SPACED VERSUS MASSED TESTING IN A COLLEGE CLASS: AN EXPLANATORY ITEM RESPONSE MODEL

A Master's Thesis

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree of

Master of Arts

By

Joseph W. Pirozzolo

May, 2016

SPACED VERSUS MASSED TESTING IN A COLLEGE CLASS: AN EXPLANATORY ITEM RESPONSE MODEL

An Abstract of a Master's Thesis

Presented to

The Faculty of the Department

of Psychology

University of Houston

In Partial Fulfillment

Of the Requirements for the Degree of

Master of Arts

By

Joseph W. Pirozzolo

May, 2016

ABSTRACT

Many studies have shown that distributed study is more effective than massed study. In the present study we were interested in the effects of frequent testing, student ability, practice quizzing, and item form (multiple choice or short-answer) at the pre-tests and the post-test. Two sections of an undergraduate psychological methods class were taught by the same professor at similar times of day. In the frequent testing class, 8 midterm exams were spaced about 1 exam every 2 weeks. In the standard class, 2 midterm exams were spaced about 1 every 8 weeks. All exams, including the final exam, consisted of both multiple choice (MC) and short answer (SA) questions. The form of questions (MC or SA) during the midterm exams was called the pre-test method and the form on the final was called the post-test method. Both classes took the same comprehensive final exam. Only final exam data was analyzed in this study. An explanatory item response model (EIRM) was used to estimate the effects of the person predictors: student ability, testing frequency, and quiz assignment, and the item predictors: pre-test and post-test method on final exam performance. Not surprisingly, student ability and post-test method explained the most variance in item responses of any of the predictors in the model. Testing frequency also significantly influenced item responses. A marginally significant interaction between testing frequency and post-test method was also observed. We conclude that frequent testing (spacing) improves performance relative to massing, however, the largest benefits are expected to be seen on recall memory tasks. We also argue that the benefits of frequent testing generalize across populations and conditions.

Keywords: Testing effect, spacing effect, desirable difficulties, explanatory item response models

ACKNOWLEDGEMENTS

I would like to thank my committee co-chairs Drs. David Francis and Donald Foss for providing critical direction and feedback during all stages of this project.

I would also like to thank committee members Drs. Tammy Tolar and Lynne Steinberg for offering me their perspectives on this project, and to Dr. Paulina Kulesz for her gracious help with statistical analysis.

TABLE OF CONTENTS

Desirable Difficulties ……..1

The Testing Effect …….…….…….3

The effect of quizzing …….…….…….…6

The Spacing Effect ……...…….…….…….7

Applications of Item Response Models ……..……..…….……...10

The Present Study ……13

Method ……**14**

Participants ……...14

Materials ……...14

Midterm exams ……14

The final exam ……15

Procedure ……..15

Person predictors ……..15

Student ability ……15

Missing data ……...16

Testing frequency ……17

Quiz Assignment ……17

Item predictors ……18

Pre-test method ……...18

Post-test method …….…….18

Analysis ……18

Hypotheses ……19

Results ……..**20**

Random Effects ……22

Fixed Effects ……24

Main effects ……24

Interactions ……..27

Discussion ……...…….**30**

Random Effects ……30

Fixed Effects ……32

Main effects ……32

Interactions ……..35

Implications……...36

Limitations ……...37

Future Directions ……39

Conclusions ……..39

LIST OF TABLES

Table 1. Two-way ANOVA Testing Mean Differences in SAT by Testing Frequency and Exclusion.

Table 2. Two-way ANOVA Testing Mean Differences in GPA by Testing Frequency and Exclusion.

Table 3. Percentage of Students by Number of Quizzes Taken.

Table 4. Variance and Variance Reduction of Models.

Table 5. Fit Statistics.

Table 6. Main Effects- Pass Rate by Condition.

Table 7. Two-Way Interactions- Pass Rate by Condition.

Table 8. Three-Way Interactions- Pass Rate by Condition.

Figure 1. A line plot showing the pass rate for students by ability level.

Figure 2. A line plot showing the pass rate by testing frequency.

Figure 3. A line plot showing the marginally significant ($p = .06$) interaction of testing frequency and post-test method.

DEDICATION

To my Parents, my wife,

and my mentors- without whom

this would not have been possible.

Spaced Versus Massed Testing in a College Class: An Explanatory Item Response Model

The primary objective of higher education is to develop knowledgeable and intelligent citizens. Some cognitive psychologists, reflecting on recent research, argue that educators neglect some of the most robust instruction and testing strategies that have been shown to improve long-term memory retention. The retention of course material over time is critical to developing successful college graduates. A goal of higher education should be to improve the performance of students in the long-term and to facilitate the transfer of knowledge and skills to related problems. Many college course designs feature too few examinations and limited variability in the types of responses required of students. There is good evidence that colleges and universities could increase student graduation rates and improve student success post-graduation by improving the ways in which students are evaluated.

The aim of this paper is to identify learning conditions that improve learning in the college classroom and evaluate the effects of the application of these principles in an original experimental design. In the following sections, research on the conditions that lead to improved long-term learning will be presented and discussed. This paper will describe the method and discuss the results of an experiment where an explanatory item response model was used to evaluate the effects of person and item predictors on final exam performance in a college course.

Desirable Difficulties

Research on learning has reliably shown that conditions that challenge the learner are often the best conditions for long-term memory retention (Bjork & Bjork, 2010). Most times the learner is unaware of the disadvantages of many common learning strategies (Kornell & Bjork, 2007). The conditions that yield the best long-term outcomes have been coined by Robert A.

Bjork as “desirable difficulties” (Bjork, 1994; Bjork & Linn, 2006). Bjork has spent a career investigating the conditions that benefit memory. Through his many investigations, Bjork argues that conditions that make learning easy and allow for immediate and rapid improvement of performance do not facilitate long-term memory. One of the most counterintuitive findings in this research area is that study time, specifically for college students, does not reliably predict test performance (Plant, Ericsson, Hill & Asberg, 2004). Study time is not a good predictor of performance because the quality of study can vary greatly for individual students. Desirable difficulties deal directly with the qualitative aspects of study and how they inform the performance of the learner.

As its name indicates, desirable difficulties are those conditions that make the learning process challenging. They are difficult in a way that is beneficial to learning (Bjork & Bjork, 2010). The central theme of desirable difficulties is that long-term memory is best facilitated under conditions where the learner engages in active forms of learning, distributes learning over time, and alters many features of learning. One desirable difficulty that will be extensively discussed here is generation. Generation of learned information can be done in many ways, including: taking a test, mental rehearsal, and oral recitation (e.g., teaching or explaining material to another person). Testing students has been shown to have benefits other than simply evaluating knowledge (Foos & Fisher, 1998; Little, 2010; McDaniel & Fisher, 1991). This is the general principle behind the finding that frequent retrieval of information leads to better memory over time, also called the testing effect. Retrieval practice or testing is a challenging and beneficial form of study and thus accounted for in the desirable difficulties approach to learning.

