Spring 2017

The Self-Study Testing Effect with Bidirectional Paired Associate Learning

Bernadette Barton
Bernadette.Barton@Colorado.EDU

The Self-Study Testing Effect with Bidirectional Paired Associate Learning

Bernadette Barton

The Department of Psychology and Neuroscience, University of Colorado at Boulder

Defense Date: April 7, 2017

Thesis Advisor: Matt Jones, The Department of Psychology and Neuroscience

Defense committee: Matt Jones,The Department of Psychology and Neuroscience. Alice Healy, The Department of Psychology and Neuroscience. Richard Olson, The Department of Psychology and Neuroscience. Ben Kirshner, School of Education.

Abstract

The *testing effect* describes a method of retrieval practice that benefits long term retention of learned material, over restudying and re-exposure (Roediger & Karpicke, 2006a). The benefits of testing have been implemented outside of psychology laboratories into classrooms around the US, but are still not being taken advantage of by students outside of lecture (Karpicke, Butler & Roediger, 2009). The current experiment seeks to acknowledge the *testing effect* as present in a widely used self-study website and further define the most effective retrieval practice students can implement within it. Fifty-eight introductory Psychology students at the University of Colorado, Boulder studied 30 Anthropology terms and their definitions in a paired associate learning format on Quizlet.com. The paired associates were split into different self-study modes on Quizlet.com: one restudy block, and two cued recall with feedback blocks that applied either a consistent direction of study to initial presentation, or inconsistent direction. Two days after the self-study session, participants took a final retention test to measure the retention of the learned paired associates. The test analyzed the bidirectional retention of the 30-paired associates with recognition questions and cued-recall questions that manipulated direction; majority of questions included the paired associates verbatim to the study session, some of the questions were not verbatim to initial study but rather applied the paired associate relationships to an example to measure for comprehensive, conceptual understanding. The results of this study were inconclusive due to a confound in study block and question items, but paved way for future implications for studies on real self-study methods of students today, and the importance of retrieval practice direction on the direction of final retention.

Introduction

In educational settings, tests are administered by teachers to evaluate and measure their students retention of learned information. In psychological research, decades have been dedicated to studying the benefits of self testing as a retrieval practice to enhance final retention; this has been coined as: *the testing effect* (Roediger & Karpicke, 2006a). The extensively studied phenomenon depicts an advantage of self testing over restudying for long term retention and heightened probability of recall, confirming self testing benefits go beyond advantages of re-exposure (Carrier & Pashler, 1992; Roediger & Karpicke, 2006a; Storm, Friedman, Murayama, & Bjork, 2014). The testing effect has been studied across various disciplines of Psychology and Neuroscience, all of which have concluded it occurs because of an active retrieval practice that modifies and enhances the memory of the retrieved information (Young, 2015). By acting as a consolidation event, the tested information is more likely to be recalled in the future than if the information was not used in retrieval practice (Storm, Friedman, Murayama, & Bjork, 2014).

Although the testing effect is widely agreed upon and has yielded consistent significant effects, those who should benefit most from the implications are not taking advantage of it. A 2009 survey of 177 students at Washington University in St. Louis found that more than half of the students (55%) rated "rereading notes or textbooks" as the most important way to study for an exam, while only 1% rated self-testing as their most important study strategy (Karpicke, Butler & Roediger, 2009). A recent study considered that students are not implementing retrieval practice into their typical exam study due to lack of awareness of benefits. This study aimed to educate college students about the testing effect in hopes that this knowledge would self-direct them to self-test in the future (Young, 2015). The 2015 study split a research methods and statistics course into two groups: one group saw a PowerPoint presentation about the importance

of self-testing and the other did not. Despite the lecture on benefits of self-testing, Young found no significant difference between groups for study methods or final grades. In two follow-up experiments, Young (2015) found that students who benefited from experimental test conditions did not implement the learned testing strategies on later study conditions. Clearly, the testing effect cannot be elicited by direction from teachers or subconscious practice, but it can be implemented in the classroom itself. Teachers are putting self-testing practice into action throughout the country with the use of iClickers. iClickers are an interactive technology used in over 700 college campuses, which allow for teachers to devote class time to testing material (Anderson, Healy, Kole & Bourne, 2011). Although these recent studies encourage teachers to devote class time to retrieval practice (Anderson et. al, 2011; Young, 2015), there are no studies that implement current technology available to students that allows them to practice retrieval outside of lecture.

The testing effect has been studied extensively and proven within laboratory and classroom settings, but has yet to take advantage of the technology available to students outside the classroom. Even though students do not report using retrieval practice to study (Karpicke, Butler & Roediger, 2009), it is likely many students are already taking advantage of the benefits of self-testing without being aware of it. Today, there are multiple online platforms that allow students to input information from their courses and study in various formats on their personal computers. A meta-analysis of 159 testing effect articles found that the most pronounced testing effect occurs from a cued-recall retrieval practice, where retrieval of one item is enhanced by a cue to that item, that provides positive and negative feedback of the recalled item (Rowland, 2014). The current study uses modes that mirror this conclusion that already exist on a common

online study site to observe pronounced testing effects in an environment more familiar to students and their current self-study habits.

Paired associate formats in testing effect literature assign a stimulus to a specific response, and then cue recall of one with its corresponding pair. Of the 159 testing effect articles, over half of the studies with significant results employed a paired associate learning format for their studies (Rowland, 2014). The current study uses a common online study website: Quizlet. The website allows its 2 million monthly users to create digital "sets" (paired associates) of to be learned information, where one term corresponds to one definition. The website provides different study modes, one in a restudy format ("note card"), and another a cued recall test with positive and negative feedback ("learn") to study the "sets" of paired associates. The paired associate testing effect studies in Rowland's (2015) meta-analysis involve a unidirectional retrieval practice and do not manipulate the direction of retrieval practice or retention but rather only the type of study mode, test or restudy. The current study has the unique opportunity to determine if the testing effect exists in online self-study aids, and also to determine the degree to which direction of retrieval practice can improve the bidirectional retention and transfer of learned paired associate relationships.

A short term memory study of bidirectional transfer of paired associates found that unidirectional forward retrieval practice produces a reverse retention of the pair only 70-90% as strong as the studied forward associate pair (Leuba, 1966). If there is not a 100% bidirectional transfer of associations in short term memory, it can be assumed that this hindered transfer will only decline with delay, alongside overall retention decline. Quizlet's paired associate "sets" allow for bidirectional test study blocks. After the initial presentation of stimulus (term) to response (definition), the current experiment used Quizlet's "learn" mode to manipulate two

different study blocks: one in a consistent direction with the original presentation, and another inconsistent with the original direction. The retention test contained questions consistent and inconsistent with direction of study per block (see Appendix) to examine if the direction of retrieval practice influences the direction of retention and if the bidirectional transfer of one study direction is superior.

