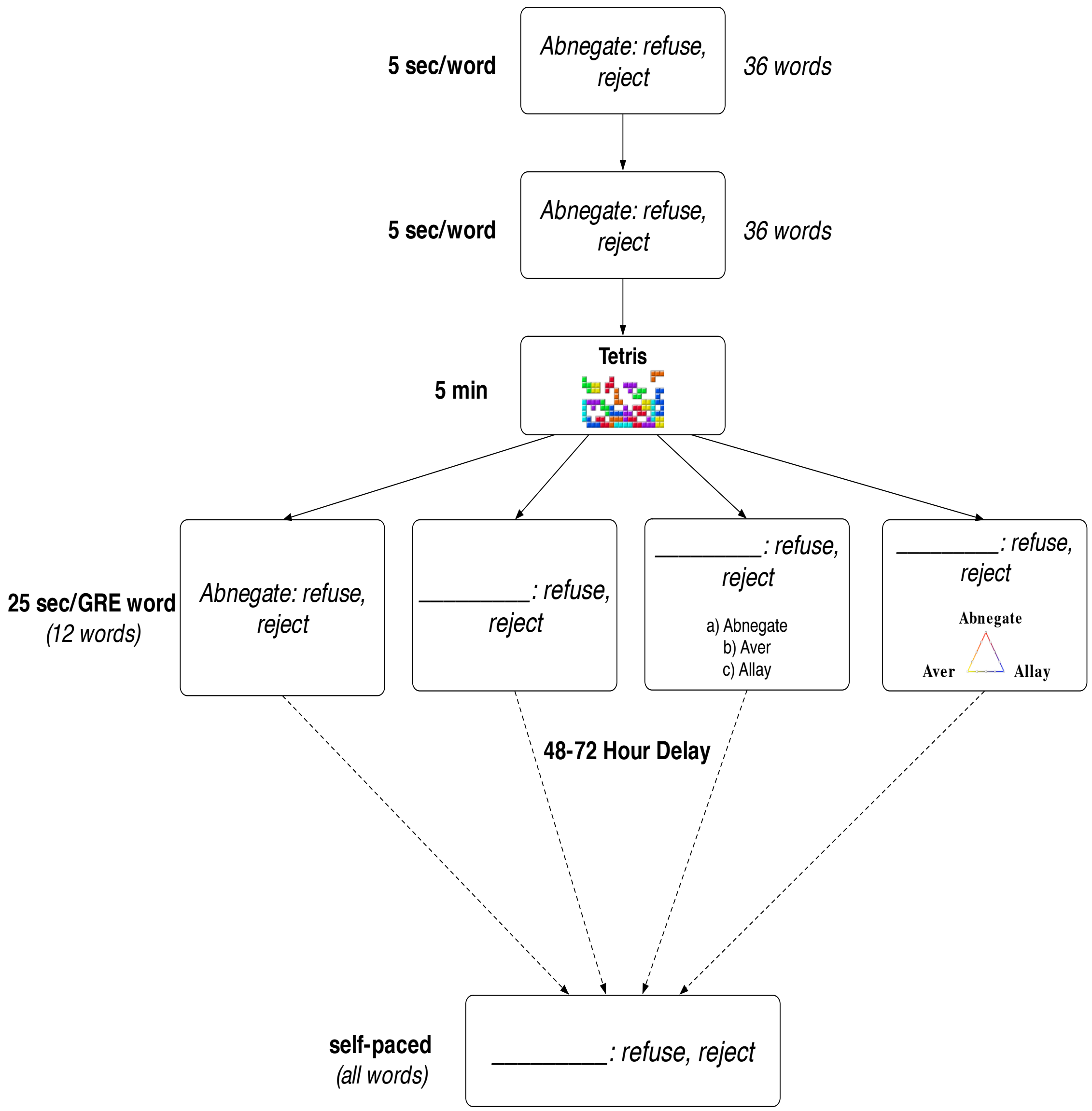

Figure 4.3. Diagram of the procedure for Experiment 5. All participants studied all of the vocabulary words two times, followed by a brief Tetris distractor. Participants next engaged in an initial test or in further study (i.e., the study-only, cued-recall, standard-multiple-choice, or confidence-weighted multiple-choice conditions) on one-third of the words. At a delay of 48-72 hours, participants were tested on all words in a cued-recall format beginning with the initially non-tested words to control for output interference.

Results

The cued recall responses on the final test were scored by a rater, blind to the participants' conditions, according to a lenient scoring guide to allow for slight misspellings and typos. No penalties for incorrect answers were assessed on the final test.

A 4 (initial activity: study-only, cued-recall, standard multiple-choice, and confidence-weighted multiple-choice conditions) x 2 (item type: directly tested/studied and not directly tested/studied) mixed ANOVA, followed by planned comparison and post-hoc *t*-tests were used to analyze the data. Results are shown in Figure 4.4.

As predicted, the omnibus ANOVA revealed a significant main effect of initial activity, such that directly tested/studied items ($M = .20$; $SD = .12$) were recalled better than not directly tested/studied items ($M = .16$; $SD = .15$); $F(1,148) = 15.54$, $p < .001$, $\eta_p^2 = .10$. The initial activity X item type interaction was also significant [$F(3,148) = 2.74$, $p = .046$, $\eta_p^2 = .05$].

For directly tested items, planned comparison *t*-tests showed there was no significant difference between the proportion of words recalled in the cued-recall ($M = .19$; $SD = .16$) and the standard multiple-choice conditions ($M = .23$; $SD = .17$) [$t(75) = -.83$, $p = .79$], or the confidence-weighted multiple-choice condition ($M = .21$; $SD = .16$) [$t(68) = -.50$, $p = .62$]. There was no significant difference between the proportion of words recalled in the cued-recall and the study-only conditions ($M = .16$; $SD = .12$)[$t(76) = 1.35$, $p = .18$]. There was a significant difference between the proportion of words recalled in the standard multiple-choice and the study-only conditions [$t(81) = 2.32$, $p = .02$, $d = .51$], while there was a marginal difference between the confidence-weighted multiple-choice condition [$t(74) = 1.95$, $p = .055$, $d = .44$]. There was not a significant difference between the proportion of words recalled in the two multiple-choice formats [$t(73) = .33$, $p = .74$]. There was a significant testing effect for directly tested items, such that being tested either in one of the multiple-choice or cued-recall formats, as an initial activity (*M* = .21; *SD* = .16) was better than studying (*M* = .16; *SD* = .12) [*t*(151) = -2.22, *p* = .03, *d* = .43].

For items that were not directly tested, there was no significant difference between the proportion of words recalled in the standard multiple-choice (*M* = .19; *SD* = .14) and the confidence-weighted multiple-choice (*M* = .22; *SD* = .13) conditions [*t*(73) = -.83, *p* = .41]. There was also no significant difference between the proportion of words recalled in the cued-recall (*M* = .11; *SD* = .11) and study-only conditions (*M* = .11; *SD* = .10) [*t*(76) = .38, *p* = .70]. There was, however, a significant difference between taking some sort of multiple-choice test (*M* = .20; *SD* = .14) versus either a cued-recall test or studying (*M* = .11; *SD* = .11) [*t*(151) = 4.71, *p* < .001, *d* = .76].

A pair of post-hoc paired-samples *t*-test with a Bonferroni correction showed that for participants taking a multiple-choice test as their initial activity, there was no significant difference in the proportion of words recalled on the final test for directly tested items (*M* = .22; *SD* = .17) and not directly-tested items (*M* = .20; *SD* = .14) [*t*(74) = 1.11, *p* = .27]. For participants taking either a cued-recall test or studying as their initial activity, there was a significant difference in the proportion of words recalled on the final test for the directly tested/studied items (*M* = .17; *SD* = .14) and the not directly tested/studied items (*M* = .11; *SD* = .11) [*t*(77) = 4.85, *p* < .001, *d* = .55].