Bjork and many other researchers in cognitive science have also found that it is important to distribute or “space” study and generation (or testing) of material throughout the learning

process. Spacing is best described as extending study and testing sessions over time, such that information is not automatically available in short-term memory from one session to the next. Spacing study contrasts with massing study, whereas massing refers to spending large amounts of time studying in one or very few sessions. In practice, one of the most common forms of massing study is colloquially referred to as "cramming," where a student delays studying until just before an examination. Research has shown that spaced study conditions are more effective than massed study conditions, especially when retrieval is delayed (Bjork & Bjork, 2010).

Desirable difficulties encompass the main topics discussed thus far: testing and spacing. In the following section, each of these desirable difficulties will be briefly reviewed, followed by the present study, which incorporated desirable difficulties to improve retention in an undergraduate level course.

The Testing Effect

It has long been known that more active forms of learning are more effective than passive strategies for long-term retention. This was, perhaps, first documented by Francis Bacon in 1620 when he wrote that a phrase can be memorized more quickly if one attempts to recite it during the learning phase, as opposed to repeated rereading (Bacon, 1620/2000; Roediger & Karpicke, 2006a). More recently, educators and psychologists have made similar observations of differences between active and passive learning (Karpicke, 2012). Testing memory is one form of active learning that has been reliably shown to improve long-term memory. This phenomenon has come to be known as "the testing effect." Most testing effect experiments feature the contrasting of conditions that requires taking a memory test of learned information and conditions where restudying or rereading is the primary learning strategy. Testing has proved to be one of the most beneficial memory tools.

In a review of the testing effect by Roediger and Karpicke (2006a), the authors cite Tulving's (1967) finding that testing memory may be just as beneficial as studying. Tulving's goal was to investigate whether tests are simply assessments of knowledge or useful training manipulations. Tulving created 3 groups of subjects that learned a word list using different procedures during 24 trials. Subjects participated in both study trials, where the words on the list were presented and learned, and test trials, where the subjects performed a free-recall test of the words they had learned. Each set of four trials was considered one cycle. It is helpful for the understanding of Tulving's method if one lets S represent a study trial and T represent a Test trial. The standard learning condition, created to represent a traditional procedure for learning, took a test after every study trial, completed as STST (study-test-study-test) for 6 cycles or 24 total trials. Subjects in the repeated-study condition studied the list 3 times and then took a test (SSST) for 6 cycles. Subjects in the repeated-test condition studied the word list and then took 3 consecutive tests (STTT). Tulving found that despite the rather large differences between the 3 learning conditions, the groups had relatively similar learning curves over the course of the experiment. The standard and repeated-study conditions both recalled about 20 words at the final test trial and subjects in the repeated-test condition recalled about 18.5. This is particularly interesting because the subjects in the repeated-study condition had 6 more study trials than subjects in the standard condition. Perhaps more surprising was that the repeated-test group performed nearly as well as both of the other groups despite their exposure to only 6 total study trials whereas the standard and repeated-study conditions had 12 and 18 study trials respectively. Additionally, Roediger and Karpicke (2006a) argue that the repeated-test condition may have not had the same benefit of sustained short-term memory during the experiment because their last test trial of each cycle was two trials removed from a study trial.

The results of Tulving's (1967) study were surprising to many researchers. The implication of this study is that a taking a test may be just as beneficial as studying. However, Karpicke and Roediger (2006) believed that Tulving's examination of the power of testing, while innovative and thoughtful, might have even underestimated the power of testing. Upon consideration of Tulving's methods, Karpicke and Roediger noticed that in order to eliminate effects due to exposure time, Tulving limited both study and test trials to 36 seconds. Karpicke and Roediger thought that 36 seconds might be too short a period to accurately assess memory using a free-recall method. The investigators also believed that testing might have an even larger effect when recall is delayed. In order to provide a more accurate measure of the effect of testing, Karpicke and Roediger replicated Tulving's study, but presented subjects with 40 words at a rate of 1 word every 3 seconds (2 minutes of study time total). The authors also equated testing time to study time, allowing 2 minutes for free-recall. Finally, to achieve a more accurate measure of true learning, Karpicke and Roediger included a delayed retrieval post-test of memory, 1 week after completion of the study. The post-test in this study was a 10 minute free-recall test. The participants were instructed to draw a line under the last recalled word written every minute, thus this design gave the researchers the opportunity to evaluate how recall develops over time for the 3 conditions.

The results of Karpicke and Roediger (2006) indicated, as Tulving's study, that the learning curves for the 3 conditions are very similar. However, the authors found that the standard group performed better than the other two groups over their last 4 tests. It was also noted that the repeated-testing group did not perform as well at the other two groups on early test trials, but quickly caught up later in the study. Perhaps most interestingly, and relevant to the target of this paper, was the finding that at the delayed recall post-test, the standard and repeated-

testing groups performed significantly better than the repeated-study group. The standard and repeated-testing group recalled 68% and 64% of words respectively, whereas the repeated-study group recalled only 57% of the words. The most important implications of this study are that: 1) testing, when combined with study trials, serves as a valuable learning experience and 2) individuals who engage in more tests of memory are more likely to retain information over time.

To explain these testing effects, researchers have theorized that each generation of learned material modifies and strengthens the memory of that information, which increases the likelihood that it can be remembered again (Bjork, 1975; Schacter, Norman, & Koutstaal, 1998; Whitten & Bjork, 1977). The processes that underlie learning from testing have also been shown to be quite different from those involved in traditional studying (Toppino & Cohen, 2009). This evidence-based theory continues to be the leading description of why testing memory is beneficial for long-term retention.

In addition to the consistent finding that tests inhibit forgetting of learned information, some researchers have also found evidence for indirect effects of testing. Indirect effects of testing can be described as those that improve learning of material but not directly related to retrieval on tests. Research on the indirect effects of testing suggests that tests make subsequent study sessions more effective (Arnold & McDermott, 2012). Similarly, intermittent memory tests have also been shown to improve students' attention to instruction (Szpunar, Khan, & Schacter, 2013).

The effect of quizzing. Although the research on the testing effect discussed thus far primarily deals with word-list stimuli, researchers have also found similar testing benefits using educational stimuli. Middle school and high school students perform better on material that appeared on a quiz or practice test than material that had not (Lipko-Speed, Dunlosky, &

Rawson, 2014; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2013). This testing effect has been found to be independent of whether the quiz questions are short-answer or multiple-choice. Most relevant to the population studied in this paper, researchers have also found testing effects in college courses using a variety of methods. Online quizzes have been shown to improve college student performance regardless of whether the course-meeting format was traditional or web-based (Pennebaker, Gosling, & Ferrell, 2013; Wiklund-Hornqvist, Jonsson, & Nyberg, 2014). Frequent testing seems to provide a manageable and affordable way to improve the retention of course material. However, there is some evidence that the degree to which quiz questions resemble test questions plays a role in the effectiveness of quizzing (Wooldridge, Bugg, McDaniel, & Lui, 2013). This finding suggests that there may be strict limitations for the transfer of knowledge gained from testing to other material. Additionally, Lipko-Speed, Dunlosky, and Rawson (2014) argue that practice tests with feedback may only have significant benefits for students who put forth the required amount of effort to generate responses on their own.