In addition to forward and reverse direction questions, the retention test contained comprehension questions that were not verbatim to the paired associate study blocks and did not manipulate direction of test. Instead, these questions applied the term with an example to test for conceptual understanding (see Appendix). By the active retrieval practice of the testing study conditions, terms were likely affected by the *generation effect*, which says that generating stimuli increases the retention of that information (Rowland, 2014). The terms in the testing study blocks were expected to benefit from active generation of the term or the definition. A 2004 study of retrieval practice also manipulated subjects self-generation; in this study, subjects saw both complete and incomplete highlighted words in prose format and retained more of the incomplete words in a fill-in-the-blank retention test. This study showed a benefit of generation on retention because the active completion of incomplete highlighted words increased their probability of recall (deWinstanley and Bjork, 2004).

As opposed to iClicker retrieval practice (usually recognition, multiple choice questions), the current online retrieval practice did not only provide a self-test method that benefits from an optimal retrieval practice, but from the advantage of self-generation of a response, prior to the positive and negative feedback of the verbatim response, whether it is term or definition.

The Present Study

The current study aimed to confirm the presence of the testing effect within a paired associate format in a commonly used online self-study aid to give students the ability to benefit from retrieval practice without interference or guidance from teachers. The results of this study aimed to show that students use retrieval practice to study for exams by asking about prior Quizlet usage in a survey, to show that retrieval practice does not need to be implemented by teachers or take up class time. The study considers the reach of the testing effect, and sought to find an equal transfer of bidirectional associations because of the manipulation of direction in an optimal self-study practice.

Main Hypothesis: testing effect present in test study blocks

There was one restudy block (notecard) and two test study blocks (learn). The test study blocks represent optimal retrieval practices because they are presented in a repetitive, cued-recall test format, with positive and negative feedback. It was predicted that study blocks would have a main effect on percent error. Specifically, material learned in a test block would be retained and understood better than material learned in the restudy format. The main effect of study block would prove the testing effect to exist in the employed online study aid mode, where a teacher is not present or influencing type of study. The study predicted that optimal retrieval practice of test blocks would correspond to a better retention across question type (recognition, cued recall, comprehension) over restudy blocks because of the practice of active retrieval that enhanced memory for those certain pairs (Young, 2015). Rowland's (2014) meta-analysis found that final recall tests produce larger testing effects than final recognition tests, regardless of initial test-study format. In accordance with Rowland's (2014) meta-analysis conclusion, the current retention test sought to find larger testing effects for cued-recall questions than recognition

questions. Additionally, test blocks were expected show a greater testing effect for comprehension questions because of the generation of the recalled pair prior to feedback. The generation of responses in the test blocks was expected to aid increased retention and the ability to recall the pair conceptually, beyond verbatim recognition or recall.

Secondary Hypothesis: direction of retrieval practice relevant to direction of retention

For students to implement the self-study retrieval practice, it is beneficial for them to know which direction of practice is best for retention. The present study predicted that the direction of retrieval practice affects the direction of retention. The study manipulated the direction of test study blocks, one in a forward direction of study (term to definition) and one in a reverse direction of study (definition to term), and the retention test question direction (forward or reverse multiple choice, forward or reverse fill-in-the-blank) to examine the effect direction has on retrieval practices and the ultimate retention of paired associates. It was assumed that items studied in a forward direction would be recalled best on forward retention questions, and items studied in reverse would do be recalled best on reverse direction questions, because that was the direction in which they were studied. It predicted that forward terms would perform equally on reverse direction questions, and comprehension questions because of the generation effect from recalling a definition before getting feedback. It was predicted that terms studied in a reverse direction would also perform equally on forward direction recognition questions, and comprehension questions because of an advantage of seeing the terms in the forward direction during initial presentation in conjunction with studying them in reverse, which would transfer at a recognition level because of optimal practice. This hypothesis understood that forward cued-recall questions would be the most difficult despite random assignment because they required recall of a multi-word, conceptual definition rather than a one-word term. It was hypothesized

that forward studied terms would have an advantage with these questions due to conceptual generation and the mirror of study practice to this specific question type.

Method

Subjects

Fifty-nine Introductory Psychology students at the University of Colorado, Boulder participated in this experiment in exchange for participation credit in the course. Subjects signed up to participate through the University of Colorado's SONA website. Subjects knew only of the duration of the study and the amount of credits attainable prior to their appointment time. Subjects signed up for both part one and two through the SONA system before initial participation. If a subject did not receive participation for part one of the experiment, part-two was canceled by the researcher. There were six possible conditions that a subject could be assigned to. These conditions resulted from a counterbalanced design of the order of presentation of the three study modes: note card, forward learn, and reverse learn. All subjects were first exposed to the entire set of 30 terms and then proceed onto counterbalanced trials of modes of studying. All subjects were exposed to the associate pairs an equal amount of times. One subject was excluded from data analysis due to failure to return for the retention test.

Materials

This study used 30 one-word terms and their corresponding definitions. The study material was taken from a lower division anthropology class at the University of Colorado Boulder (ANTH 2200: The Archaeology of Human History) textbook, *The Past In Perspective* (Feder, 2011). All terms were one word. Definitions ranged from three to seven words.

Two of Quizlet's study modes were used for all three study blocks in the first part of the experiment. One mode was the "notecard" mode, which was used for the initial presentation and

the restudy block. Quizlet's "notecard" mode reflects presentation and restudy blocks in a paired associate format from previous studies as it presents users with the term and then reveals the corresponding definition. It does not require any active retrieval from the subject. In the presentation block, all subjects were exposed to all 30-paired associates in a forward direction, no matter their assigned condition order. Following the initial presentation of all 30 pairs, the paired associates were split into three equal study blocks: one "notecard" (restudy) block and two "learn" (test) blocks. Regardless of successful recall, repeated retrieval practice produces positive effects on retention (Karpicke & Roediger, 2009) to optimize this result, each study block was completed twice.

The learn mode is used as an immediate cued recall test condition (with feedback) in both forward directional and reverse directional paired associate format. Quizlet's learn mode reflects immediate cued-recall test conditions by presenting the subject with either a term or a definition from the initial presentation and cues the recall of the corresponding pair. For forward direction testing, the site presents users with a term and cues recall of it's corresponding definition. For reverse direction testing, the site presents users with a definition and prompts the recall of the corresponding term. After typing in the corresponding information, Quizlet's "learn" mode presents a feedback page with either a correct screen or incorrect screen. In the incorrect screen, Quizlet's "learn" mode compares the entered response, in red, to the verbatim response from the "set," in green, on the screen for the subject to view. In the correct screen, Quizlet "learn" mode compares the entered response to the verbatim response on the screen for the subject to view, both in green. All modes are dependent upon the subject to interact with the screen, there is no fixed time for the notecard mode or test mode input but rather the subject's own option to

manually flip the note card or input their response. All study trials ended by showing a completion page that indicated all terms in this page had been studied.