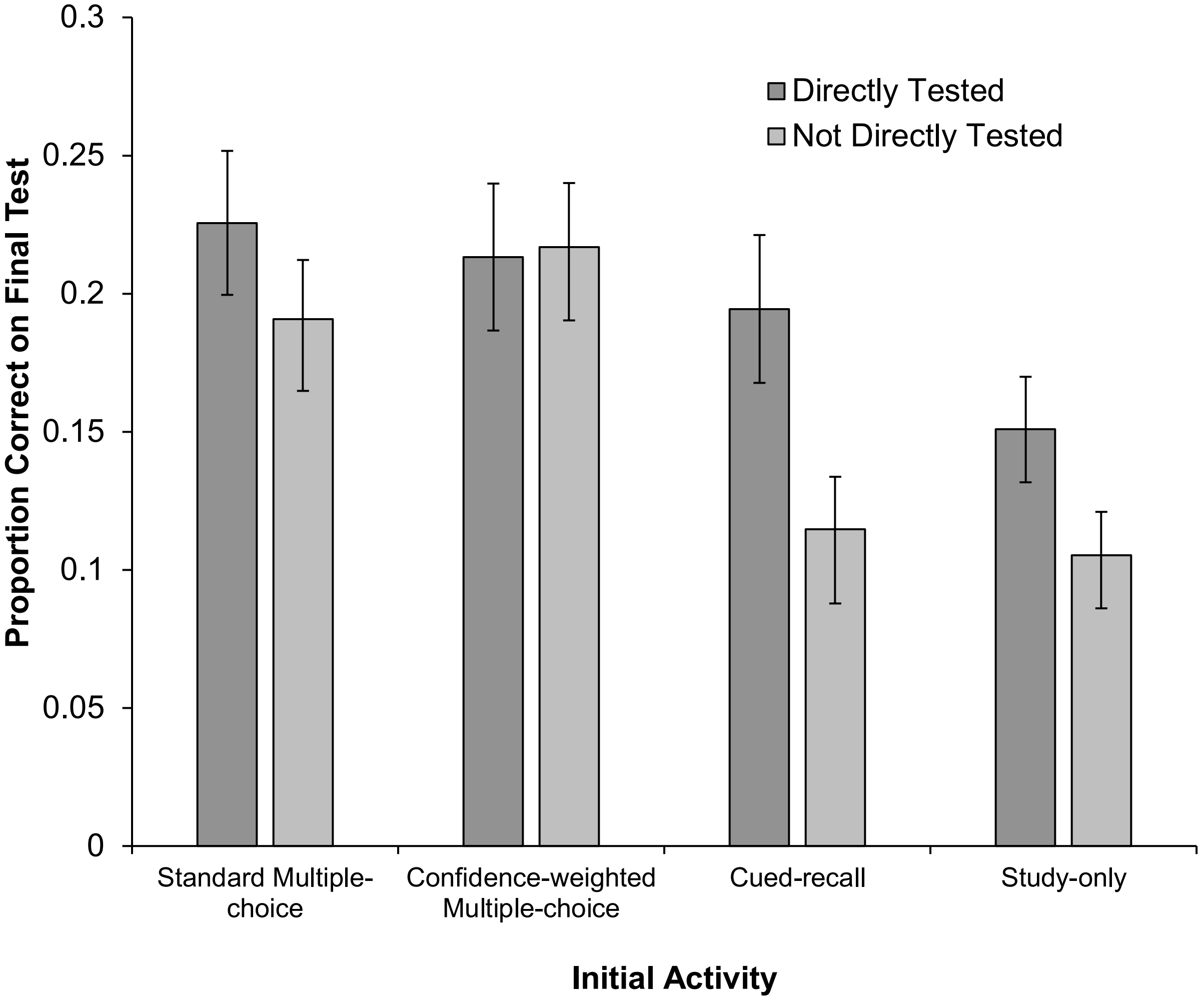

Figure 4.4. Results from Experiment 5. Proportion of words correctly recalled on the final cued-recall test as a function of the type of initial activity (standard multiple-choice, confidence-weighted multiple-choice, cued-recall, or study-only). Darker bars represent the proportion correct of the 12 words that were directly tested/studied. Lighter bars represent the proportion correct of the 24 words that were not directly tested/studied. Error bars represent ± 1 standard error of the mean.

For practical purposes, to see whether overall more words were recalled in the multiple-choice conditions, a one-way ANOVA on the total number of words recalled regardless of condition showed a significant effect of main activity [*F*(3, 149) = 5.83, *p* < .001]. Planned comparison independent samples *t*-tests showed that more words were recalled in the standard multiple-choice (*M* = 7.29, *SD* = 4.62) and confidence-weighted multiple-choice conditions (*M* = 7.79, *SD* = 4.76) than in the cued-recall condition (*M* = 5.08, *SD* = 4.40) [*t*(75) = 2.14, *p* = .035; *t*(68) = 2.48, *p* = .016]. There was no difference in the overall number of words recalled between the two multiple-choice conditions [*t*(73) = -.46, *p* = .65]. There was also no significant difference between the total number of words recall in the cued-recall and study-only conditions (*M* = 4.33, *SD* = 3.43) [*t*(76) = .85, *p* = .40]. See Table 4.1 for a comparison of means.

Table 4.1

Mean number of words recalled by initial activity condition in Experiment 5

Initial Activity	**Mean Number of Words Recalled (*SD*)**
Standard Multiple-choice	7.29 (4.62)
Confidence-weighted Multiple-choice	7.79 (4.76)
Cued-recall	5.08 (4.40)
Study-only	4.33 (3.43)

Discussion

For items that were directly tested, the typical pattern that testing is superior to studying can be seen in Experiment 5. Taking some sort of initial test (standard multiple-choice, confidence-weighted multiple-choice, or cued-recall) after studying the words was superior to restudying. Unlike in Experiment 4, however, taking a cued-recall test did not lead to better final test performance than taking a multiple-choice test for the directly tested items. Some notable differences between Experiments 4 and 5 that could explain the discrepancy include the presence of feedback, learning in the context of a sentence during the initial activity phase, a sentence-completion multiple-choice final test, selecting against a given incorrect alternative twice (rather than just once) during the initial activity phase, and a shorter delay between the initial activity and the final test.

Experiment 4 was also conducted entirely within-subjects, potentially making it easier to detect differences between the types of initial activities for directly tested words. Also, by removing the "easiest" (as determined by final test performance in Experiment 4) from each word group, the words remaining in Experiment 5 were perhaps more competitive with one another. As a result of increased competitiveness, retrieval in the presence of alternatives and then the need to discriminate between each of them during the multiple-choice testing might have been just as beneficial as cued-recall testing for the directly tested words even though cued-recall initial tests are generally thought to be better learning events (e.g., Duchastel, 1981; Foos & Fischer, 1988; Hamaker, 1986; McDaniel, Anderson, Derbish, & Morrisette, 2007).

Most notably, the multiple-choice conditions outperformed the cued-recall and study-only conditions for the not directly tested/studied items, replicating the main result of Experiment 4 under different conditions while expanding on the findings of Little et al. (2012).

Multiple-choice testing has a benefit over cued-recall testing, namely when competitive incorrect alternatives are later tested, even when the materials are not presented as a coherent text passage. Whereas in Experiment 4, overall performance did not differ between any of the initial activity conditions, in Experiment 5, overall performance was superior in the multiple-choice conditions due to enhanced recall of words that were not directly tested and the fact that those words represented two-thirds of the to-be-recalled words on the final test.

These findings support the prediction that multiple-choice tests are a more efficient way of studying, particularly when there is a large pool of items to be learned, as each test trial might lead to the retrieval of the definition for not only the correct answer, but also for the definitions of the incorrect answers. In the cued-recall condition, when participants are given practice tests on the12 to-be-learned words, they most likely attempt to retrieve just those 12 definitions. In contrast, in the two multiple-choice conditions, participants—in addition to retrieving those 12 definitions—most likely also retrieve the definitions of 24 additional words in order to reject them, with the incorrect alternatives acting as a guide for these retrieval attempts.

Even more interestingly, it appears that multiple-choice testing might benefit the incorrect alternatives just as much as the directly tested words when students are required to learn the definitions of difficult and confusable vocabulary words, as there was no significant difference between final recall performance for the directly tested and the not directly tested words in either of the two multiple-choice conditions. In contrast, a large difference in final recall performance was observed between the directly tested/studied items and the not-directly tested studied words for the cued-recall and study-only conditions (with an even greater discrepancy for the cued-recall condition as the directly tested words benefit from the testing effect). This pattern of results suggests that participants are retrieving definitions to the words presented as incorrect alternatives when the initial activity is some type of multiple-choice test. At the same time, they are not thinking about the definitions to other words that they have learned when the initial format does not have incorrect alternatives presented, further supporting the idea that multiple-choice testing promotes a broad retrieval strategy while cued-recall testing promotes a narrow one.

While participants in the confidence-weighted multiple-choice condition did not significantly outperform those in the standard multiple-choice condition, the pattern of results was numerically in the direction that was predicted. Performance on the final test was low (*M* = .19 for the standard multiple-choice format and *M* = .22 for the confidence-weighted multiple-choice format) compared to performance on the final test after learning factual information from passages. The numerical increase of 3% for the confidence-weighted multiple-choice condition over the standard multiple-choice condition in Experiment 5, however, represents a percent increase in improvement of approximately 15%. This increase is similar to the size of the percent increase in performance observed for the confidence-weighted multiple-choice condition versus that observed for the standard multiple-choice condition seen in Sparck et al. (2016). The issue of whether confidence-weighted multiple-choice tests can be significantly more effective at increasing the related benefit to vocabulary word learning should be further investigated, perhaps with the use of an easier or smaller set of vocabulary words to increase final test performance and thereby allowing a better opportunity for participants to demonstrate the anticipated added benefit.

If the confidence-weighted condition does actually enhance the learning of information that is not directly tested relative to the standard multiple-choice condition, it has even larger implications for flashcard design, as simply listing competitive alternatives in the standard format multiple-choice format might not be enough to achieve the maximum benefits for the learning of this information. Additionally, as part of obtaining a better understanding of the processes by which the confidence-weighted multiple-choice testing enhances learning more generally, investigating whether the point system implemented plays a crucial role in the degree of learning achieved is important. Gaining such a fuller understanding is critical for making valid and helpful recommendations to teachers and students about how best to use confidence-weighted multiple-choice testing.

Traditional paper flashcards, for example, typically do not incorporate scoring during learning (although computer applications can implement a scoring system quite easily, if it is required to see the benefit). Additional effort on the part of flashcard constructors may thus be required to produce the best outcomes. The results of Experiments 1 and 2, however, could shed some light on whether learners could begin by using a system with confidence-weighted multiple-choice questions and then switching over to using standard multiple-choice questions while still reaping the most learning benefits, although more research is needed.

General Discussion

The experiments in Chapter 4 apply the principles of designing effective multiple-choice practice tests to a new domain— vocabulary learning. Cued-recall testing appears to help the learning of only directly tested words, whereas multiple-choice tests can not only help the learning of directly tested words but also the learning of words that are not directly tested, but appear as incorrect alternatives on the initial multiple-choice test. These results have broad implications for the design of flashcards and how to best engage in retrieval practice when there is a large set of materials to learn, time is a limited resource, and therefore, only a limited number of test trials are available for study. Importantly, the related question benefit appears to last at a delay (48-72 hours) that is more similar to the delay between studying and testing in educational settings. Previous experiments showing the benefits of multiple-choice testing (e.g., Little et al., 2012; Sparck et al., 2016) have all been at short delays of 5 minutes.

There are several natural follow-up experiments to the present experiment that would systematically explore many of the variables included in Experiment 4 with the control of Experiment 5. First and foremost, researching the role of feedback, including how and when feedback should be given, in these dynamics is an important next step, as realistically, students rely on feedback when studying with flashcards. Research has shown that receiving feedback after taking a multiple-choice test can increase positive effects and reduce negative ones (see Marsh, Roediger, Bjork, & Bjork, 2007, for a comprehensive discussion of the potential negative consequences) for directly tested information (Butler & Roediger, 2008).

Fewer studies have focused on how feedback affects non-tested, but related information. Little et al. (2012) found no difference in the size of the related information benefit whether feedback was given after the initial multiple-choice test. Unpublished follow-up research to Sparck et al. (2016), however, found a reduced benefit for related information on a final cued-recall test when participants were given feedback after both standard and confidence-weighted multiple-choice tests relative to receiving no feedback at all. It was hypothesized that because participants might have known they were going to see the correct answer to all of the questions after selecting their answer, they might have put less effort into retrieving information about the various alternatives before selecting their answer. If inclines to adopt this less effortful strategy, related information would not then see any sort of boost in its later recall.