The Spacing Effect

The spacing effect is described by researchers as the finding that long-term memory is improved for information that is distributed into multiple presentations rather than learned all at once or in very few presentations (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Crowder, 1976; Dempster, 1988). The benefit of distributing practice over time has been studied since the late 19th century. Hermann Ebbinghaus may have been one of the first scientists to publish work on the benefit of distributing practice as opposed to massing practice. Ebbinghaus found that it took him fewer trials to memorize a long syllable series when he spaced his practice over 3 days than when he repeated practice trials sequentially (Dempster, 1988; Ebbinghaus, 1885/1917).

Though the benefits of the "spacing effect" have been known for a relatively long time, there is little evidence that it is being utilized in educational instruction. There are several reasons for the lack of spacing in education, one of which is that the benefits of spacing have not been adequately demonstrated in educational settings (Dempster, 1988).

It was once thought that spacing significantly benefited recall, but did not benefit inductive learning, the process of learning a stimulus category by examining exemplars. However, Kornell and Bjork (2008) found that participants identify paintings by artist name more accurately if the paintings are presented in a spaced condition than if they are massed. In this paradigm, spacing refers to the order of stimulus presentation. In the massed condition all of each artist's paintings are presented sequentially (e.g., A-A-A B-B-B C-C-C), whereas in the spaced condition paintings made by different artists are spaced or interleaved (e.g., A-B-C A-B-C A-B-C). The spacing and massing conditions were manipulated within participants such that, for each participant, paintings by a specific artist were either presented in a spaced or massed fashion. During the study phase participants studied slides of paintings with the name of the artist listed below the painting. After the study phase participants were asked to identify novel paintings by an artist that they had studied during the experiment. The results of this experiment show that the participants were significantly more accurate in identifying novel paintings by artists who were studied in the spaced condition. Additionally, participants were asked if spacing or massing led to better learning or both conditions were equal. The majority (approximately 65%) of participants believed that massing was a better learning condition even though over 75% of participants performed better on artists that were spaced. These findings show that spacing the presentation of learning stimuli improves learning, however, people can be largely unaware of the benefits of spacing.

In a review of the implications of the spacing effect, Dempster (1989) reported that the spacing effect is not only limited to stimulus presentation or studying, but also applies to testing. Dempster's review found that spacing tests improves memory at a later point as compared to massing tests (Whitten & Bjork, 1977). Spacing tests at expanding intervals (increasing the retention interval after each test) also seems to optimize the benefits of test spacing (Landauer & Bjork, 1978). Rothkopf (1966) discovered that spacing not only affects repeated items, but also can improve students' performance on questions that have not been tested before. Dempster highlights the many conditions where spacing improves memory and also some specific conditions where spacing does not have significant benefits, for example in pre-school children (Dempster, 1989; Toppino & DiGeorge, 1984).

Several researchers have investigated why distributing practice tends to lead to better performance. One theory of spacing states that with longer distances between encoding information the likelihood of encoding being qualitatively different increases (Wichawut, 1972). Therefore, spacing encoding increases the likelihood that separate sessions will be encoded differently, potentially leading to more associations. Additionally, some argue that spacing encoding at longer distances between sessions is more effective than shorter distances until some limit. Spacing presentations and testing seems to be beneficial for learning because it allows for improved encoding and forces retrieval of information that is not automatically available (Bjork & Allen, 1970).

Spacing studying and testing has been reliably shown to have rather large benefit for long-term learning. Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) performed a relatively exhaustive examination of the literature on 10 techniques that are often used in classrooms to benefit learning. The authors reported that both testing and spacing have high

utility as educational techniques. These were the only two techniques of the 10 reviewed that received a high utility rating. Other conventional methods were found to have inconsistent or suspect utility when being applied to real learning situations.

Applications of Explanatory Item Response Models

Item response theory (IRT) models have been used to evaluate the effectiveness of tests for decades. IRT models were designed using the basic principle that responses to questions are the product of both person and item characteristics. IRT models are nonlinear regression models that express the probability of endorsing specific responses to test items in terms of the examinee's ability and one or more parameters that describe the relation between the item and ability (Hambleton & Jones, 1993). These characteristics of IRT models make it possible to link item responses, such as correct or incorrect answers to test items, to examinee ability and to understand how items differ in difficulty and their ability to discriminate between examinees of different ability levels (Hambleton & Jones, 1993).

Explanatory item response models (EIRMs) are a specific type of IRT model that offer researchers the opportunity to ask and answer interesting questions that explain item difficulty in terms of characteristics of items and person ability in terms of measured characteristics of individuals. Recently, the field of education has begun to use EIRMs to understand the effects of person and item predictors on test performance at the item level. Person predictors are variables that describe individual differences. Some common person predictors are: estimates of general ability (e.g., grade point average), experimental condition, gender, and many others that help explain the specific ability measured by the test (e.g., knowledge of psychological methods). Item predictors are factors that explain differences in item difficulty, such as different levels of item content or item format. One important feature of EIRMs is that they model the likelihood of

a correct response to an item based on both person and item predictors. This gives researchers the ability to determine the likelihood of correct responses for different levels of person and item factors. De Boeck and Wilson (2004a) note that the advantage of analyzing data using EIRM is that it offers an interesting joint analysis of both person and item factors.

There are several different types of EIRMs. One of the simplest models is the 1-parameter logistic model, also called the 1PL model or Rasch model. This model is called a 1-parameter model because it is designed to estimate only item difficulty, and therefore effectively control for the fact that some items may be more difficult than others. Most directly related to the present study, another benefit that EIRMs offers researchers is the ability to control the factors included in the model in estimating the effects of other factors of interest. This attribute gives researchers statistical control over variables, especially in situations where random assignment of people or items to conditions is not possible and the characteristics of interest are not independent. Overall, EIRMs offer an interesting method of analysis that allows for individual differences among people and items. The statistics provided by EIRMs also allow for an interpretation that is easily scalable and applicable to real life situations.

Rasch models can be used with outcomes that are binary, ordered-category, and nominal category (Kulesz, 2014). The ability of the Rasch model to analyze binary data makes it most appropriate to utilize for educational data, where items are generally scored as correct or incorrect (as in the present study). The Rasch model can expressed as

$$p_i(\theta) = \frac{1}{1+e^{-Da(\theta-b_i)}} \quad (1)$$

where p_i represents the probability of passing (answering correctly) item *i* for a given level of ability, θ (Allen & Yen, 2002). In the above representation of the Rasch model, *D* is a constant and *a* represents the item discrimination function, which accounts for the change in probability across different levels of ability. Since *a* in this model is constant across all items, the Rasch model assumes that all items have an equal discrimination, thus the Rasch model is a one-parameter model logistic regression model. The constant D serves to scale the probability function so that the Rasch model expressed in the logistic form in Equation 1 is approximately equivalent to expressing the Rasch model using the normal ogive function. The expression $(\theta - b_i)$ in the Rasch model simply means that the probability of passing an item is a function of the difference between examinee ability θ and item difficulty b_i. In other words, as the difference between ability and difficulty increases in a positive direction (i.e., ability is greater than item difficulty), then the probability of correctly answering a question increases; as the difference increases negatively, then the probability of correctly answering a question decreases.