Considering Rowland's (2014) conclusion that increased testing effects correspond with delayed intervals of retention, the retention test occurred two days after part one. Of the 159 testing effect articles, 58% of final tests were conducted in either cued recall or recognition format (Rowland, 2014) none of which considered direction of study to direction of retention. The present retention test included both recognition questions (multiple choice) and cued-recall questions (fill-in-the-blank). These two verbatim question types manipulated direction, either consistent with study or inconsistent with initial study (see Appendix).

The retention test also considered the possible reach of the generation effect in test blocks by including recognition and cued recall questions in a comprehension format. All of the questions besides the comprehension questions were verbatim in content from the study blocks of part one. The comprehension questions applied the term conceptually without the use of any previously studied words associated with the paired associate. Comprehension questions did not manipulate direction, and therefore were balanced between the two test types: Test A had three comprehension questions in fill-in-the-blank format, and three in multiple choice; Test B reversed the question types (see Appendix). Comprehension questions were the only questions balanced for test type across subjects. At the end of the test, there was a seven-question survey that covered prior exposure to material and past Quizlet usage (see Appendix). The tests were scored blind of condition and question type. The 30-question test was split into five question types: multiple-choice forward, multiple-choice reverse, fill-in-the-blank forward, fill-in-the-blank reverse, and comprehension (Table 1).

Random.org was used to assign paired associates to each study block. The assignment to study block was the same for all subjects. The presentation of information per block was randomized by Quizlet. The order of presentation of the study blocks was counterbalanced. Random.org was used to determine how each term would be tested. The assignment to test type was the same for all subjects.

Procedure

Subjects signed up for two, 30-minute timeslots, with up to four subjects per timeslot. When a subject arrived for their appointment, he or she was asked by the research assistant to carefully read a consent form and sign if he or she felt comfortable with the requirements of the experiment. The consent form gave a brief description of the experiment and informed participants that there was no risk associated with participation. The description informed the subject that the experiment sought to study different self-study learning methods effects on long-term retention. The consent form informed participants that researchers would record their initials next to their subject number until he or she returned for part two. Subjects were made aware that no personal information would be recorded in the data. It did not mention the conditions of the experiment. Upon giving consent, researchers recorded the subject's initials next to their subject number and condition. The experimenter gave subjects brief instructions about the procedure of the study. Subjects were told in part one they would be studying 30 terms online that would be tested for retention in two days during part two. Subjects were told to follow the cues on the screen to study each term, and when the screen indicated one round had completed, to continue on to the next tab. Subjects were told the initial webpage tab was an introduction session to all the terms that would be tested, in which he or she was asked to read the term and its corresponding definition and move on. Subjects were told after the introductory

session, he or she would be studying subsets of these terms in different study formats in the following six web tabs. Subjects were asked to complete all seven tabs in consecutive order. Before starting the experiment, subjects told to keep cell phones and all other distractions away for the duration of the experiment. Then subjects were led into a private room with a computer to complete the experiment. Next to the computer, there was a visual example of a round completion page with a reminder that this page indicated he or she should move onto the next tab in the browser.

After completing all seven tabs, the initial introduction tab and six study module tabs, the subject exited the experiment room to receive credit for participation in part one. Subjects received one credit for their participation and a reminder sheet with the day and time they signed up to complete part two of the experiment. After the subject left the laboratory, the researcher entered the experiment room and marked the run sheet for any incomplete rounds or web errors on the screen.

When a subject arrived for part two, the researcher found their initials on the experiment run sheet to determine which test to administer (A or B) and what subject number to record on the retention test. Subjects were told they would be taking a 30 question retention test over the material studied in part one, and were asked to also complete a small seven question survey following the completion of the test. Subjects were asked keep cell phones and other distractions away for the duration of the test. Then, subjects were led into a quiet room with a pen to take the test. After completing the retention test and survey, subjects gave their test to the experimenter and were given their second credit for participation. Subjects were given a feedback form explaining the hypothesis of the study and its goals as well as the purpose to the six different study block orders.

Results

This study contained two different types of analysis: a subject analysis and an item analysis. The subject analysis was performed to interpret the results randomized across subjects. The item analysis was performed to interpret the results randomized across items. Both analyses were conducted in order to generalize the results of this study to larger populations and other items, because the paired associate items selected for this experiment were confounded to only one study block and tested in with only one question type, and for multiple choice and fill-in-the-blank questions, in only one question direction. The terms in this study were confounded to their study block and test question, and in order to look at the results objectively, both analyses must be included. These analyses consisted of a 3 x 5 subject analysis and a 3 x 5 item analysis that considered the main effect of study blocks, the main effect of test type (including both verbatim questions and comprehension questions), and the interaction between study blocks and test types. They also included a 3 x 2 x 2 subject analysis and a 3 x 2 x 2 item analysis, these analysis considered only the 24 questions that manipulated direction or retention, and excluded the six comprehension questions. These three-way analyses considered the main effect of study block, the main effect of test type (multiple choice or fill-in-the-blank), the main effect of test direction (forward or reverse), and the interactions between: study blocks and test types, study blocks and test directions, test types and test directions, and study blocks, test types and test directions. A separate subject analysis was conducted on only comprehension questions in order to isolate questions that did not manipulate direction and remove all verbatim type questions from these results. Results from the seven-question survey were scored by frequency.

The repeated measures factorial 3 x 5 subject analysis ANOVA found a significant main effect of study block on percent error, $F(2, 57) = 3.12$, $p < .05$. A Fisher's PLSD paired

comparison post hoc test of study block revealed a significant mean critical difference between the restudy and reverse blocks; restudy vs. reverse, $p = .02$ (Table 2). The subject analysis yielded a significant main effect of test type, $F(4, 228) = 36.68$, $p < .0001$. This main effect across subjects represents a larger mean for percent error on fill-in-the-blank test types and comprehension test type, and foreshadows the effect of test direction found in a later analysis (Figure 1). The analysis yielded a significant interaction between study blocks and test type, $F(8, 456) = 9.1$, $p < .0001$. This interaction represents that levels of the study block influenced the effect of levels of test type, and vice versa (Figure 2). This interaction highlights the difference of the test study blocks on questions of recognition, where unidirectional transfer only occurred for reverse study blocks, and test type was only relevant to test study blocks.