In the present experiment, no feedback was given after the test trials, so as to better isolate the benefits of testing, without feedback acting as a confounding variable. From a more practical educational perspective, however, it is imperative that the role of feedback on non-tested, related information be fully investigated. Nearly 70% of students report seeking feedback during self-testing as a means of assessing how well they know something (Kornell & Bjork, 2007). Although Experiments 4 and 5 cannot be directly compared to show exactly what role feedback plays given a variety of design differences, it should be noted that in Experiment 4 where feedback was given, vocabulary words in the multiple-choice condition that were not directly tested were used correctly in the context of a sentence more often than their counterparts in the cued-recall and study-only conditions.

A second follow-up experiment could include a multiple-choice final test where participants are presented a sentence with a missing GRE-level vocabulary word and have to fill in the correct, studied vocabulary word from a list of alternatives provided (as done in the actual GRE and was done in Experiment 4). Such an experiment would provide greater ecological validity in support of multiple-choice testing acting as an effective method of studying for the GRE. One concern with such a sentence-completion final test (and why the final version of the test was changed from Experiment 4 to 5) is that dynamics similar to what are expected to occur during the initial test (i.e., retrieval of definitions to competitive words) could play out during the final test if it consisted of multiple-choice questions, thus potentially affecting later tested words from that group. In other words, participants could actually be learning from their experience taking the final test, which would make it difficult to determine how much of the benefit is coming from the initial learning phase and how much of it is coming from the final testing phase.

As was done in Experiment 5, all not directly tested/studied words would need to be tested first, and one of the not directly tested/studied words from each group would have to be among the first block of 12 words tested on the final multiple-choice sentence-completion test. Performance on the not directly tested/studied words from the first block of the final test should be compared with the not directly tested/studied words tested from the second block of the final test, as it might be expected that overall performance on the second not directly tested/studied words from each group would be even greater if productive retrieval occurs during the first block of the final test. Learning on a final cued-recall test, however, is not of concern, as there is no evidence that cued-recall testing enhances the retrieval of related information without the presence of a mediator (Chan et al., 2006; Little et al., 2012).

Another related follow-up experiment should explore the role of learning the definitions in the context of a sentence, another possible *desirable difficulty*, which may lead to superior performance on the final test, particularly if it is formatted similarly to the actual GRE (which as previously discussed, requires using contextual clues to select the most appropriate fit from a list of words). Learning definitions in the context of a sentence, thus, may be more in line with the notion of transfer appropriate processing where final test performance improves when learning and encoding processes are more similar to retrieval processes (e.g., Morris, Bransford, & Franks, 1977), and a more effective way to learn difficult vocabulary words. Knowing the associated definitions of all of the potentially used words is only the first part of being able to successfully answer questions on the GRE.

College students are an ideal population to study the acquisition of GRE-level vocabulary words, as they represent the majority of GRE test-takers who are preparing to enter graduate school. Ideally, further research could capitalize on this fact by recruiting a group of motivated soon-to-be GRE test-takers and offering a multiple session learning program that combines the benefits of testing (specifically multiple-choice testing) with other desirable difficulties known to improve long-term learning including, but not limited to, expanding retrieval, interleaving, and errorful generation, first to see if these variables interact, and second, to see the optimal way to combine them. For example, Experiments 4 and 5 each began by having all participants go through a study phase where they read intact word and definition pairs. Instead of simply reading the pairs, it would be interesting to study how having participants first generate guesses about what each word means would affect retention and how this sort of production might interact with the multiple-choice testing benefit. Perhaps instead of relying on a single technique, a more effective flashcard system could use a combination of *desirable difficulties* that might lead to additional learning benefits.

A multi-session study would also be ideal for studying the effects of spacing. Maximizing the benefits of spacing is difficult to do in single session experiments (like Experiment 4) or even a two-part experiment where the second session only consists of the final test (like Experiment 5). In addition, it would offer more ecological validity and allow for even better recommendations as to how GRE test-takers should study since the retention interval between beginning to study for the GRE and actually taking the GRE for most takers is likely weeks or months with many study sessions in between.

Finally, a multi-session experiment would allow for more words to be learned, another important consideration, as well-prepared students must be familiar with more than 36 GRE-level words to receive high marks on the test. Across the various sessions, previously studied words could reappear but now as incorrect alternatives in the practice tests for newly appearing words (potentially at increasingly spaced intervals) to keep those previously studied words easily accessible while adding to a larger lexicon. Results from further investigation of how to optimize the learning of GRE type vocabulary words and their definitions in more realistic situations could have a great impact on the lives of the half million students who study for and take the GRE each year.

Although the present research has focused on the application of multiple-choice testing to flashcards for use with studying GRE-level vocabulary words, such research could be expanded to include a variety of other content. Students do not use flashcards solely to study vocabulary words, but also to learn and to test themselves on core concepts across many subjects. Extending research on the benefits of incorporating multiple-choice testing into systems of flashcards relevant to other areas of knowledge would be an exciting direction to follow and one that could eventuate in improving the self-regulated studying of learners in a variety of fields.

Chapter 5: Overview and Conclusions

When taking practice tests, learners can only realistically be tested on a subset of the information they are expected to be responsible for at the final criterion test. While retrieval practice is incredibly powerful for items that are directly tested, the story of what happens to non-tested information remains less clear. Multiple-choice initial tests, and format variations such as confidence-weighted multiple-choice tests, potentially have the power to encourage students to use broad, effective retrieval strategies that can benefit non-tested information if constructed properly. Broadly, the research presented in the present dissertation focuses on the study of instances where initial testing opportunities can also lead to the facilitation of information that is related, but not directly tested.

More specifically, the goals of the present experiments were to assess whether we can create more effective multiple-choice practice tests and pretests by using confidence-weighted multiple-choice testing and whether this type of testing can be applied to a learning situation other than studying a text passage—vocabulary learning—and to determine whether multiple-choice tests might be a more effective way to learn when faced with limited opportunities for study relative to the number of items to-be-learned. Essentially reported research sought to discover ways to maximize learning through the presentation of incorrect alternatives during practice testing.

Overview of Findings

Chapter 2 explored whether taking a confidence-weighted multiple-choice test later influences how learners take subsequent standard multiple-choice tests. Namely, does the benefit to recall of related information on standard multiple-choice tests increase following experience with a confidence-weighted multiple-choice test? Results suggest that answering questions using the confidence-weighted multiple-choice format after reading a text does impact later learning, although it appears the benefit may come from two sources—use of more productive retrieval processes during the taking of subsequent multiple-choice tests and more effective encoding of the newly studied information. Prior experience with the confidence-weighted multiple-choice format may encourage test-takers to think more deeply about the incorrect alternatives when presented with a standard multiple-choice question than they otherwise would have. The experience of answering questions using the confidence-weighted format may have also made any high confidence errors more salient or highlighted how uncertain they were in their answers after reading the text, which could drive learners to read a following text of similar difficulty more carefully.

Chapter 3 investigated whether the observed greater benefits of the confidence-weighted multiple-choice testing format over the standard multiple-choice format for recall of related information would also occur with pretesting. No evidence in support of such an increased benefit to related information over and above standard multiple-choice testing was seen. While the confidence-weighted multiple-choice format showed no additional advantages over standard multiple-choice testing, a robust pretesting effect was noted for all multiple-choice conditions relative to a baseline study-only condition, offering support to the findings of Little and Bjork (2016) that multiple-choice pretests can be effective learning tools. This finding of the present research also offers more evidence in support of the idea that while multiple-choice posttests benefit the learning of related information by encouraging test-takers to retrieve information they have learned about each of the alternatives (and confidence-weighted multiple-choice posttests seem to boost this further, encouraging learners to do this productive retrieval more often), multiple-choice pretests might operate by a different mechanism given that the to-be-learned information has not yet been studied.

Chapter 4 concentrated on determining whether the benefit to the learning of related information in text materials seen with multiple-choice post-testing as well as the increased benefit seen with confidence-weighted multiple-choice post-testing extends to the learning vocabulary words. Although confidence-weighted multiple-choice testing did not show an improvement over standard multiple-choice testing, a related information benefit was seen such that words—not directly tested in one of the multiple-choice formats—were better recalled than not directly tested words in the cued-recall or study-only conditions. A secondary goal assessed whether multiple-choice tests in general are a more efficient way to learn under conditions of limited test trials. Results suggest that when given a fixed number of test trials, learners can gain access to a greater number of vocabulary words and their associated definitions after taking multiple-choice practice tests relative to cued-recall tests, which are analogous to traditional flashcards.

Implications and Future Directions for Multiple-choice Testing in General

The benefits of multiple-choice testing for non-tested, related information have now been extended from text-based, fact learning to word-definition associative learning, and encouragingly, have been shown to last for at least several days. When constructed in the proper manner with competitive alternatives, multiple-choice tests have the additional advantage over cued-recall tests of increasing the efficiency of learners. Such benefits are particularly useful when constructing tests for instructive purposes, as is the focus of the research discussed in this dissertation. This new finding offers additional evidence that multiple-choice testing encourages a broad retrieval strategy, making a wider range of information retrievable in a single retrieval episode (or flashcard). Knowing the instructional benefits of multiple-choice testing and understanding when their use is appropriate over other testing formats, is particularly important for students and educators as awareness that the act of testing is a powerful pedagogical tool, and not only a summative assessment, grows.

Students wishing to optimize their own self-regulated learning could incorporate multiple-choice testing into their flashcards with dramatic results. Rather than simply writing a vocabulary word (or some other topic) on one side of the card and the definition (or associated information) on the other, flashcard creators could incorporate multiple-choice questions with competitive alternatives on one side and the correct answer on the other (perhaps even including explanations of why the incorrect alternatives are incorrect or circumstances under which they might be correct). Such a design could improve students' productivity and provide a better "bang for their buck" while studying. As students become busier with coursework and extracurricular activities, the importance of studying "smarter," not "harder," is welcome advice. Although results on the overall effectiveness of the technique have been mixed (with few well-controlled studies), the actual construction of the multiple-choice questions may have learning benefits over and above answering questions written by someone else, another variable that should be further explored in conjunction with the new proposed flashcard system (Bottomley & Denny, 2011; Palmer & Devitt, 2006; Sircar & Tandon, 1999).

The Importance of the Relationship Between Initial and Final Test Items

Educators wishing to harness the benefits of multiple-choice testing can provide practice tests and other resources that are well constructed such that answers are competitive. In order to directly see the learning benefits for related information after taking initial multiple-choice tests, it is also important to make sure that final tests are constructed in a way such that the previously incorrect answers are now correct answers. Otherwise learning may occur, but it may not be evident without the appropriate final test questions to measure it.