When the Rasch model is used as an explanatory item response model, it is stipulated that the probability of a correct response depends on both person and item characteristics. Person and item characteristics are used as covariates or predictors of performance. The EIRM of interest in the present study is a logistic regression model and can be expressed as

$$\eta_{pi} = \sum_{j=1}^{J} \theta_j Z_{pj} + e_p - \sum_{k=1}^{K} \beta_k X_{ik} + e_i \tag{2}$$

In this expression of the explanatory Rasch model η_{pi} represents the log odds of a correct response for person p on item i (De Boeck & Wilson, 2004b). The first expression to the right of the equal sign is a representation of the person contribution in the model, where θ represents examinee ability, Z represents the fixed effect of a person predictor, and e_p represents the error in explaining ability as a function of person covariates. Similarly, for the item contribution β item difficulty, X represents fixed item effects, and e_i represents error in explaining item difficulty in terms of item characteristics. In full, this EIRM stipulates that the log odds of passing an item is equal to the total person contribution minus the total item contribution. Using EIRMs one can make inferences about the effects of person and item predictors and their interaction.

The Present Study

Frequent testing and distributed practice are often not present in many common study strategies. Reading and rereading a textbook is perhaps one of the most common forms of studying for an examination (Kornell & Bjork, 2007). However, this approach to learning has poor long-term results. In the following experiment we attempted to incorporate frequent testing and a spaced-learning paradigm to improve long-term retention in an undergraduate course. We also investigated the effect of taking practice quizzes on final exam performance. In this study the effects of spacing is are measured by the performance differences between two college classes with different testing schedules (i.e., testing frequency) and the testing effect is evaluated by the differences in performance on repeatedly asked and novel questions (i.e., the difference between pre-test method and post-test method). Using an EIRM framework, we identified student ability, testing frequency, and quiz assignment as person predictors that play roles in

final exam performance. We also identified item pre-test method (i.e., Short-Answer, Multiple-Choice, or Untested) and post-test method (i.e., Short-Answer, or Multiple-Choice), and their interaction as item predictors that may influence item difficulty at the post-test. In the present study, we were interested in how these person and item predictors influence final exam performance and interact in a real, semester-long college class as evidenced by their effects on person ability and item difficulty in an EIRM of the cumulative final exam.

Method

Participants

Undergraduate students (N = 237; frequent = 144) enrolled in two sections of a psychological methods course at the University of Houston participated in this study. Students took exams as part of the normal requirements for completion of the semester-long course. We required that students 1) completed more than half of all scheduled exams and 2) completed the final exam for their data to be included in the study. Of the 237 students enrolled in one of the two classes, 22 students (frequent = 12, 8.3%; standard = 10, 11%) were excluded because they took half of the scheduled exams or less and 47 students (frequent = 36, 25%; standard = 11, 11.8%) were excluded because they did not take the final exam. Twenty-nine students (frequent = 25, 17.4%; standard = 4, 4.3%) officially withdrew from the class, and therefore met one or both of the exclusion criteria. In total, 48 (frequent = 37, 25.7%; standard = 11, 11.8%) students were excluded for meeting one or both of the exclusion criteria. After exclusion, data from 189 (frequent = 107) students were analyzed.

Materials

Midterm exams. Midterm exams and the final exam were comprised half of multiple-choice (MC) and half of short answer (SA) questions. On MC questions students were instructed

to select the best of 4 or 5 alternatives. SA questions were fill-in-the-blank or cued recall questions, which could be sufficiently answered in 1-2 sentences. Students in both classes had the opportunity to take 48 midterm exam questions during the semester, albeit on different schedules in order to operationalize the spacing effect. Specifically, the number of total midterm exams and the number of questions per exam differed across the two classes so that classes differed in the spacing of exams, but did not differ in the number of times that any given material was tested. The instructor gave midterm exam feedback during lecture time, approximately one week after each exam.

The final exam. A comprehensive final exam consisted of 64 total questions. The final exam was comprised of 32 MC questions and 32 SA questions. Both the MC and SA question sets were comprised of 16 questions that had been tested previously on the mid-term exams (repeated questions) and 16 questions that were novel, consisting of topical information that had not been tested previously. Of the 16 repeated MC and SA questions, 8 were identical to questions participants had taken previously. The remaining 8 questions had their presentation method inverted, covering the same material, but in the opposite question form (MC or SA). For example, an MC question on the final exam that had been previously taken as an SA question on a mid-term exam is an inverted question.

Procedure

Person predictors. The person predictors in this design were: 1) student ability, 2) testing frequency, and 3) quiz assignment.

Student ability. Student SAT, ACT**,** and GPA scores were obtained from the university registrar. Each score was not available for every student. One hundred twenty five (frequent = 61) students had GPA scores, 103 (frequent = 62) students had SAT scores, and 35 (frequent =

16) students had ACT scores. Seventy-four (frequent = 43) students had only one score, 63 (frequent = 36) students had only two scores, 21 (frequent = 8) students had all three scores, and 31 (frequent = 20) had none of the three scores. To utilize all student ability data and to increase the accuracy of our measures, a composite student ability score was created from all three scores. Each score was standardized across the entire sample for each student using the mean for each score as the center of the distribution. After standardization of each score, all available z-scores were then added and divided by the number of scores available for each student. This process yielded an average ability score for all the students for which ability data was available.

Missing data. For the 31 students for whom ability data was not available, we employed multiple imputation through PROC MI (SAS 9.4) to address the problem of missing data (Shafer, 1997). Multiple imputation can be viewed as a way of using computer simulation to capture the uncertainty in the values that are missing from the data and incorporating that uncertainty into the analysis. PROC MI offers several approaches to imputation. In the present study, we used a regression based imputation method to estimate the values for missing data, while accounting for the variance of scores in the sample. Data from the subjects who have complete data is used to fit a regression model of the variable with missing values onto a set of predictors. This model is then applied to the cases with missing data to construct a set of plausible values for each case with missing data based on their standing on the observed predictors. The result of this procedure is a set of plausible values for each person with missing data that does not arbitrarily decrease the variance in the sample. In the present study, student ability was regressed on the cumulative midterm exam total ($r = .61$). PROC MI produced five datasets for the estimated values of missing data, resulting in each participant having five student ability estimates (participants who were not missing data thus had five duplicates of their student ability estimate). We ran analyses

by imputation number such that five models were estimated (one model for each imputed student ability estimate). The five models did not differ in terms of random effect and fixed effect estimation. Because results did not vary across the five runs, imputed values were averaged for each student who was missing data to yield one imputed student ability estimate. Typically, one would average the results of the five runs and describe the variability between and within runs, and compute significance tests taking into account the variability across runs, but the magnitude of missing data was sufficiently small that between run variability was negligible.