The 3 x 2 x 2 subject analysis ANOVA analyzed study blocks, multiple choice and fill-in-the-blank test types, and forward and reverse test directions. This analysis yielded a significant main effect of study block, $F(2, 57) = 4.1$, $p < .05$. This main effect of study block illustrates a possible testing effect advantage, where percent error was higher on restudy items than items studied in both reverse and forward test study blocks (Figure 3). This main effect of the testing study blocks vs. restudy was analyzed using a Fisher's PLSD, which revealed a significant mean critical difference between reverse study block and restudy study block; restudy vs. reverse, $p = .006$. However, the Fisher's PLSD critical difference between forward study block and restudy study block was only marginally significant; forward vs. restudy, $p = .053$ (Table 3). The three-way subject analysis found a significant main effect of test type, $F(1, 57) = 81.28$, $p < .0001$. This main effect shows a higher percent error on fill-in-the-blank questions (Figure 4) that was significantly different than percent error on multiple-choice questions. There was a significant main effect of test direction, $F(1, 57) = 33.15$, $p < .0001$, where forward test

direction questions had significantly higher percent errors than reverse test direction questions. A significant interaction between study block and test type occurred, $F(2, 114) = 10.59$, $p < .0001$. This interaction reflects that the difference between study blocks is influenced by test type, and vice versa. There is still less percent error on multiple choice questions, but the study block influences the degree of this error across subjects (Figure 5). An interaction between study block and test direction was significant, $F(2, 114) = 15.7$, $p < .0001$. This interaction shows a possible testing advantage only present in reverse direction questions, which will also be analyzed further in the discussion (Figure 6). The interaction between test type and test direction was not significant, where the fill-in-the-blank test type retains a higher percent error regardless of direction, and forward test direction retains a higher percent error regardless of test type (Figure 7). The interaction between study block, test type and test direction was significant, $F(2, 114) = 5.29$, $p < .01$. This interaction highlights the inconsistency found in other significant interactions because of the varying degrees of influence between variables on overall performance (Figure 8).

In the 3 x 5 factorial items analysis ANOVA, the study blocks and all 5 question types were analyzed across all 30 paired associate items. In this analysis, there was no significant main effect of study type. There was a significant main effect of question type, $F(2, 54) = 8.66$, $p < .001$. This indicates a difference between the items assigned to each question type, but that these differences could be due to the question type or do the items assigned to question type. There was not a significant interaction between study type and question type. This interaction indicates that the main effect of question type is real, but cannot be isolated beyond the confound of item to question type.

A 3 x 2 x 2 item analysis ANOVA was used to analyze study blocks, test type, and test direction across items. The main effect of study block was not significant. The main effect of test

type was significant, $F(1, 12) = 18.3$, $p < .01$. The main effect of test direction was significant, $F(1, 12) = 6.03$, $p < .05$. The interaction between study block and test type was not significant. The interaction between test type and test direction was not significant. The interaction between study block, test type, and test direction was also not significant. This analysis yielded a marginally significant interaction between study block and test direction, $F(2, 12) = 3.75$, $p = .054$ (Figure 9).

Additionally, a subject analysis ANOVA was conducted for study blocks and comprehension question percent error. There was no item analysis conducted for comprehension questions because the question type was balanced across items. The subject analysis yielded no significant results. This represents no effect of study block on comprehension question performance. The null results indicate that the main effect of study block is not present on questions that applied the paired associates conceptually, without the manipulation of direction.

The seven-question survey found no prior familiarity relevant to the present experiment. 37 students who participated in this study reported using Quizlet before. 21 of participants reported using Quizlet for almost every class exam. Over half of the subjects reported using the notecard restudy mode as a retrieval practice (see Appendix).

Discussion

The results of this study are inconclusive because the 30-paired associate items were confounded. Although the items were randomly assigned to study blocks, they were not balanced across subjects. Although the items were randomly assigned from study blocks to question type, they were not balanced across items. The direction of each question type was also not balanced. Any significant results found in data analysis are inconclusive because the manipulation is flawed; significant differences between levels of the independent variables of study block, question type, and direction are not definitive because the differences found among the three study types could either be due to the present manipulation or to differences among the items assigned to the three study conditions. The hypothesis of a testing effect present in online self-study, and the hypothesis that the direction of self-study practice is relevant to the direction of retention are legitimate and significant results would be applicable to all current educational settings. A future study seeking significant results of these hypotheses should counterbalance the items presented within study blocks across subjects, where some subjects study 10 terms in restudy while others study the same 10 terms in reverse or forward test study. Future studies should also balance across question type, where one term is asked in forward direction multiple choice for one subject, reverse multiple-choice for another, forward fill-in-the blank for another, reverse fill-in-the blank for another, and comprehension for another. Future studies should have at least 5 different versions of the test. If the way an associate pair was studied and tested was balanced across subjects, any significant results could indicate a difference due to manipulation, and eliminate the possibility of differences due to items confounded to one level of the independent variable(s).

Despite the ambiguous nature of the results, results that were significant across subjects and items analyses are worth discussing. A significant main effect of test type occurred in both subject and item analyses that included and did not include direction as a factor. This main effect highlights the real difference between test questions of recall and of recognition. There were significantly more errors on recall questions than recognition questions (Figure 4), which confirms the common sense notion that recall questions are more difficult than recognition questions. This main effect of test type was present across subjects and items, which shows that a significant difference exists between questions of cued recall and of recognition for likelihood to remember previously studied stimuli (Figure 4). An interesting result was the difference of performance on fill-in-the-blank questions for the forward study block items in comparison to the performance on fill-in-the-blank questions for the other study blocks items (Figure 5). The forward study block was hypothesized to benefit from a generation effect, or at least have a higher level of understanding due to a retrieval practice of an entire concept. Although they did worse overall than the reverse study block, they had the least amount of errors on fill-in-the-blank questions than any other blocks. Despite the finding that matching initial tests to final tests does not boost testing effects (Rowland, 2014), it is possible that the forward studied terms benefit on the hardest question types, forward fill-in-the-blank questions, because the terms were studied in a rigorous, and generative forward cued-recall format. From the interaction of study type and test direction (Figure 6), we see that the most difficult questions, forward fill-in-the-blank questions (Figure 1), can be recalled slightly better when terms are studied in the forward retrieval practice. All of these interpretations are simply speculation because of the confound between terms and study blocks, and terms and question type. Because of the confound, it is

possible that the terms selected for the forward study were easier, or the forward terms randomly selected to questions of recall were easier than other recall terms.

A significant main effect of test direction occurred in both the item and subject analyses. This main effect and its interaction with test type highlights an underlying assumption of the experiment's directional hypothesis: regardless of test type, forward direction test questions are the most difficult (Figure 7). It is reasonable to assume that recalling a definition ranging from three to seven words is more difficult than recalling or recognizing a single term. Regardless of the test question, it should be fundamentally easier to remember a term when you are given the correct definition. Although the main effects of test type and test direction were significant across subjects and items, it is impossible to extend these results to new items. All main effect significant in subject analysis but not in item analysis are inconclusive because effects are impossible to generalize to new items.