One such case where a lack of connection between incorrect alternatives on the initial test and correct answers on the final test might be obscuring learning from multiple-choice testing is McConnell, St-Onge, and Young (2015). The researchers constructed pedagogical tests on the advice of the Medical Council of Canada (2010) as part of the training to prepare medical students for their licensure exam. Medical students were given scenario-based multiple-choice questions involving hypothetical patients and their medical cases (also known as context-rich multiple-choice questions) on some topics and cued-recall questions (i.e., the same stems as the context-rich multiple-choice questions with no alternatives provided) on other topics, with topic and its associated question type counterbalanced across students.

A final mock licensure exam of all context-rich multiple-choice questions was given at the end of the training. Some questions on the exam were verbatim-repeat context-rich multiple-choice questions. Some questions tested the same learning objective as the originally studied context-rich multiple-choice questions but provided new vignettes and alternatives. And finally, some questions tested learning objectives (some describing scenarios verbatim from the learning session and others describing scenarios that were related) initially tested using cued-recall questions.

There was a robust testing effect on the final test for context-rich multiple-choice questions that were repeated verbatim as well as cued-recall questions that were repeated (but now presented in a multiple-choice format with alternatives present). Overall, performance on verbatim-repeated questions was significantly higher than performance on new, but related questions that tested the same learning objective, suggesting some specificity of learning that did not transfer to similar medical cases. Most importantly, there was no benefit from testing for the related questions for topics initially studied in the context-rich multiple-choice condition. This finding suggests that the multiple-choice testing benefits first reported in Little et al. (2012) did not transfer to related questions under these circumstances.

While inconsistent with Little et al. (2012), construction of the multiple-choice alternatives may explain the difference and show why it is so important to construct final tests that will measure learning from multiple-choice initial tests. If the incorrect alternatives in McConnell et al. (2015) included appropriate diagnoses, tests to run, next steps for the medical professional to take, and so forth, in a slightly different scenario (e.g., with a patient of a different age or gender, or a patient with slightly different presenting symptoms), the final test question should be based on that scenario and one of the previously incorrect alternatives should be the correct answer. While the scenarios might have touched the same broad topic, it is not clear whether or not the alternatives for some or all of the initial questions actually had competitive alternatives that were directly related to a future question. Prior research has shown no multiple-choice testing benefit for new questions (e.g., Nungester & Duchastel, 1982). Given the uncertainty in how the training questions and how the final test questions were connected to one another, it thus becomes unclear whether a benefit would be expected in McConnell et al. given the need to construct the initial multiple-choice tests and the final tests in a specific way to see a result on the final test (Little & Bjork, 2015).

Similarly, Pan et al. (2015) found improvements in retention only for specific pieces of studied, multiterm history and biology facts that were directly tested via initial multiple-choice tests, but they found no benefit to other pieces of those same facts. For example, the fact *Winston Churchill was Prime Minister of the United Kingdom during World War II*, is comprised of four key pieces of information (*Winston Churchill*, *Prime Minister*, *United Kingdom*, and *World War II*). After a study phase where each of the to-be-learned facts is read, participants might be asked to select the answer *Winston Churchill* from one of four alternatives to the fill-in-the-blank question stem, ______________ *was Prime Minister of the United Kingdom during World War II* as part of the initial test. In such a case, when asked to fill in the answer (with no alternatives presented) to the question *Winston Churchill was Prime Minister of the* ______________ *during World War II* (correct answer: *United Kingdom*) on the final test, no benefit was found to this potentially "related" information.

However, based on the guidelines set forth by Little and Bjork (2015) that alternatives must be competitive to trigger productive retrieval to systematically reject them in an active and thoughtful manner, it would not be expected that asking about which country Winston Churchill led would demonstrate the type of related information learning from the mechanism focused on in the present research. In other words, when selecting *Winston Churchill* as the correct answer for ____________ *was Prime Minister of the United Kingdom during World War II* on the initial test, there is no need to consider or retrieve information about the United Kingdom and thus no memory benefit would be expected. Participants might, however, be expected to answer questions about other prime ministers of the United Kingdom during the early 20th century or other world leaders during World War II (if those were the alternatives presented) at a greater rate on the final test than participants who did not experience the question in a multiple-choice format. Crucially, the final test would have to be set up to ask about those other leaders in order to measure whether or not such retrieval occurred during the initial test.

Related information for the text passages in the research presented here and in Little et al. (2012) and Sparck et al. (2016) is defined on the basis of being about the same specific topic (e.g., geysers in Yellowstone National Park). Related information for the vocabulary words in the present studies is defined on the basis of starting with the same letter and being of the same part of speech. The confusability of these items and the need to discriminate between similar items underlie these defined relationships. There are, however, other ways to operationalize relatedness that deserve further research.

Hamaker (1986) specified five additional ways in which final questions may be related to initial questions that might offer guidance. First, all questions may come from a restricted category and the final questions are new questions from the same category (e.g. Rothkopf & Bisbicos, 1967, where participants performed better on new proper names when they had previously answered adjunct questions on proper names). Second, questions might be close in proximity (not overlapping in content) and thus when trying to recall the answer to the initial question, facilitation may spread to information presented nearby (e.g. Chan, McDermott, & Roediger, 2006; Chan, 2009; Frase, 1968; Rothkopf & Billington, 1974). Third, they might be related such that they are different examples of a more general principle from the reading (e.g. Watts & Anderson, 1971). Fourth, the initial questions may test specific factual information and the later questions address a new application of information that can be answered by retrieving that specific factual information (e.g. Andre, Mueller, Womack, Smid, & Tuttle, 1980, Experiment 4). Finally, the final question may be related such that it is a paraphrased version of the initial question (e.g. Andre et al., Experiment 6). Changing how relatedness is operationalized is an important step in understanding how the present boost to retention of related information generalizes and can be used more broadly.

Multiple-choice Testing and Transfer

Discussion of the relationship between initial and final questions required to observe the benefit of learning from multiple-choice testing leads to another important consideration, particularly when trying to scale benefits up to educationally relevant settings— namely, how does multiple-choice initial testing affect transfer? Answering related questions and transferring or applying learned information to new situations are both desired in educational settings.

Within the field of psychology, the term transfer may be broadly defined as the application of acquired knowledge to novel contexts. As described by Barnett and Ceci's (2002) taxonomy for transfer, what defines a novel context also widely varies. A novel context may refer to a change within or across the knowledge domain, the physical context, the temporal context, the functional context, the social context, or the modality. Transfer across and within the knowledge domain, is most directly related to classroom learning expectations. A review by Carpenter (2012) as well as a recent meta-analysis by Pan and Rickard (2018), suggest that the benefits of testing generally do extend to new types of problems, but there is still relatively little research that focuses more specifically on harnessing the benefits of multiple-choice tests to new types of problems.

Transfer may further be broken down into either near or far transfer. In the context of testing, near transfer might describe the relationship between a vocabulary term and its definition. For example, on an initial test, an individual might be presented with a definition and asked to produce the vocabulary word (as was done in Experiment 5). Then on the final test, the individual might be presented with the vocabulary word and ask to produce the definition. Technically, this question would be considered to involve transfer because the information retrieved on the final test is not the exact same information recalled on the initial test; however, since the cue and target essentially reverse each other in these questions, applying knowledge from the question should occur quite frequently and relatively easily, and is thus near transfer. The effect of this simple manipulation could be explored. Perhaps a multiple-choice flashcard that dynamically incorporates both retrieval of the definition and the word could yield even better memory for the to-be-learned vocabulary words. Previous work on learning foreign language paired associates suggests that contextual interference during encoding is a *desirable difficulty* that leads to durable retention and flexible use of the foreign vocabulary words (Schneider, Healy, & Bourne, 2002; Soderstrom, Sparck, & Bjork, 2016).

More interesting, however, is the effect on truly novel questions. The final test questions in all of the work previously described (e.g. Chan et al., 2006; Chan, 2009; Little, et al., 2012; Sparck et al., 2016) and in Chapters 2 and 4 of the present dissertation tested factual information that was studied prior to initial testing; however, the information was just not directly tested during the initial testing phase of the experiments. A final test of near transfer would also involve retrieving studied information in a slightly different context from the initial test. A final test of far transfer, however, might ask test-takers to solve some problem that may appear to be very different and make inferences that go beyond what was initially learned after being tested on a basic concept (e.g., Johnson & Mayer, 2009; Watts & Anderson, 1971). Retrieving the initially tested information, though not the correct answer on the final test, and drawing conclusions from that information can help solve the novel problem on the final test. Far transfer requires more of the learner, particularly that the learner notices the underlying connections between the learned information, the initial question, and the final question. Understanding if and how much initial multiple-choice testing affects this type of learning would be incredibly valuable for a variety of educational applications.

Another future direction important to improving the utility of multiple-choice testing is to make the benefit of non-tested related information more generalizable to a variety of educational applications. This direction involves understanding how more conceptual materials or ones that require the understanding of a process where current understanding depends on prior learning (within the same learning session) fare after multiple-choice testing. For example, to-be-learned lessons could describe the formation of lightning or the life cycle of a star (e.g., Johnson & Mayer, 2009; Yue et al., 2015). To fully understand the later stages of one of these processes, a learner needs to first understand the preceding stages. Understanding how multiple-choice testing might impact the understanding of these interrelated, dependent steps is an interesting and important step to further scale up this area of research.

Defining Competitiveness

Even if we develop a better understanding of these issues, it remains important to note that competitiveness may vary among individuals with different levels of background knowledge. Research shows that students' background knowledge affects their comprehension of prose materials (e.g., McNamara, Kintsch, Butler, Songer, & Kintsch, 1996) and expertise affects the ability to classify and solve problems (e.g., Chi, Feltovich, & Glaser, 1981). These findings give reason to suspect that prior knowledge and expertise may also play important roles in whether or not people retrieve information appropriately during initial multiple-choice tests to see the benefit for related information.