Testing frequency. Two types of testing schedules were manipulated between subjects: frequent (spaced testing) and standard (massed testing). Participants enrolled (not randomized) in two separate sections of the same undergraduate psychological methods course taught by the same professor at approximately the same time of day. Because random assignment of students to condition was not possible, a coin was flipped to determine which class would receive each treatment. The frequent class took eight 6-question midterm exams over the semester, spaced about 1 exam every two weeks. The standard class took two 24-question midterm exams, spaced about 1 every eight weeks. Students in both classes took the same comprehensive final exam. Over the course of the semester students in both classes took the same questions (a total of 112 test questions).

Quiz assignment. A portion of students in each class (total = 105; frequent = 57) took short practice quizzes over course material in the laboratory sections associated with the lecture course. The quizzes had no weight on the course grade for any students. A coin was flipped to determine which laboratory sections had the opportunity to take quizzes. The laboratory sections were taught by graduate student instructors who administered all quizzes and offered brief feedback on the correct answers shortly after quizzes were completed. Students in the quiz

condition took six 6-question quizzes (a total of 36 questions). Each quiz consisted of half MC and half SA questions. The content of the quizzes was considered closely related to material on the midterm exams and the final exam. Students who were enrolled in a laboratory section where practice quizzes were offered were coded as students who took quizzes.

Item predictors. Item factors were: 1) pre-test method and 2) post-test method.

Pre-test method. Half of the questions on the final exam had appeared (in MC or SA form) on a previous midterm exam. Repeated questions on the final exam either appeared in the same form as on the midterm exam (e.g., MC-MC and SA-SA) or in the opposite form (e.g., SA-MC and MC-SA). Final exam questions that were not previously taken on a midterm were called novel questions. Final exam questions that appeared on a midterm as an MC question were coded as pre-test method MC, questions that appeared on a midterm as an SA question were coded as pre-test method SA, and questions that did not appear on midterm exams were coded as pre-test method NT (not tested).

Post-test method. Half of all final exam items were MC and half were SA. Using pilot test data, final exam questions were equated for difficulty at < +/- 1% for MC items and < +/- 4% for SA items. For example, on the final exam all combinations of pre-test by post-test method MC questions were adjusted for item difficulty within the item type on the final exam (MC or SA). MC items on the final exam were coded as post-test MC questions and SA items were coded as post-test SA questions.

Analysis

A 1-parameter explanatory item response model (Rasch model) was used to estimate the effects of student ability, testing frequency, quiz assignment, pre-test method, and post-test method on the probability of correctly answer an item of average difficulty. The analysis was

performed only on responses to final exam questions. As our design features factors manipulated between-subjects and within-subjects, this EIRM used a generalized linear mixed model (GLMM) approach. PROC GLIMMIX (SAS 9.4) was used to model the probability of correctly answering an item given various levels of person and item predictors.

A series of models was used to compare the effects of different types of models. A fully unconditional model was estimated using no predictors to show the total variance in person and item random effects. The item predictors model included item predictors pre-test and post-test method and the person predictors model included person predictors student ability, testing frequency, and quiz assignment. The full model included all item and person predictors and 5 selected interactions: testing frequency by post-test method, testing frequency by pre-test method, testing frequency by quiz assignment, testing frequency by student ability, and a 3-way interaction of testing frequency by pre-test method by post-test method. Models were compared in terms of residual variance reduced in order to observe how well each model was predicting performance.

To aid in model interpretation, estimates of model parameters were used to estimate pass rates in order to display the effects of different factors in the design (e.g., the testing effect, the spacing effect, etc.). The pass rate or probability of correctly answering an average item for various conditions was computed by transforming the log odds provided by PROC GLIMMIX into a probability statistic.

Hypotheses

We hypothesized that students in the frequently tested class would have a higher probability of correctly answering items. We also predicted that students who received practice quizzes would perform better on the final exam. In addition to these effects, we believed that all students would perform better on items that had been pre-tested on the midterm exams. Finally, based on previous data we hypothesized that students in the frequent class would outperform students in the standard class on SA items.

Results

Tables 1 and 2 show results from two 2-way ANOVAs used to test for differences in SAT and GPA scores between the two levels of testing frequency and between students included in the analysis and those who were excluded. In these two ANOVAs the interaction term (testing frequency by exclusion) tests whether differential attrition occurred, that is, whether students who were excluded and students who were included in testing differed across the two conditions (frequent vs. standard) potentially biasing comparisons of the two conditions when those comparisons are based only on the included students. Table 1 shows that no statistical difference was found between levels of testing frequency between included and excluded students across conditions. Table 2 shows that no statistical difference was found between levels of testing frequency and exclusion on GPA. Tables 1 and 2 also show that the interaction of testing frequency and exclusion was not significant for both SAT and GPA variables.

Table 1

Two-way ANOVA Testing Mean Differences in SAT by Testing Frequency and Exclusion

Factor	Num. DF	Den. DF	Type III Sum of Squares	Mean Square	F	p > F
Testing Frequency	1	120	330.69	330.69	0.01	0.91
Exclusion	1	120	35537.76	35537.76	1.29	0.26
Testing Frequency*Exclusion	1	120	19732.85	19732.85	0.72	0.4

Note: Testing Frequency- whether a student was in the frequent or standard testing condition; Exclusion- whether data from a student was excluded from the analysis (see exclusion criteria in participants section).

Table 2

Two-way ANOVA Testing Mean Differences in GPA by Testing Frequency and Exclusion

Factor	Num. DF	Den. DF	Type III Sum of Squares	Mean Square	F	p > F
Testing Frequency	1	149	0.0017	0.0017	0	0.95
Exclusion	1	149	0.2681	0.2681	0.59	0.44
Testing Frequency*Exclusion	1	149	0.1395	0.1395	0.31	0.58

Note: Testing Frequency- whether a student was in the frequent or standard testing condition; Exclusion- whether data from a student was excluded from the analysis (see exclusion criteria in participants section).

Table 3

Percentage of Students by Number of Quizzes Taken

Quizzes Taken	%
0	0%
1	1%
2	5%
3	18%
4	25%
5	20%
6	31%

Note: Percentage represents the percentage of the students who took a given number of quizzes of those students who were given the opportunity to take quizzes.

As mentioned above, quiz assignment was measured using an intent-to-treat framework. To investigate the extent of participation in the quiz condition, the number of quizzes taken for each student was calculated. Students in the quizzing condition took a majority of the six possible quizzes (M = 4.52, SD = 1.29). Approximately 76% of students in the quizzing condition took four or more of the 6 possible quizzes. Table 3 shows the distribution of participation in the quiz condition.