The interaction between study block and test direction was significant in the subject analysis and marginally significant in the item analysis and is worth discussing. A significant interaction of study block and test direction reveals the possibility that the main effect of study block did produce a testing effect in the forward and reverse study blocks. Better results on reverse test questions over forward test questions exists only for the study blocks that practiced retrieval (Figure 9). For the notecard study block, the forward questions have a slightly better performance than reverse test questions, which is the opposite trend of the test study blocks. This indicates a possible testing effect present in the reverse direction questions. This trend in conjunction with the main effect of test direction indicates percent error on forward questions was significantly different than performance on reverse questions; it is possible that the forward questions were too difficult across study blocks and reached a performance ceiling. It is possible

that for the restudy block, direction of study did not matter whereas it did for test study blocks. All interpretations are inconclusive due to confounded paired associates; a follow up study looking at bidirectional transfer of paired associates studied in a repetitive and effective retrieval practice might find that forward direction of retrieval practice could produce a larger, more significant testing effect bidirectionally.

Although the hypothesis of a testing effect present in both forward and reverse study blocks on Quizlet and the hypothesis that direction of retrieval practice will influence direction and transfer of final retention cannot be confirmed or denied by the results of this study, the hypotheses themselves are novel and should be followed up. In initial follow up studies, experimenters should balance the items to study groups and test questions. Eliminating all confounds to this study should be the priority for any follow up experiment. A study employing Quizlet and the generation effe*ct* should examine the benefits of subjects creating their own Quizlet "set" over studying a previously made "set" on final retention, even if exposure is increased for the premade "set." A study of this nature could also include more comprehension type questions, rather than only six, to study the degree to which direction of retrieval practice or repetitive retrieval practice itself can affect the level of understanding of a concept. A significant finding would strengthen the phenomenon of the generation effect and also give real life implementations of its effects for current students.

It is possible that the notecard mode could have been used as a retrieval practice rather than restudy if students tried to recall the corresponding definition before reading it. One of the survey questions asked during part two of the experiment considered this possibility: "did you try to recall the corresponding definition on the note card before revealing to yourself?" Over half of participants in this study reported doing so. This possibility leaves room for follow up studies to

interpret if notecards can be used as an effective retrieval practice. In the 2009 study by Karpicke, Butler & Roediger, 40% of students reported using notecards as a study strategy, and it is arguable that rather than these students using notecards as a restudy method, that they are using it as a retrieval practice. Although the results of this study do not indicate any sort of advantage of notecards being used as a retrieval practice, an experiment that could collect data from the study session (for example, record the amount of time a participant stayed on the first side of the notecard before revealing the response) could determine if there is an advantage for terms studied this way.

Conclusion

The primary motivation of the current study was to acknowledge that effective self-study methods exist online and that students today are using them. Although the results of a testing effect were inconclusive, 37 students who participated in this study reported using Quizlet before. 21 of participants reported using Quizlet for almost every class exam. Additionally, students reported using notecards in an effective manner. The study was not able to conclude any real findings from the results, but it did confirm the common sense notion that what should be hard, is hard. The current experiment also leaves a lot of room for follow up experiments and further interpretations that take into account the real life study habits of students today with past findings, and seek to apply even more effectiveness and detail in the direction of memory and retention into a phenomenon that is tried and true throughout laboratory and classroom environments.

Reference List

Anderson, L. S. , Healy, A. F. , Kole, J. C. , & Bourne, L. E. , Jr. (2011): *Conserving time in the classroom: The clicker technique*, The Quarterly Journal of Experimental Psychology, 64:8, 1457-1462.
doi: http://dx.doi.org/10.1080/17470218.2011.593264

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior, 8*(4), 463-470. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1016/S0022-5371(69)80090-3

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642. doi:http://dx.doi.org.colorado.idm.oclc.org/10.3758/BF03202713

de Winstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*(6), 945-955. doi:http://dx.doi.org.colorado.idm.oclc.org/10.3758/BF03196872

Feder, K. L. (2011). *The past in perspective: An introduction to human prehistory*. New York: Oxford university Press.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1037/0022-0663.81.3.392

Karpicke, J. D., Butler, A. C., & Roediger,Henry L., I.,II. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471-479. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1080/09658210802647009

Leuba, H. R. (1966). Symmetry in paired associates. *Journal of Experimental Psychology, 72*(2),

287-293. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1037/h0023452

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*(4-5), 494-513. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1080/09541440701326154

Murdock, B. B., Jr. (1958). "Backward" associations in transfer and learning. *Journal of Experimental Psychology, 55*(2), 111-114. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1037/h0044180

Roediger, Henry L., I., II, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre, & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education (55); the psychology of learning and motivation: Cognition in education (55),* 1-36, Chapter xiv, 313 Pages. Elsevier Academic Press, San Diego, CA. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1016/B978-0-12-387691-1.00001-6

Roediger,Henry L., I.,II, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1111/j.1745-6916.2006.00012.x

Roediger, Henry L., I.,II, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1037/a0037559

Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker□assisted memory enhancement in college classrooms. *Applied Cognitive Psychology, 26*(4), 635-643. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1002/acp.2843

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 115-124. doi:http://dx.doi.org.colorado.idm.oclc.org/10.1037/a0034252

Young, A. P. (2015). *The discovery and adoption of the testing effect: Challenges of eliciting self-testing behavior in students* (Order No. 10003722). Available from Dissertations & Theses @ University of Colorado, Boulder; Dissertations & Theses @ University of Colorado System; ProQuest Dissertations & Theses A&I. (1760171660).

Table 1: study block assignment to question types

	Forward		Reverse		Comprehension	
	MC	Fill	MC	Fill	MC	Fill
Forward Study	2	2	2	2	1	1
Restudy	2	2	2	2	1	1
Reverse Study	2	2	2	2	1	1

Note: Key: MP= multiple choice, Fill= fill-in-the-blank. Each subject studied 10 terms in each study block. Each term was randomly assigned to a question type (MP, fill, and comprehension). The assignment to study block was the same for all subjects. The assignment to direction question types was the same for all subjects. The comprehension question types were balanced across subjects.

Table 2: 3 X 5 Subject Fisher's PLSD comparison of Study Block critical mean difference

	Mean Difference	P- Value
Forward vs. Restudy	-0.045	0.0973
Forward vs. Reverse	0.021	0.4419
Restudy vs. Reverse	0.066	0.0161

Note: The 3 x 5 subject analysis of variance had a significant main effect of study block. The Fisher's PLSD showed that significant difference between study blocks occurred between the restudy block and reverse study block.

Table 3: 3 x 2 x 2 Subject Fisher's PLSD comparison of study block critical mean difference

	Mean Difference	P- Value
Forward vs. Restudy	-0.06	0.053
Forward vs. Reverse	0.026	0.4037
Restudy vs. Reverse	0.086	0.0061

Note: The 3 x 2 x 2 subject analysis of variance had a significant main effect of study block. The Fisher's PLSD showed that the significant difference between study blocks occurred between the restudy and reverse block. It also shows a marginally significant difference between the forward and restudy blocks, which was not present in the ANOVA that excluded retention question direction as an independent variable.