For example, take the basic geography question *What is the capital of British Columbia, Canada?* used as a sample question in Sparck et al. (2016) and seen in Figure 1.2. The correct answer is *Victoria*. *Berlin* is obviously the wrong choice, noticeable to anybody with the most basic knowledge of geography. For demonstrative purposes, let us replace that alternative with

Toronto. For someone who knows anything about Canadian geography, the answer choice *Toronto* would not be competitive since it is located in Ontario (in the eastern portion of Canada). For someone with very little knowledge about Canadian geography, the mere fact that Toronto is a city in Canada might make it competitive enough to be considered the correct answer. The alternative *Toronto* thus might engage someone with low levels of prior knowledge in productive retrieval processes, but it likely would not for someone with high levels of prior knowledge (e.g., a Canadian citizen).

The answer choice *Vancouver*, on the other hand, is likely a competitive alternative even for those who are relatively familiar with Canadian geography, as it is a major city also located in British Columbia and begins with the same letter, both of which could potentially lead to interference between the two cities. Using *Vancouver* as an alternative should then lead to productive retrieval of information about both Victoria and Vancouver, as the test-taker tries to discriminate between the two cities starting with the letter "V." When later asked *What is the largest city in British Columbia?,* the answer (*Vancouver*) should be more accessible according to the mechanisms described in this dissertation.

Defining what qualifies as a competitive alternative then becomes tricky as teachers try to develop questions at the appropriate level of difficulty for their students' knowledge base. Questions written to act as pedagogical tools for first-year medical students may look very different than those written for third-year medical students, residents, or even current physicians. Instructors may even have students with a wide variety of backgrounds enrolled in a single course, and thus appropriate materials may be even more dependent on individuals themselves than classes of individuals.

When considering the results of any study looking at the related question benefit, it is therefore important to note that incorrect alternatives may drastically differ in competitiveness for different populations and may be an important variable influencing experimental outcomes. The results from Experiment 3's self-assessment of the to-be-learned topics prior to being tested on or reading the passages suggest that prior knowledge for Saturn and Yellowstone National Park in our sample of psychology undergraduates is rather low, and that it is not likely an important factor in interpreting the findings of Little et al. (2012), Sparck et al. (2016) or the text-learning studies outlined in Chapters 2 and 3. In an actual classroom, however, the effect of prior knowledge becomes increasingly more likely, as many classes are explicitly intended to build upon (what is thought or supposed to be, but is not always) prerequisite knowledge.

The Number of Alternatives Presented

Another important consideration to more comprehensively understand the benefit to non-tested, but related information is the number of competitive alternatives that are presented with each question stem. In theory, a three-item multiple-choice question could lead to the retrieval of three separate pieces of information (two of them related), while a four-item multiple-choice question could lead to the retrieval of four separate pieces of information (three of them related), and so on. In light of the present findings, instructors may naturally think it appropriate to include a large number of alternatives, expecting to see large learning benefits after providing students with relatively few questions. There are, however, many reasons why such a strategy would be misguided.

As established by Little and Bjork (2015), alternatives must be competitive for any related question benefit to be seen. Psychometric analysis of test question banks suggests that many of the presented alternatives are rarely selected, suggesting that these alternatives are not truly competitive (Dickinson, 2013). Several explanations may be offered as to why adding more options do not make for a more competitive test.

First, the questions may have been poorly constructed. Adding additional alternatives just for the sake of adding alternatives will not yield positive results. Second, the questions might not be written at the appropriate difficulty for the people taking the test. The previous discussion of the role that prior knowledge plays suggests that may play a role. Third, and most interesting, depending on the topic, there might not be enough alternatives that are truly competitive. Competitiveness for the GRE-level vocabulary words in the present research was defined as being of the same part of speech and starting with the same letter. There are arguably a large number of potentially competitive alternatives, so perhaps increasing the number of alternatives might be an effective strategy to encourage more productive retrieval under such circumstances. On the other hand, when asking questions about planets or geysers in Yellowstone National Park, there are only a limited number of planets in our solar system and relatively few notable geysers found in Yellowstone.

Even if, in a perfect world, the benefit for related information was directly proportional to the number of alternatives, from a practical sense, adding a large number of alternatives and expecting learning benefits might still be faulty. Specifically, fatigue might be a concern. In the present research, as well as prior studies that have shown a benefit for multiple-choice testing (Little et al. 2012; Little & Bjork, 2015; Sparck et al., 2016), only three options were presented during the initial multiple-choice test. If learners feel overwhelmed during the initial test by having to retrieve many pieces of information, they may not persist with an effective strategy that will lead to the related question benefit. Additionally, research on the multiple-choice format and directly tested information suggests that increasing the number of multiple-choice lures on initial tests can have negative consequences in the form of smaller testing effects (Roediger & Marsh, 2005).

Although well-controlled research on the topic is sparse, ultimately applied research on final criterion tests focusing on assessment suggests that changing the number of options on an exam, does not reliably change multiple-choice test scores (e.g., Dickinson, 2013; Schneid, Armour, Park, Yudkowsky & Bordage, 2014); and thus, we might also not expect an increase in the benefit to related information by adding more and more alternatives to practice tests. An advantage of using fewer alternatives is that students are able to answer more questions per hour, thus improving content coverage and validity of an exam. A meta-analysis by Rodriguez (2005) suggests that offering three options is the optimal number of alternatives to assess learning. Other research offers similar conclusions, noting that writing more alternatives takes time and leads to diminishing gains in reliability (Baghaei & Amrahi, 2011). These findings imply that test constructors should spend more time constructing fewer, but better and truly competitive alternatives.

Directly translating these results to maximize the benefit of related information after engaging in retrieval practice is not immediately straightforward. Additional research investigating the relationship between the number of alternatives presented on practice tests and the benefit to related information on final tests is thus needed to fill in the gaps. It should also be noted that the confidence-weighted multiple-choice format, by design, is bound to only have three alternatives.

Researching ways to optimize the use of multiple-choice tests as a study tool in general thus remains a fruitful field of research. Before making broad generalizations to students and educators about how to construct effective multiple-choice tests, more research on these topics must be completed. To better understand the benefits of multiple-choice testing, it is important to understand how competitiveness varies, the type of problems multiple-choice testing benefits, in addition to understanding the role of feedback, final test format, and a variety of other educationally relevant variables.

Implications and Future Directions for Confidence-weighted Multiple-choice Testing

Sparck et al. (2016) first established that confidence-weighted multiple-choice testing can enhance later performance on questions about related information to a greater extent than standard multiple-choice testing, naturally helping learners spontaneously engage in productive retrieval about incorrect alternatives. One concern about the use of confidence-weighted multiple-choice testing is that the format could be hard for instructors to implement in their classrooms given its required structure. The present research however shows that confidence-weighted multiple-choice testing can potentiate future learning by both encouraging more productive retrieval about each of the alternatives on a subsequent standard multiple-choice test (i.e., the adoption of a better test-taking strategy) and a more effective reading and encoding of the subsequent text.

These results suggest that instructors may be able to train students using confidence-weighted multiple-choice tests to the benefit of standard multiple-choice tests. While promising, further research must be done to test how long-lasting the benefits are (e.g., if students receive confidence-weighted multiple-choice training on Monday, will the benefits on future learning still be seen on Friday?), how much experience is needed (e.g., is just answering a few confidence-weighted multiple-choice questions enough to elicit future changes?), and how transferrable the benefits are (e.g., if students receive confidence-weighted multiple-choice training in their biology class, will the learning benefits transfer to their history class?).

Understanding each of these questions individually as well as the potential interactions between these variables is important for implementing confidence-weighted multiple-choice testing in classrooms. For example, answering a few confidence-weighted multiple-choice questions may be enough to elicit strategy changes in the test-taker for a short time, but more extensive training may be needed to see longer lasting, widespread future benefits. By more fully understanding the roles of these variables, we can improve the recommendations made to educators hoping to equip their students with useful learning tools and to students hoping to optimize their self-regulated studying.

The added benefits of confidence-weighted multiple-choice testing do not appear to extend to pretests. Although the confidence-weighted multiple-choice pretests did not improve subsequent learning better than standard multiple-choice pretests, a pretesting effect for all multiple-choice conditions (relative to baseline performance) was seen, suggesting that confidence-weighted multiple-choice can be just as effective as standard multiple-choice tests. This finding also supports the hypothesis put forward in Sparck et al. (2016) that confidence-weighted multiple-choice testing benefits learners by being more likely to retrieve information about each of the alternatives that they have recently learned about. In a true pretest, the learner will know little to no relevant information.

Although there was not a significant difference between the standard and confidence-weighted formats found for studying vocabulary words in Experiment 5, there was a numerical difference in the direction expected. Overall performance on the final cued-recall test was low with high variability for all of the learning conditions, suggesting that the task was difficult for the learners. Low performance and high variability might suggest that more participants would be needed to detect a true effect.

To better understand how task difficulty affects the presence and size of the confidence-weighted multiple-choice benefit, research should establish the difficulty of learning different sets of materials and then compare confidence-weighted multiple-choice testing to standard multiple-choice testing after learning materials of variable difficulty to see if the difference between the two testing formats consistently appears. If the added benefit appears, does it vary or remain proportionally constant across materials of different difficulties? While the difference did not reach significance in Experiment 5, the percent increase from standard multiple-choice to confidence-weighted multiple-choice was similar in size to that seen in Sparck et al. (2016). It might be expected that when the materials are more difficult, there might be a greater focus on discriminating among items on the initial test, in which case we would expect a larger benefit for related information. On the other hand, when the materials are more difficult, the test-taker might be expected to have a more difficult time in general recalling the correct answer on a final cued-recall test. Competing dynamics could be at play in this instance. Another possibility exists such that confidence-weighted multiple-choice testing, which is thought to be about making relational judgments, may be better for probing about material learned from a coherent text where the relationships are more strongly emphasized. Being tested on difficult vocabulary might also words might encourage beneficial processing more naturally.

Research on confidence-weighted multiple-choice is currently quite limited. All of the questions that remain for multiple-choice testing more generally also apply to confidence-weighted multiple-choice testing. Providing answers to some of these questions is an important step in helping the research community to make recommendations to educators and students who want to study “smarter.”

Concluding Remarks

Research involving the improvement of testing as a pedagogical tool, has practical implications for the field of education as teachers increasingly turn to research in cognitive psychology to maximize the learning of their students, and students seek out new study methods to optimize their own self-regulated learning. While historically maligned by the research community, when properly designed, prior research (e.g., Bjork, et al., 2014; Little et al., 2012; Sparck et al., 2016) as well the present studies show that multiple-choice tests and their varieties can have benefits similar to and beyond those of other test formats. Understanding the nuances of how initial testing format interacts with other variables, particularly when the final test contains new and related information is an important area of study within the study of test-enhanced learning more broadly.