The following results are reported by the amount of variance explained by each EIRM, the strength and direction of main effects and interactions, and the probability that a student with certain characteristics would answer an item correctly (pass rate). The pass rate can be obtained by calculating: $\frac{\exp(\text{expected}\log odds)}{1+\exp(expected\log odds)}$ (see De Boeck & Wilson, 2004b, p. 49; Kulesz, 2014). All pass rate statistics should be interpreted as the probability of passing an item of average difficulty for the corresponding conditions, while controlling for all other effects in the model.

Random Effects

Table 4 shows the residual variance in person abilities and in item difficulties, standard errors, and the variance reduction in both person abilities and item difficulties attributed to each predictor model. First a fully unconditional model was used to estimate both person and item variance when no predictors are used. The following models use the unconditional model as a baseline to test model fit and how predictors in each model influence person and item variance. The item predictors model, which included the pre-test and post-test method of final exam questions, reduced person variance by 5% and item variance by 28%. The person predictors model, including student ability, testing frequency, and quiz assignment, reduced person variance by 47% and item variance by 23%. To evaluate the effect of person by item

interactions, the full model included all main effects and selected interactions of person and item predictors. In total the full model included 5 main effects and 5 interactions. The interactions included in the full model were: testing frequency by student ability, testing frequency by pre-test method, testing frequency by post-test method, testing frequency by quiz assignment, and a three-way interaction between testing frequency, pre-test method, and post-test method. The full model reduced person variance by 48% and item variance by 28%, suggesting that person by item interactions reduced the person variance by an additional 1%, but had no effect on item variance, relative to the model that included only item predictors.

Table 4
Variance and Variance Reduction of Models

	Person Side			**Item Side**		
Estimated Model	*Variance*	*SE*	*Percent Variance Reduction*	*Variance*	*SE*	*Percent Variance Reduction*
Unconditional	1.259	0.15		1.499	0.25	
Item Predictors	1.192	0.14	0.05	1.086	0.17	0.28
Person Predictors	0.662	0.08	0.47	1.149	0.18	0.23
Full Model	0.660	0.08	0.48	1.086	0.17	0.28

Note: The unconditional model has no predictors. The Item Predictors model includes predictors: Pre-test and Post-test Method. The Person Predictors model includes predictors: Student Ability, Testing Frequency, and Quiz Assignment. The Full Model includes both person and item predictor main effects and interactions of person and item predictors. Percent variance reduction is relative to the unconditional model.

Table 5 shows the corresponding Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for each of the models in the first series. Both the AIC and BIC are used by researchers to assess how well a model fits the data, however, each of these statistics has a different method of penalizing for model complexity. Lower values of both the AIC and BIC indicate better model fit. As predicted by the percentage of variance reduced by each model, in general the models with smaller residual variance have lower (i.e., better) fit statistics. In model

series 1, the only major discrepancy in model fit between the AIC and BIC statistics is for the full model, where the AIC favors the person predictors model and the BIC favors the full model.

Table 5
Fit Statistics

Estimated Model	*AIC*	*BIC*
Unconditional	12710	12706
Item Predictors	12687	12667
Person Predictors	12583	12571
Full Model	12596	12552

Note: AIC- Akaike Information Criterion;
BIC- Bayesian Information Criterion.

Fixed Effects

Main Effects. Not surprisingly, student ability, $F(1, 11807) = 116.44$, $p < .0001$, and post-test method, $F(1, 11807) = 5.76$, $p = .02$, showed strong main effects. Testing frequency, $F(1, 11807) = 3.99$, $p = .05$, showed a moderate main effect. Quiz assignment, $F(1, 11807) = .32$, $p = .57$, and pre-test method, $F(2, 11807) = .05$, $p = .95$, failed to reach significance.

Figure 1 shows pass rate statistics by student ability level. Because student ability is a continuous measure, we have chosen arbitrary values of the ability measure at which to estimate the pass rate. Specifically, we estimate the pass rate at low, average, and high levels of ability, where low and high student ability were defined as being 1 standard deviation below and 1 standard deviation above the mean, respectively. Low ability students appear to be performing disproportionately worse than average and high ability students. High ability students have a pass rate of 84%, average ability students have a pass rate of 69%, and low ability students have a pass rate of 48%. Figure 2 demonstrates the main effect of testing frequency by pass rate, where

students in the frequent class have a pass rate of 72% and students in the standard class have a pass rate of 66%.

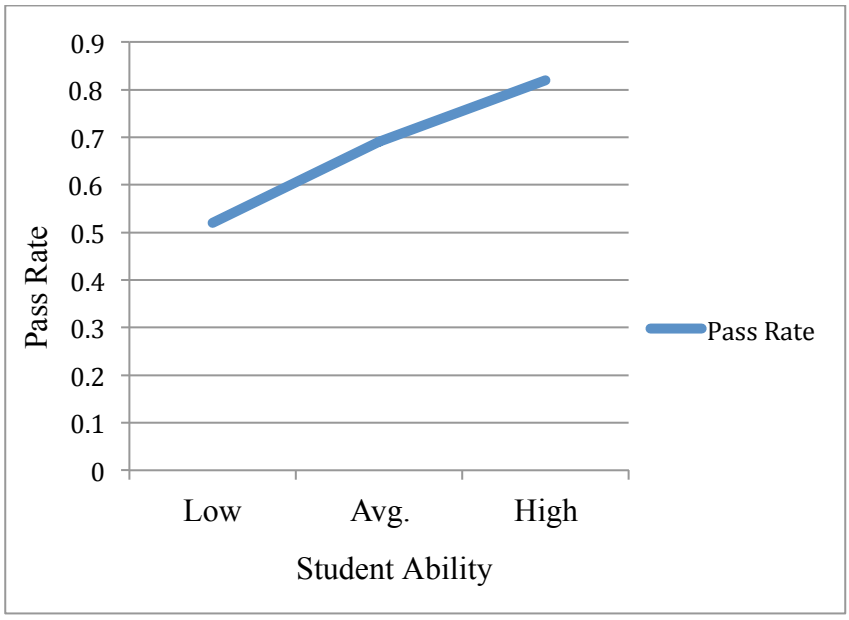

Figure 1. A line plot showing the pass rate for students by ability level. Low = low student ability (one SD below the mean); Avg. = average student ability; High = high student ability (one SD above the mean).