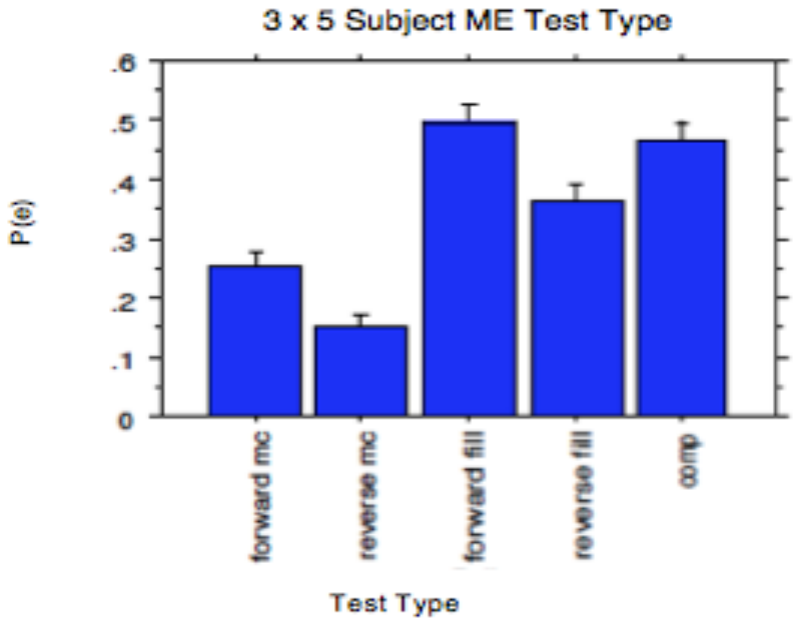

Figure 1: 3 x 5 subject ANOVA main effect of test type, where a larger percent error is significantly different for fill-in-the-blank and comprehension questions.

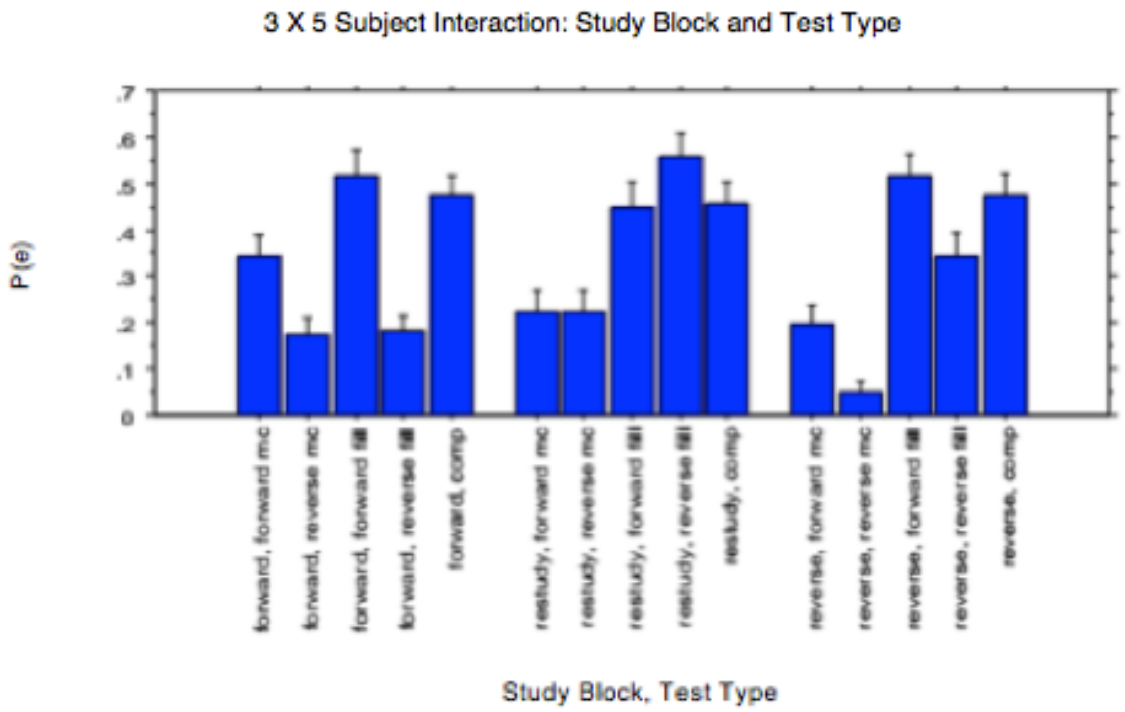

Figure 2: the 3 x 5 subject ANOVA significant interaction between study block and test type. This interaction represents the effect of study block on test type, and the effect of test type on study block. Where recognition questions show a different trend in percent error by study block than cued recall questions.

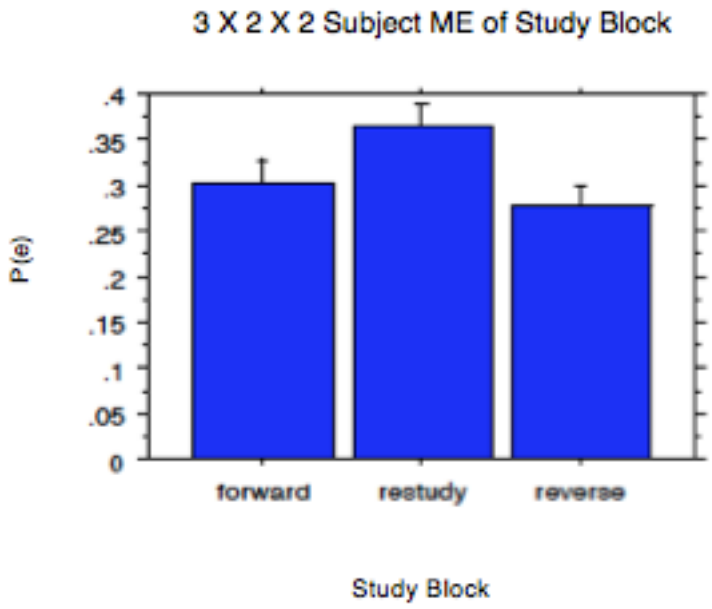

Figure 3: 3 x 2 x 2 subject ANOVA showed a significant main effect of study block, where percent error is higher in restudy blocks than test study blocks.

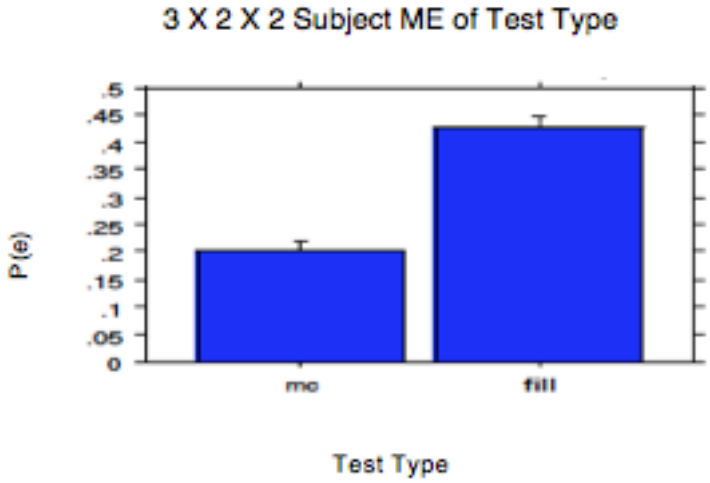

Figure 4: 3 x 2 x 2 subject ANOVA showed a significant main effect of test type, where percent error on fill-in-the-blank questions was significantly higher than multiple choice questions. This significance illustrates that cued recall questions are more difficult than recognition questions.