Testing is one of the most effective ways to encourage lasting learning (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Moving forward, research should continue to focus on improving its utility to maximize benefits for learning. Such findings could have a large impact on the educational community, all the way from elementary school to college and beyond.

Appendix A

Saturn

Saturn is the sixth planet from the Sun and the second largest planet in the Solar System. The planet is most well known for its beautiful system of planetary rings, which consist largely of water ice particles with smaller amounts of rocky debris and dust. Along with Jupiter, Uranus, and Neptune, Saturn is classified as a gas giant (also known as a Jovian planet, after the planet Jupiter).

The existence of Saturn has been known since prehistoric times: Saturn is the most distant planet that can be seen with the naked eye. Saturn gets its name from the Roman god Saturnus: the god of agriculture and harvest. The Romans considered Saturnus to be the equivalent of the Greek god, Kronos. Ancient Chinese cultures designated the planet Saturn as the 'earth star,' based upon the five elements which were traditionally used to classify natural elements. In Hindu astrology, Saturn is known as 'Sani' or "Shani'—the judge among all the planets.

Saturn's rings were first observed by Galileo in 1610. With his telescope, Galileo was able to see Saturn's rings, but was not able to observe them as such, instead declaring

Saturn to be composed of three bodies that almost touch each other. It was in 1655 that Christian Huygens was the first to suggest that a ring surrounded Saturn. In 1675, Giovanni Domenico Cassini determined that Saturn's ring was actually composed of multiple smaller rings with gaps between them; the largest of these gaps was later named the Cassini Division. In 1959, James Clerk Maxwell hypothesized that the rings could not be solid or they would become unstable and break apart. Maxwell proposed that the rings must be composed of numerous small particles, all independently orbiting Saturn. Maxwell's theory was proven correct in 1895 through spectroscopic studies of the rings carried out by James Keeler.

Saturn orbits the sun at an average distance of 1.4 billion kilometers. One Saturnian year (the amount of time that it takes Saturn to rotate around the Sun) occurs about every 30 Earth-years. One Saturnian day (the amount of time that it takes the planet to rotate on its axis) takes approximately ten Earth-hours. This can be contrasted with Mercury, which has a relatively fast revolution around the sun and a relative slow rotation on its axis. A day on Mercury (the time it takes the planet to rotate on its axis) takes 176 Earth-days, but a Mercurian year (the time it takes the planet to orbit the sun) takes 88 Earth-days—on Mercury, a day is longer than a year.

Saturn is composed of hydrogen, with small amounts of helium and other trace elements. Although there is little direct information about Saturn's internal structure, it is thought that the interior of the planet consists of a small core of rock and ice, surrounded by a thick layer of metallic hydrogen and a gaseous outer layer. Because of its massive gaseous layer, Saturn has a low density compared to other planets: Saturn is the only planet in the Solar System that has a density less than that of water. Saturn has a planetary magnetic field that is stronger than that of Earth, but not as strong as that of Jupiter. Not every planet has a magnetic field. Venus, for example, is a special case of a rocky planet with no magnetic field.

Sixty known moons orbit Saturn. Only seven of Saturn's known moons are massive enough to have collapsed into hydrostatic equilibrium under their own gravitation. Titan is Saturn's largest moon and the Solar System's second largest moon (only Jupiter's moon, Ganymede is larger). Titan was discovered in 1655 by Huygens. Titan is the only moon in the Solar System to possess a significant atmosphere. Mimas and Enceladus were discovered by William Herschel in 1789. Saturn's celestial body atmosphere exhibits a banded pattern similar to Jupiter's, but Saturn's bands are much fainter and are also much wider near the equator.

Saturn’s rings, unlike the rings of other planets, are very bright. Evidence suggest that the rings of Saturn possess their own atmosphere, which is independent of the planet itself; this atmosphere is largely made of the oxygen gas (02) that is produced when ultraviolet light from the Sun interacts with ice in the rings. Two prominent rings (A and B) and one faint ring (C) can be seen from Earth. There are several gaps within the rings; two such gaps are opened by known moons embedded within them. The gap between the A and the B rings is known as the Cassini Division. The much fainter gap in the outer part of the A ring is known as the Encke Gap. The Maxwell Gap lies within the outer part of the C ring. The area between the A ring and the F ring is known as the Roche Division. While the largest gaps in the rings (such as the Cassini Division and the Encke Gap) can be seen from Earth, it has been discovered that the rings actually have an intricate structure of thousands of thin gaps and ringlets. Additionally, Saturn's second largest moon Rhea may have a fragile ring system of its own. Until 1980, the structure of the rings of Saturn was explained exclusively as the action of gravitational forces. The *Voyager* spacecraft found radial features in the B ring, called spokes, which could not be explained in this manner, as their persistence and rotation around the rings were not consistent with orbital mechanics. Spokes appear to be a seasonal phenomenon, disappearing in the Saturnian midwinter/midsummer and reappearing as Saturn comes closer to equinox.

Saturn was first visited in September 1979; *Pioneer 11* flew within 20,000 km of the planet's cloud tops. In November 1980, the *Voyager 1* probe visited the Saturn system; the probe revealed previously unseen surface features of various moons. *Voyager 1* data indicated that wind speeds on Saturn could reach 1,800 km/hour—significantly faster than those on Jupiter. *Voyager 2* probed slightly later, in 1981. In 2004, the *Cassini-Huygens* spacecraft conducted a flyby of Titan, capturing radar images of its large lakes and their coastlines as well as numerous islands and mountains. In March of 2006, NASA reported the existence of geysers: liquid water reservoirs that erupt on Saturn's moon Enceladus; pockets of liquid water may exist no more than tens of meters below the surface of the moon. In 2006, a *Cassini-Huygens* probe saw the first proof of hydrocarbon lakes near Titan's north pole: the largest of these is almost the size of the Caspian Sea. The *Cassini-Huygens* probe's primary mission ended in 2008 when the spacecraft had been expected to have completed 74 orbits around the planet Saturn.

Saturn Questions: Set A (Correct Answer denoted with *)

What planet lacks an internal magnetic field?

- Venus *
- Mercury
- Jupiter

Saturn's rings were first observed in what year? (at the time, however, they were not known to be rings)

- 1610 *
- 1675
- 1789

In 1655, who became the first scientist to suggest that Saturn is surrounded by a ring?

- Maxwell
- Huygens *
- Keeler

The atmosphere of Saturn's rings is primarily composed of what element?

- Oxygen *
- Hydrogen
- Helium

What is Saturn's second largest moon (a moon that is believed to have its own fragile ring system)?

- Titan
- Rhea *
- Mimas

Saturn was first visited in September of 1979 by which space probe?

- Voyager 1
- Pioneer 11 *
- Voyager 2

Only seven of Saturn's known ________ are massive enough to have collapsed into hydrostatic equilibrium.

- moons *
- rings
- spokes

How long does it take Saturn to rotate once on its axis?

- 10 Earth hours *
- 88 Earth days
- 30 Earth years

The area between the A ring and the F ring is known as the ________.

- Cassini Division
- Encke Gap
- Roche Division *

Saturn gets its name from Saturnus, the ________ god of agriculture and harvest.

- Roman *
- Chinese
- Hindu

Saturn Questions: Set B (Correct Answer denoted with *)

On what planet is a day longer than a year?
- Venus
- Mercury *
- Jupiter

In what year did William Herschel discover Mimas and Enceladus, two moons of Saturn?
- 1610
- 1675
- 1789 *

Who first proposed that Saturn's rings aren't solid, but must instead be composed of many small particles?
- Maxwell *
- Huygens
- Keeler

The body of Saturn is primarily composed of what element?
- Oxygen
- Hydrogen *
- Helium

What is Saturn's largest moon?
- Titan *
- Rhea
- Mimas

Which space probe collected data demonstrating wind speeds on Saturn exceeding 1,800 km/hour?
- Voyager 1 *
- Pioneer 11
- Voyager 2

________ appear to be a seasonal phenomenon, disappearing in the Saturnian midwinter and midsummer.
- moons
- rings
- spokes *

How long does it take Saturn to revolve around the sun?
- 10 Earth hours
- 88 Earth days
- 30 Earth years *

The area between the A ring and the B ring is known as the ________.

- Cassini Division *
- Encke Gap
- Roche Division

What ancient culture designated Saturn as the 'earth star'?

- Roman
- Chinese *
- Hindu

From Little, Bjork, Bjork, and Angello (2012)

Yellowstone

Established in 1872, Yellowstone became America's first national park. The park is located at the headwaters of the Yellowstone River, for which it takes its name. In the eighteenth century, French trappers named the river "Roche Jaune" which is probably a translation of the Minnetaree name for "Rock Yellow River." Approximately 96% of the land area of Yellowstone National Park is located in the state of Wyoming, but the park extends into neighboring states of Idaho and Montana. Yellowstone is widely known for its wildlife and geothermal features: the park, itself, contains half of the world's geothermal features.

Evidence suggests that Aboriginal peoples have lived in the Yellowstone region for at least 11,000 years. The region is home to several Native American tribes including the Nez Perce, Crow, and Shoshone. European explorers first entered the region in the early nineteenth century. In 1806, John Colter left the Lewis and Clark Expedition to explore the region with a group of fur trappers. Upon seeing Yellowstone, he described it as a place of "fire and brimstone" due to the boiling mud, steaming rivers, and petrified trees. Colter continued to explore the region for another four years, finally leaving the wilderness when two of his partners were killed by Blackfeet Indians. Over the next forty years, numerous reports from mountain men and trappers told of geothermal features of Yellowstone, yet most of these reports were believed at the time to be myth. After an 1856 exploration, mountain man Jim Bridger reported observing boiling springs, sprouting water, and a mountain of glass and yellow rock. These reports were largely ignored as Bridger was known for being a "spinner of yarns." His stories did arouse the interest of explorer and geologist Ferdinand Vandeveer Hayden, who, in 1859, started a two-year survey of the upper Missouri River region. Bridger and United States Army surveyor W. F. Raynolds acted as guides. After exploring the Black Hills region in what is now the state of South Dakota, the party neared the Yellowstone River, but heavy snows forced them to turn back.