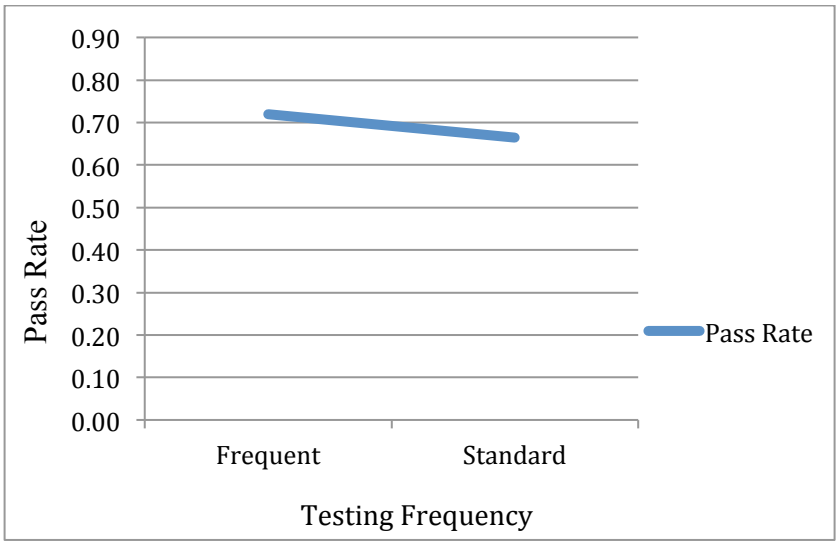

Figure 2. A line plot showing the pass rate by testing frequency. Frequent = the frequent class; Standard = the standard class.

Table 6

Main Effects- Pass Rate by Condition

Condition	Estimate	SE	Pass Rate
Student Ability			
High	1.682	0.171	0.84
Avg.	0.792	0.167	0.69
Low	-0.098	0.141	0.48
Testing Frequency			
Frequent	0.943	0.137	0.72
Standard	0.682	0.145	0.66
Quiz Assignment			
Quiz	0.849	0.144	0.70
No Quiz	0.776	0.137	0.68
Pre-Test Method			
MC	0.834	0.199	0.70
NT	0.839	0.196	0.70
SA	0.764	0.199	0.68
Post-Test Method			
MC	1.074	0.166	0.75
SA	0.551	0.165	0.63

Note: Pass rate is the calculated probability of a correct final exam response for various conditions. Estimate- log odds associated with each condition; High- student ability one SD above the mean; Avg.- mean student ability; Low- student ability one SD below the mean; Frequent- students in the frequent class; Standard- students in the standard class; Quiz- students in the quiz condition; No Quiz- students not in a quiz condition; MC- multiple choice; SA- short answer; NT- not tested.

Table 6 shows the pass rate associated with all main effects in the analysis. Large differences in pass rates between levels are shown for student ability and post-test method. A moderately large difference in pass rates was observed due to testing frequency factor. Quizzing had a negligible effect on average item pass rates. Specifically, students who took quizzes passed items of average difficulty at 70%, whereas students who did not take quizzes passed items of average difficulty at 68%. This difference was not statistically significant. Small differences

were found in the pass rates due to pre-test methods. These differences were also not statistically significant.

Interactions. Five interactions were tested in the analysis. No interactions reached significance at the $p = .05$ level. The interaction of testing frequency and post-test method, $F(1, 11807) = 3.66$, $p = .06$, was the strongest interaction. Quiz assignment did not moderate the effect of testing frequency, $F(2, 11807) = .38$, $p = .53$, and student ability did not moderate the effect of testing frequency, $F(1, 11807) = .04$, $p = .84$. The effect of testing frequency was also unaffected by pre-test method, $F(2, 11807) = .62$, $p = .54$. The 3-way interaction of class, pre-test method, and post-test method, $F(4, 11807) = .67$, $p = .61$, was also not statistically significant.

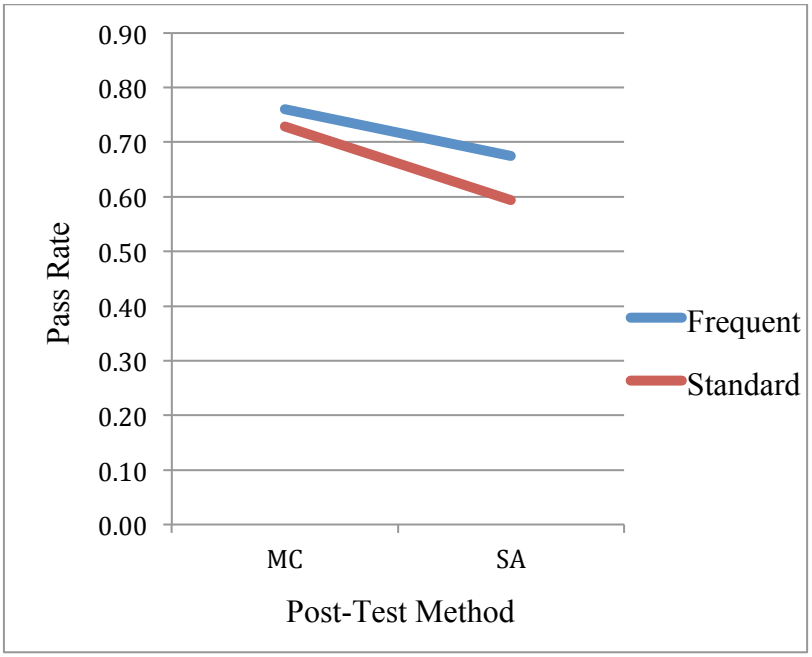

Figure 3. A line plot showing the marginally significant (p = .06) interaction of testing frequency and post-test method. Frequent = the frequent class; Standard = the standard class; MC= multiple choice question; SA = short answer question.

Figure 3 shows a line plot of the marginally significant interaction of testing frequency and post-test method. Table 7 shows the pass rates associated with the conditions for the 2-way

interactions. Disproportionately differences are observed between the standard and frequent class on SA items. The main effect of testing frequency is noticeable in all interactions in table 7, where the frequent class has higher pass rates for all comparable conditions. Table 8 shows the pass rates for the 3-way interaction of testing frequency, pre-test, and post-test method. Main effects of testing frequency and post-test method can be seen in the 3-way interaction. Higher pass rates are associated with being in the frequent class and taking an MC post-test item, however, no statistically significant interaction between all three variables occurred.

Table 7

Two-Way Interactions- Pass Rate by Condition

Condition	Estimate	SE	Pass Rate
Testing Frequency x Quiz Assignment			
Frequent x Quiz	1.020	0.158	0.74
Frequent x No Quiz	0.866	0.164	0.70
Standard x Quiz	0.678	0.166	0.65
Standard x No Quiz	0.685	0.183	0.65
Testing Frequency x Post-Test Method			
Frequent x MC	1.158	0.177	0.76
Frequent x SA	0.727	0.176	0.67
Standard x MC	0.989	0.184	0.73
Standard x SA	0.374	0.183	0.59
Testing Frequency x Pre-Test Method			
Frequent x MC	1.000	0.210	0.73
Frequent x NT	0.943	0.205	0.72
Frequent x SA	0.885	0.210	0.71
Standard x MC	0.668	0.217	0.66
Standard x NT	0.735	0.211	0.68
Standard x SA	0.642	0.216	0.66
Testing Frequency x Student Ability			
Frequent x High	1.674	0.161	0.84
Frequent x Avg.	0.943	0.137	0.72
Frequent x Low	0.209	0.163	0.55
Standard x High	1.386	0.182	0.80
Standard x Avg.	0.682	0.145	0.66
Standard x Low	-0.025	0.171	0.49

Note: Pass Rate is the calculated probability of a correct final exam response for various conditions. Estimate- log odds associated with each condition; MC- multiple choice; SA- short Answer; NT- not tested; Quiz- students in the quiz condition; No Quiz- students not in a quiz condition; Frequent- students in the frequent class; Standard- students in the standard class.