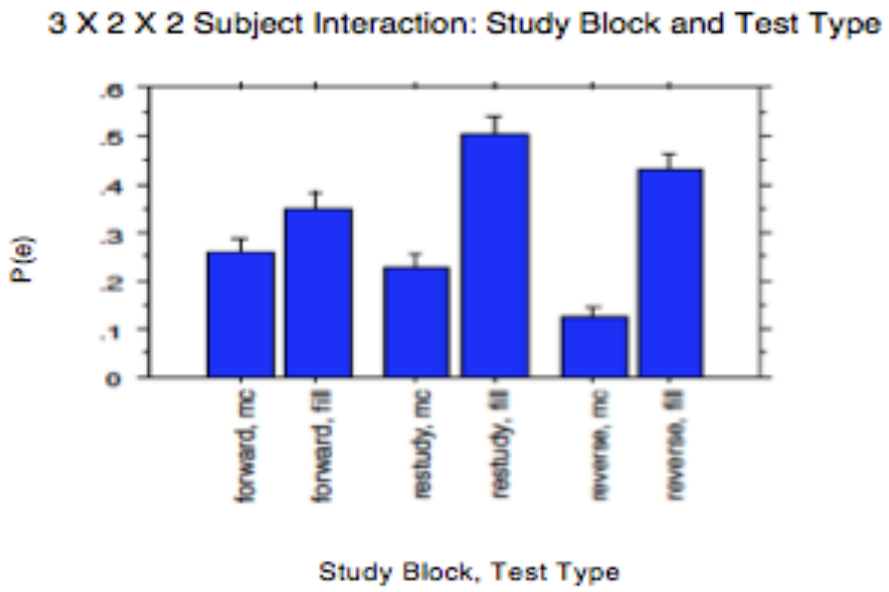

Figure 5: 3 x 2 x 2 subject ANOVA showed a significant interaction between study block and test type. Recognition questions still had significantly less errors but the effect of study block influences this percent error across subjects.

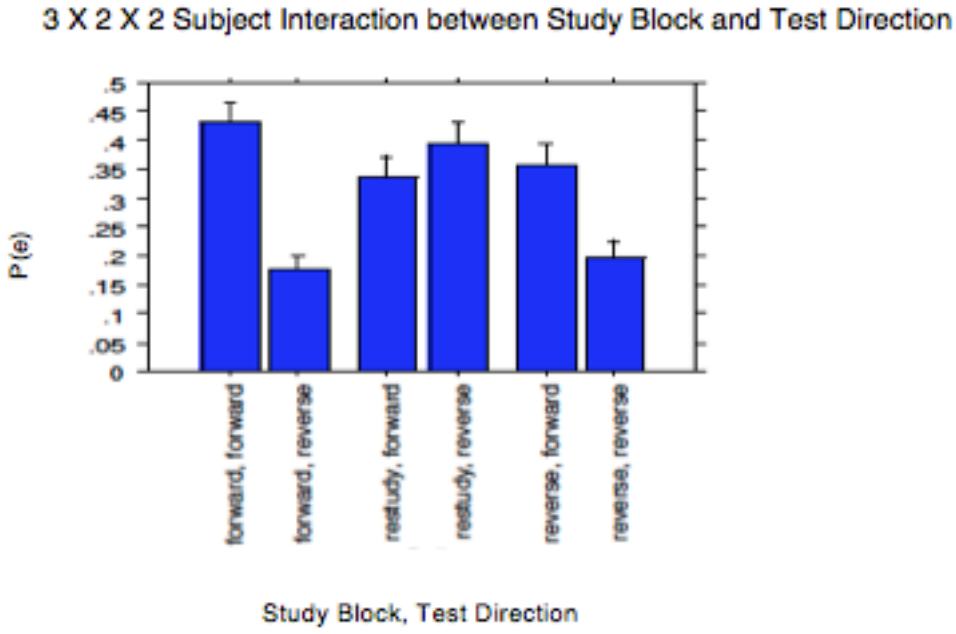

Figure 6: 3 x 2 x 2 subject ANOVA showed a significant interaction between study block and test direction. Direction of study influences direction of retrieval in test study blocks.

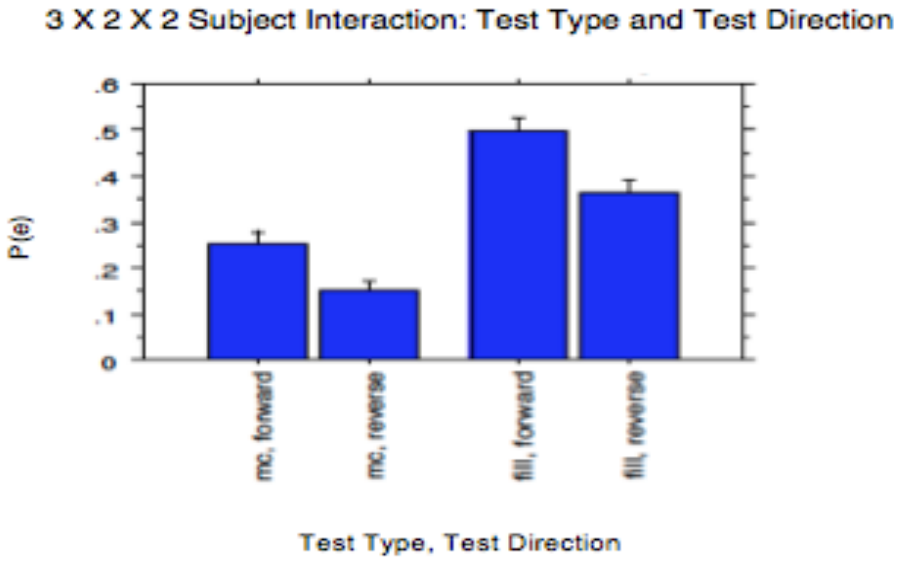

Figure 7: 3 x 2 x 2 subject ANOVA non-significant interaction between test type and test direction. Fill-in-the-blank test type retains a higher percent error regardless of direction, and forward test direction retains a higher percent error regardless of test type.