In 1872, President Ulysses S. Grant signed a bill into law that created Yellowstone National Park. Nathaniel Langford was appointed as the park's first Superintendent, serving for five years although denied a salary, funding, and staff. Langford lacked the means to improve the land or to properly protect the park, and without formal policy or regulations, he had few legal methods to enforce such protection. This left Yellowstone vulnerable to poachers, vandals, and others seeking to raid its resources. In 1916, pioneering industrialist and conservationist, Stephen Tyng Mather, along with journalist and writer Robert Sterling Yard, spearheaded a publicity campaign to promote the creation of a federal agency to oversee National Parks. Mater eventually became the first director of the National Park Service under the United States Department of the Interior. Under the dynamic leadership of Mather, several national parks, including the Grand Canyon and Zion National Park, were established. In 1917, administration of Yellowstone was transferred to the National Park Service.

Yellowstone is the centerpiece of the 20 million acre Greater Yellowstone Ecosystem—a region that includes Grand Teton National Park and adjacent National Forests. The Greater Yellowstone Ecosystem is the largest remaining continuous stretch of mostly undeveloped land in the United States (outside of Alaska) and is considered to be the world's largest intact ecosystem in the northern temperate zone. Yellowstone is home for a variety of animals including elk, moose, and bison: threatened species include the endangered gray wolf, the threatened lynx, and grizzly bears. An estimated 600 grizzly bears live in the Greater Yellowstone Ecosystem, with more than half of the population living within Yellowstone. The grizzly is currently listed as a threatened species. Population figures for elk are in excess of 30,000— making them the largest population of any large mammal species in Yellowstone. Numbering fewer than 50 in 1902, but more than 4,000 as of 2003, bison as also well represented in the park.

Yellowstone National Park spans an area of approximately 3,500 square miles. The park rests at an average altitude of 8,000 feet above sea level. The highest point in the park is atop Eagle Peak, which is almost 11, 400 feet above sea level. The most prominent summit on the Yellowstone Plateau (although not the highest) is Mount Washburn at 10, 243 feet. Nearby mountain ranges include the Gallatin Range to the northwest, the Beartooth Mountains in the north, the Absaroka Range to the east, and the Teton Range and the Madison Range to the southwest and west.

Forests comprise 80% of the park's land area—the remaining land area is primarily comprised of grasslands. 1.700 species of trees and plants are native to the park. Of the eight conifer tree species documented, Lodgepole Pine forests cover 80% of the total forested areas. Other conifers, such as the Douglas Fir and Whitebark Pine, are found in scattered groves throughout the park. As of 2007, the Whitebark Pine is threatened by a fungus known as white pine blister rust; however, this is mostly confined to forests well to the north and west. Aspen and willow are the most common species of deciduous trees in the park.

Yellowstone National Park sits atop the Yellowstone Caldera, which is considered an active volcano. The most recent volcanic activity occurred about 70,000 years ago. Geothermic activity continues to occur in geysers and other thermal features in the park. There are 300 geysers in Yellowstone and a total of at least 10,000 geothermal features. A geyser is a type of hot spring that erupts periodically: geyser activity is caused by surface water gradually seeping down through the ground until it meets rock heated by magma. The most famous geyser in the park is Old Faithful. The tallest active geyser in the park, as well as in the world, is Steamboat Geyser. Daisy Geyser usually erupts every 90-110 minutes and is very predictable. Castle Geyser is thought t o be the oldest geyser in the world. The size of Castle Geyser's cone, in a shape that reminds people of a castle, indicates that it may be somewhere between 5,000 and 40,000 years old. Interestingly, although prominent in the park, geysers are actually the least common geothermal feature there. Besides geysers, Yellowstone has other geothermal features including hot springs, mud pots, and fumaroles. A fumarole is an opening the earth's crust that emits steam and gases including carbon dioxide. Hot springs are the most common hydrothermal features in the park. A prominent hot spring known for its beauty is Morning Glory Pool.

Yellowstone is a popular tourist destination. Since the mid-1960s, at least 2 million tourists have visited the park almost every year.

Yellowstone Questions: Set A (Correct Answer denoted with *)

What explorer left the Lewis and Clark Expedition to explore the region with a group of fur trappers?

- Colter *
- Bridger
- Hayden

About 600 of what threatened species live within the Greater Yellowstone Ecosystem?

- elk
- bison
- grizzly bears *

Attacks by what tribe caused Colter to leave the Yellowstone region?

- Minnetaree
- Shoeshone
- Blackfeet *

What is the tallest geyser in Yellowstone National Park?

- Old Faithful
- Steamboat Geyser *
- Castle Geyser

In what year did Colter first explore Yellowstone with a group of fur trappers?

- 1806 *
- 1856
- 1872

The majority of Yellowstone National Park resides in what state?

- South Dakota
- Wyoming *
- Montana

What is the highest peak in Yellowstone National Park?

answers:

- Eagle Peak *
- Mt. Washburn
- Beartooth

Who was Yellowstone National Park's first Superintendent, who served for five years devoid of salary, funding, and staffing?

- Mather
- Grant
- Langford *

Which species of tree covers 80% of the total forested areas in Yellowstone National Park?

- Douglas Fir
- Lodgepole Pine *
- Whitebark Pine

What type of geothermal feature is an opening in the earth's crust that emits steam and gasses?

- Geyser
- Hot Spring
- Fumarole *

Yellowstone Questions: Set B (Correct Answer denoted with *)

What mountain man reported observing boiling springs, sprouting water, and a mountain of yellow rock, but was largely ignored due to a reputation of being a 'spinner of yarns?"

- Colter
- Bridger *
- Hayden

What species makes up the largest population of a large mammal species in Yellowstone National Park?

- elk *
- bison
- grizzly bears

French trappers named Yellowstone River "Roche Jaune," probably a translation of what Native American tribe's name for Yellow Rock River?

- Minnetaree *
- Shoeshone
- Blackfeet

What geyser is thought to be the oldest in the world?

- Old Faithful
- Steamboat Geyser
- Castle Geyser *

In what year did Yellowstone become a National Park?

- 1806
- 1856
- 1872 *

The Black Hills region is found primarily in what state?

- South Dakota
- Wyoming *
- Montana

At 10,243, what is the most prominent peak (but not the highest) in Yellowstone National Park?

- Eagle Peak
- Mt. Washburn *
- Beartooth

Who signed a bill into law making Yellowstone the first national park?

- Mather
- Grant *
- Langford

As of 2007, which tree, found only in scattered groves, is threatened by a specific fungus?

- Douglas Fir
- Lodgepole Pine
- Whitebark Pine *

With only about 300 examples in the park, what is the least common type of hydrothermal feature in Yellowstone?

- Geyser *
- Hot Spring
- Fumarole

From Little, Bjork, Bjork, and Angello (2012)

Appendix B

* indicates a word removed from Experiment 5

** indicates a word added to Experiment 5

GRE WORD	**DEFINTION**	**SENTENCES** (Exp. 4 only)
A-word group		
Abnegate	refuse, reject	If you wish to be an abstinent monk, you must have the will and ability to _____________ worldly possessions. By talking about the many harms of drugs, the school counselor hoped she could encourage her students to _____________ drug use.
Aver	assert or affirm strongly	When talking to her parents, the teenager tried to _____________ her right to her privacy by asking her parents not to monitor her computer. In his passionate and eloquent speech, the attorney hoped to _____________ his client's innocence to the judge.
Allay	put fears to rest, calm, mitigate	Hopefully, the company's soaring stock price will _____________ the concerns of nervous stockholders. By praying day and night, the people hoped to _____________ the anger of the Gods.
Abet *	to encourage or aid, usually in wrongdoing	During the press conference, the president vowed severe consequences for any person or group who chose to _____________ the terrorists. Even though he didn't directly _____________ the murderers, he still offered them protection and was therefore charged.

C-word group

Compunction	feeling of guilt, regret	Helen divorced her husband because he never seemed to feel any _____________ about lying to her. When the Petersons failed to make their mortgage payments, the bank manager showed no _____________ and quickly foreclosed on the couple's home.
Calumny	untrue statement made to damage someone's reputation	Although Charles does not personally like Henry, he is not the type of person to spread a _____________ about his enemy. When Jeremy felt he was about to get in trouble, he would often distract his parents with a _____________ about his older brother.
Conciliation	act of placating, reconciling, winning over	He was successful at his attempt at _____________ by buying his girlfriend her favorite flowers. Having acted as a mediator for many conflicts over the years, she was now an expert at the _____________ of angry partisans.
Complacence *	satisfaction with existing situation, often unaware of potential danger or defects	The citizens' _____________ with the current government, despite its obvious lack of proficiency and adroitness, was baffling. After seeing the untroubled _____________ of the rich kids in the private school, the new teacher decided to educate them more about the realities of the outside world.

E-word group

Expiate	do something as a way of showing you are sorry	To _____________ for breaking his neighbor's window, John shoveled snow for three months. Because Bill lacks an income source, he can only _____________ for his crime by collecting trash on the side of the roads.
Expurgate	change a written text by removing offensive parts	The producer agreed to _____________ some of the R-rated scenes so that the movie could be shown on network television. Even though Wikipedia can be a helpful resource, the interactive ability allows anyone to add inaccurate information or _____________ accurate information just because they're offended by it.
Excoriate	criticize harshly	There are many who _____________ him and his government for their policies on refugees, and for their commitment to the war in Iraq. Because Ann is an atheist, she will probably _____________ the decision to allow prayer in schools.
Expatiate *	speak or write at length or in detail	The lesson usually overruns by thirty minutes because the professor loves to _____________ on his favorite subjects. Instead of being concise and succinct, the speaker chose to _____________ on the topic endlessly, going on many tangents.

I-word group

Inchoate	only partly in existence, not perfectly formed	Having just come into existence a few years ago, the new political party is considered _____________ by many historians. If only you could come up with a complete plan and not just an _____________ idea.
Inveterate	always happening, habitual, ingrained	An _____________ reader, she always carried a book with her wherever she went. Having become _____________ in the everyday language of teenagers, the new word could now be heard everywhere in schools, parks, and social media outlets.
Irascible	easily angered	It wouldn't take too much effort to aggravate an _____________ person. Being an _____________ boss, he would blow up in outrage at the sight of even the slightest loss in revenue.
Incumbent *	necessary as a duty and responsibility	Both his lack of skill and his temper made him incapable of maintaining the attitude supposed to be _____________ on a president; and his tongue was never a carefully governed one. It is _____________ on all citizens to be aware of the political situation and to vote.