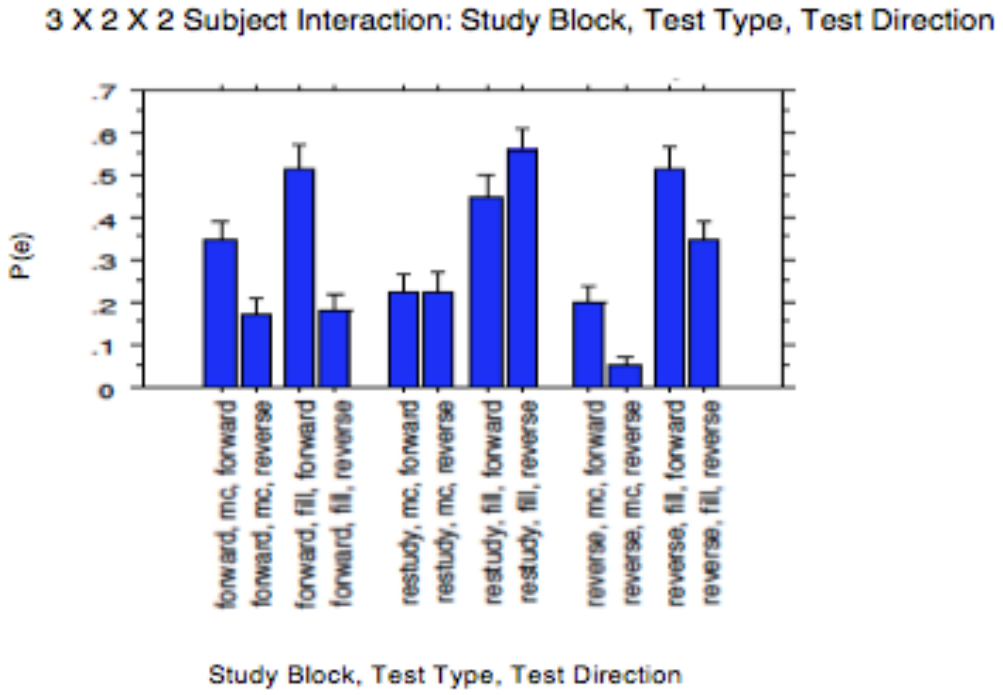

Figure 8: 3 x 2 x 2 subject analysis ANOVA interaction between study block, test type and test direction was significant. This illustrates the inconsistency of results because of the varying results and limited questions per block and direction.

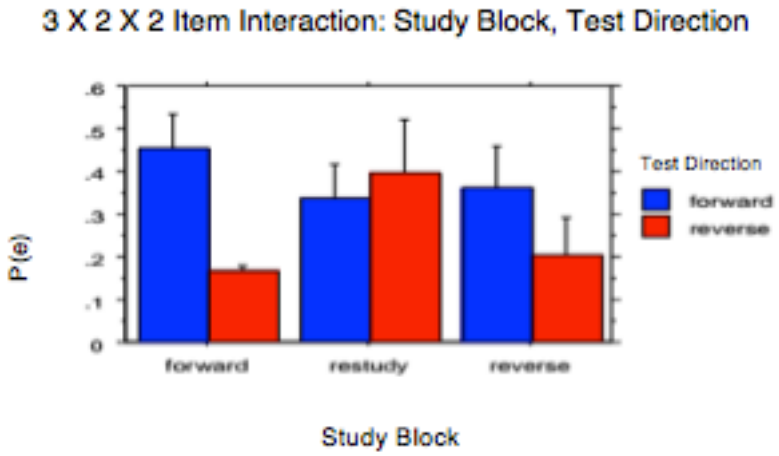

Figure 9: 3 x 2 x 2 item analysis, marginally significant interaction of study block and test direction. This interaction shows that better results on reverse test questions over forward test questions exists only for the study blocks that practiced retrieval, meaning that directional study only mattered in testing study blocks and that a testing effect may be present. This indicates a possible testing effect present in the reverse direction questions.

Appendix

Paired Associates:

Forward Paired Associates:

1. arboreal: life in the trees
2. mastaba: tombs for the elite
3. dendrochronology: tree ring dating technique
4. midden: area used for trash
5. stadial: period of increased glaciation
6. archaeology: study of humans material past
7. prehistory: period of history before written records
8. provenance: history of ownership of an artifact
9. excavation: act of digging up earth for study
10. site: any place with human material remains

Reverse Paired Associates:

11. ecofact: environmental remains of humans
12. artifact: manufactured remains of humans
13. microlith: small flaked stone tool
14. cache: objects purposefully buried
15. token: pieces of shaped clay
16. pluvial: period of increased rainfall
17. animism: belief that natural objects have spirits
18. morphology: analysis of the shape of bones
19. debitage: waste left from manufacturing stone tools

20. varves: annual deposits made by glacial periods

Restudy Paired Associates:

21. strata: levels of earth studied

22. diffusion: transmission of ideas from culture to culture

23. cuneiform: first system of written language

24. subsistence: material necessary to sustain life

25. bipedalism: humans final evolution to walking upright

26. paleolithic: early phase of the stone age

27. tumulus: mound built above tombs

28. burn: stone tool for engraving

29. petroglyph: designs etched into rocks

30. knapper: manufacturer of stone tools

Survey Questions

Your answering of these questions simply helps inform and further the research.

1. What year are you in college?
    a. Freshman
    b. Sophomore
    c. Junior
    d. Senior
    e. 5th year senior
2. Have you taken the course at CU Boulder ANTH 2200: The Archaeology of Human History and, or are familiar with the test *Past in Perspective*?
    a. Yes
    b. No
3. Are you familiar with archeology at college level?
    a. Yes
    b. No
4. On a scale of 1 to 7, how familiar with these terms were you prior to your participation?
    (1 = not familiar at all; 7 = very familiar)

    1 2 3 4 5 6 7

5. When in the note card study mode, did you try to remember the corresponding definition before you flipped the card?
    a. Yes
    b. No
    c. Don't remember
6. Have you used quizlet before to study?
    a. Yes
    b. No
7. (If you answered yes to #6) On a scale of 1 to 7, how often do you use quizlet when studying for classes?
    (1= only used it once or twice; 7= use for every class)

    1 2 3 4 5 6 7

Test Question Examples

Multiple-choice forward:

14. Stadial means:
    a. Annual deposits made by glacial periods
    b. A period of increased rainfall
    c. Levels of earth studied
    d. A period of increased glaciation

Multiple-choice reverse:

16. A stone tool for engraving is called a:
    a. Knapper
    b. Burn
    c. Cache
    d. Token

Fill-in-the-blank forward:

15. An artifact is _____

Fill-in-the-blank reverse:

11. The manufacturer of stone tools is called _____

Comprehension question (Test A)

9. In order to live, humans need three things- water, food, and shelter. These are items of:
    a. Diffusion
    b. Cuneiform
    c. Subsistence
    d. Provenance

Comprehension question (Test B)

9. In order to live, humans need three things- water, food, and shelter. These are items of _____