O-word group

Obsequious	servile, obedient, compliant	The princess was quite happy having _____________ servants who constantly showered her with attention and service. The teacher did not like having anyone question his ways or teachings; he wanted only _____________, simple-minded students in his class.
Obstreperous	noisy and defiant	After winning the battle, the _____________ troops had to be calmed down by their commanding officer. The _____________ puppy whimpered all night and kept everyone awake.
Officious	intrusive and offering services even though they are unwanted, meddlesome	While the _____________ sales clerk may have believed he was giving me some helpful advice, he was just wasting my time by telling me things I already knew. Although Cathy doesn't claim to be an expert on anything, she is still happy to provide _____________ advice on every topic under the sun.
Obdurate*	stubbornly persistent, unwilling to change ways	Even though his fellow teammates urged him to accept the new coach, the star basketball player remained _____________ and refused to follow the coach's directions. The protestors were _____________, not moving from the town square even when the police showed up with tear gas.

P-word group

Pernicious	highly damaging	Because the chemicals you are using in the lab are _____________, you should be very careful during your experiment. Because of its high winds, the hurricane was quite _____________ to the small town.
Punctilious	precise about doing things in an accurate way	While I enjoy cleanliness, I am not so _____________ that I get upset about a little dust on my furniture. Surgeons must be very _____________ during ther operations because the mistakes they make could result in fatalities.
Perspicacious	having keen insight and understanding	Although the detective was a _____________ woman, she was not able to identify the killer's motive. The _____________ teacher had no problem figuring out which students had cheated on the exam.
Precarious *	unstable	He had no tenure and held a _____________ position in the university, likely to be fired any day. After the economic crisis, our financial situation became quite _____________ and we had to start saving considerably.

S-word Group

Staid	serious, boring, old-fashioned	Since my aunt is a nun, she always wears such a _____________ look on her face. Mike's _____________ apartment is bare of anything exciting and completely lacking in color.
Sedulous	diligent, devoted	Although she told him that she was already in a relationship with someone, Ethan refused to abandon his _____________ efforts to get a date with Felicia. As a jockey, his _____________ activities helping to train and care for the horses have led to more blue ribbons than anyone else.
Scurrilous	obscene, vulgar	In an attempt to ruin the mayor's reputation, the newspaper editor wrote several _____________ articles on the politician's spending habits. Because Elliott was angry with his ex-girlfriend, he began to spread _____________ rumors about her that were not true.
Spurious *	not genuine or authentic	After receiving a low appraisal on my diamond ring, I realized the suspicious-looking jeweler had sold me a _____________ jewel. Recently, some weight loss drugs were taken off the market because of _____________ statements made by the manufacturers.

T-word group

Tenebrous	hard to understand, obscure	He wished to delve into the _____________ depths of her mind, wanting to understand the mysterious source of her fears and anxiety. The fullness of the moon did little to alleviate the _____________ nature of the forest darkened by night.
Taciturn	reserved in speech, tending to be silent	In cultures where modesty and restraint are highly valued, it's no wonder that people turn out to be more _____________ compared to the talkative members of our Western culture. He was a _____________ leader who didn't say much, yet he was effective anyway; people admired him simply for his confident stance and calm demeanor.
Trifling	trivial, insignificant	The director of the firm had too many important tasks on hand to be bothered with _____________ matters like the broken sink in the bathroom. Having done a Master's and worked in the business for several years, the applicant was not going to settle for such a _____________ salary in the new job.
Trite *	not effective anymore due to being repeated too often	Having seen the same kind of scenario over and over again, the film critic found the movie's plot to be _____________ and banal. The so-called ladies' man was full of clichés and often used _____________ phrases to try to flirt with women.

V-word group

Visceral	obtained through intuition rather than through reasoning	The advertising creates a _____________ sensation of fear that is hard to explain logically. As a detective with an almost perfect record for solving difficult cases, she relies on her _____________ sense, in other words her gut feelings, to lead her in the right direction.
Venal	corruptible, bribable	The _____________ police officer accepted the money the drug dealers gave him to look away from their illegal transactions. Because the mayor was a _____________ man, he had no problem welcoming bribes from real estate developers.
Veracious	honest, unwilling to tell lies	Describing historical events accurately, he is credited with being a _____________ historian. She has a reputation for being _____________, and so people usually take her word.
Volatile *	liable to change rapidly and unpredictably	Support for the politician is extremely _____________ at this time, with opinion polls showing different results every week. The stock market is presently quite _____________, scaring off the more timid investors.

M-word group

Maculate**	spotted; blotched
Maladroit**	clumsy, not skillful
Maudlin**	effusively sentimental

Q-word group

Querulous**	complaining
Quiescent**	temporarily inactive
Quixotic**	idealistic but impractical

R-word group

Recumbent**	reclining
Redolent**	odorous, fragrant
Refulgent**	brightly shining; gleaming

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063-1087.

Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 608-629.

Andre, T. Mueller, C., Womack, S., Smid, K., & Tuttle, M. (1980). Adjunct application questions facilitate later application, or do they? *Journal of Educational Psychology, 72*, 533-543.

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

Baghaei, P, & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple-choice items. *Psychological Test and Assessment Modeling, 53*(2), 192-211.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? *Psychological Bulletin, 128*, 612–637.

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition, 3*, 165-170.

Bjork, E. L., Soderstrom, N. C., & Little, J. L. (2015). Can multiple-choice testing induce desirable difficulties? Evidence from the laboratory and the classroom. *American Journal of Psychology, 128*, 229-239.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Bottomley S., & Denny, P. (2011). A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. *Biochemistry and Molecular Biology Education, 9*(5), 352–361. doi: 10.1002/bmb.20526.

Bruno, J. E. (1989). Using MCW-APM test scoring to evaluate economics curricula. *The Journal of Economic Education*, *20,* 5-22.

Bruno, J. E. (1993). Using Testing to Provide Feedback to Support Instruction: A Reexamination of the Role of Assessment in Educational Organizations. In D. A. Leclercq and J. E. Bruno (Eds.), Item Banking: Interactive Testing and Self-Assessment (190-209). Springer Berlin Heidelberg.

Butler, A. C., & Roediger, H. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279-283.

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.

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19*, 580-606.

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

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

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. C. Carterette & M. P. Friedman (Series Eds.) & E. L. Bjork & R. A. Bjork (Vol. Eds.), *Handbook of perception and cognition: Vol. 10. Memory* (pp. 317-344). San Diego: Academic Press.

Dickinson, J. R. (2013). How Many Options do Multiple-Choice Questions Really Have? *Developments in Business Simulation and Experiential Learning, 40*, 171-175.

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology, 6,* 217-226.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4-58.

Educational Testing Service. (2014). A snapshot of the individuals who took the GRE revised general test.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. Journal of *Educational Psychology, 80,* 179-183.

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology, 59*(4), 244-249.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505-13. doi:10.3758/s13421-011-0174-0

Hamaker, C. (1986). The effects of adjunct question on prose learning. *Review of Educational Research, 56*, 212-242.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin Review, 19*, 126-134.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*, 290-304.

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, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101,* 621-629.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558.

Karpicke, J. D., Butler, A. C., & Roediger, H. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory 7*, 471-479.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187–194. http://dx.doi.org/ 10.1037/0278-7393.31.2.187.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 989-998. doi:10.1037/a0015729

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory. New York: Academic Press, 1970*, pp. 333-373.

Landauer, T. K., & Bjork, R. A. (1978). Optimal rehearsal patterns and name learning. In M. M. Gruenberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory. London: Academic Press, 1978*, pp. 625-632.

LaPorte, R., & Voss, J. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67,* 259-266.

Little, J. L. (2011). Optimizing multiple-choice tests as learning events. (Doctoral dissertation). Retrieved from ProQuest Information & Learning. (AAI3493389)

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*(1)*,* 14-26. doi: 10.3758/s13421-014-0452-8

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition, 44*(7)*,* 1085-1101.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344.

Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory, 19*(4), 346-359.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology*, *77*, 463-473.

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

McConnell, M. M., St-Onge, C., & Young, M. E. (2015). The benefits of testing for learning on later performance. *Advances in Health Science Education, 20*(2), 305-320.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

McNamara, D. S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction, 14*(1), 1– 43.

Medical Council of Canada. (2010). Guidelines for the development of multiple-choice questions. Ottawa: Medical Council of Canada.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*(1), 18-22.

Palmer, E., & Devitt, P. (2006). Constructing multiple choice questions as a method for learning. *Annals, Academy of Medicine, Singapore, 35*(9), 604–608.

Packman, J. L., & Battig, W. F. (1978). Effects of different kinds of semantic processing on memory for words. *Memory & Cognition*, *6*, 502–508.

Pan, S. C., Gopal, A., & Rickard, T. C. (2015). Testing with feedback produces potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 107*(4).

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710-756. doi:10.1037/bul0000151

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*(2), 644-667.

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*(5), 1139-1153.

Rodriguez, M. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Education Measurement: Issues and Practice, 24*(2), 3-13.

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

Roedgier, 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*(5). 1155-1159.

Rothkopf, E. Z., & Billington, M. J. (1974). Indirect review and priming through questions. *Journal of Educational Psychology*, *66*, 669–679.

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology, 58*, 56-61.

Rowland, C. A. (2015). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi: 10.1037/a0037559.

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R. & Bordage, G. (2014) Reducing the number of options on multiple-choice questions: Response time, psychometrics, and standard setting. *Medical Education, 48*, 1020-1027.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*, 419-440.

Sircar, S. S., & Tandon, O. P. (1999). Involving students in question writing: a unique feedback with fringe benefits. *American Journal of Physiology, 277*(6), S84–S91.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99–115.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*, 176-199.

Soderstrom, N. C., Sparck, E. M., & Bjork, E. L. (2016). Variable practice enhances learning of foreign language vocabulary. Poster presented at the 57th annual meeting of the Psychonomic Society, Boston, MA, USA.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principle and Implications, 1.* doi: 10.1186/s41235-016-0003-x

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval enhances long-term retention. *Memory & Cognition, 38*(2)*,* 244-253.

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62*, 387–394.

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 127-134.

Yue, C. L., Soderstrom, N. C., & Bjork, E. L. (2015). Partial testing can potentiate learning of tested and untested material from multimedia lessons. *Journal of Educational Psychology, 107*(4)*,* 991-1005.