Table 8

Three-Way Interactions- Pass Rate by Condition

Condition	Estimate	SE	Pass Rate
Testing Frequency x Pre-Test x Post-Test			
Frequent x MC x MC	1.335	0.288	0.79
Frequent x MC x SA	0.665	0.285	0.66
Frequent x NT x MC	1.123	0.280	0.75
Frequent x NT x SA	0.762	0.278	0.68
Frequent x SA x MC	1.016	0.286	0.73
Frequent x SA x SA	0.754	0.285	0.68
Standard x MC x MC	1.013	0.293	0.73
Standard x MC x SA	0.322	0.291	0.58
Standard x NT x MC	0.980	0.285	0.73
Standard x NT x SA	0.490	0.283	0.62
Standard x SA x MC	0.974	0.293	0.73
Standard x SA x SA	0.311	0.290	0.58

Note: Pass Rate is the calculated probability of a correct final exam response for various conditions. Estimate- log odds associated with each condition; MC- multiple choice; SA- short Answer; NT- not tested; Frequent- students in the frequent class; Standard- students in the standard class.

Discussion

Random Effects

Model comparisons are useful to show how well each model is predicting responses and how well each model fits the data. The item predictor model (pre-test and post-test) method reduced person variance by 5% and item variance by 23%. Since the only item predictors we included in our model were the method (MC or SA) of pre-test and post-test, it is intuitive that this model does not explain the majority of item variance. It is possible that item factors such as content type (topic) and the distance from instruction (time) on the item content would further reduce variance, however, this hypothesis was not evaluated in this study.

The person predictor model (student ability, testing frequency, and quiz assignment) reduced person variance by 47% and item variance by 23%. These results indicate that the person predictor model is a good model for predicting differences across individuals in their final exam performance. The full model (including all main effects and 5 selected interactions) reduced variance across individuals by an additional 1% relative to the person predictor model, but had no effect on item variance relative to the item predictor model. These findings suggest that person/item interactions account for a very small percentage of the overall person and item variance. In other words, difficult items are difficult for all students and easy items are easy for all students, and the factors that affect item difficulty do so in a manner that is comparable for all students. Likewise, the student factors that affected student ability as measured by the final exam, did so in a manner that was similar for both MC and SA items and for novel and re-tested items.

It is interesting to note that both the item predictors model and the person predictors model reduced variance on the opposite side of the data. That is, the person predictor model reduced item variance and vice versa. Although this appears to suggest that item factors interact with person factors, variance reduction on the opposite side is likely due to the unbalanced nature of our experiment. Because we were not able to randomly assign students to conditions and had limited control over experiment dropout, the treatment conditions were unbalanced. As a result people and items were correlated to some degree. In the unconditional model, the intercept term reflects the average pass rate across items, and the item variance reflects the extent to which pass rates for individual items differ from this average. However, in the model with person predictors, the intercept becomes the average pass rate across items for individuals who are at the mean of all person predictors. This average pass rate may differ from the intercept in the unconditional

model and the variance in pass rates around this average value could increase relative to the variance around the unconditional average. Similarly, the mean ability conditional on item characteristics and the mean ability ignoring item characteristics are not necessarily the same, and variance across individuals around these means can differ. Hence, person variance can change even when only item predictors were added in the model. Put another way, when person information is missing from the model, knowing something about the items tells us something about average ability. Similarly, when information about items is missing from the model, knowing something about the people tells us something about average item difficulty. In short, item difficulty and person ability are correlated. This ambiguity resolves when both person and item predictors are included in the model.

Fixed Effects

Main effects. It is intuitive that student ability (GPA, SAT, and ACT scores) would have a substantial effect on performance on a final exam, whether measured as the number or percent correct, or the probability of passing a test item with certain characteristics. In addition to student ability, testing frequency was found to have a significant effect on final exam responses, favoring the frequently tested class and supporting our hypothesis. It is furthermore interesting that this effect remains significant after the effect of student ability (and all other effects in our model) was factored out. This finding suggests that frequent testing benefits college students and further shows the positive effects of spacing study and testing. Though the difference between the frequent and standard classes in this study resulted in only a 6% performance increase, this effect should not be considered a small one. The frequent class outperformed the standard class without any designed changes in instruction, preparation, or incentives. The fact that these results were

observed while 1) only changing the frequency and amount of material on each exam and 2) controlling for student ability, shows the magnitude of the spacing effect and frequent evaluation.

The negligible effects due to quizzing was surprising. This result may illustrate some of the differences between frequent testing and quizzing as we defined them. Firstly, students in the quizzing condition were given no grade incentives to prepare and take quizzes with integrity. Students may have had little motivation to perform well, other than to use the quizzes to their own benefit for future graded exams. A large body of research shows that students differ in their motivation to succeed in school. Some students are motivated to learn material for their own good (mastery goal motivation), while others are motivated to achieve a level of academic success to improve their social status, or to avoid negative repercussions (Wolters, Denton, York & Francis, 2013). It follows that students who differ in educational motivation might perform and attend differently on quizzes that have no immediate baring on their course grade. Unfortunately, students' motivation orientation was not measured in this study.

Second, as demonstrated by Wooldridge, Bugg, McDaniel, and Lui (2013), the relation between quiz questions and subsequent exam questions is critically important. Although we believed the quiz questions in our study to be closely related (though not exactly repeated items) to the exam questions, it is possible that our materials may not have successfully met the constraints of the testing effect. It remains interesting that students in the quizzing condition did not perform better, despite having the opportunity to take and receive feedback on 36 more questions than students who were not in the quizzing condition. Although the number of

questions was confounded with the quiz condition, we suspected that answering more questions might mediate the effect of quizzing.

Post-test method exhibited a strong main effect as expected (in other words MC questions were easier than SA questions). However, pre-test method did not predict performance on the final exam. Given that the pre-test was where the level of item novelty was tested, it is surprising that pre-test method was not a strong predictor of performance. The null effect of pre-test method contradicts the literature on the testing effect, where we predicted that students would perform better on final exam questions that had been answered previously on a mid-term exam. This finding may highlight the fact that the testing effect is most reliable under strictly controlled conditions. The testing effect has been shown in many experiments using conditions similar to the conditions in this experiment. The benefits of testing have been shown in middle school classrooms (e.g., McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013), using authentic educational materials in the laboratory (e.g., Wooldridge, Bugg, McDaniel, & Lui, 2013), and in college level psychology courses (e.g., Pennebaker, Gosling, & Ferrell, 2013). Although testing effects have been shown in similar conditions, the factors in this experiment that were not controlled may have played a large role.

The lack of a pre-test method effect could have been influenced by the fact that our data was analyzed under an intent-to-treat framework, where we recognize that all students had the opportunity to take midterm exams, but not require that students completed all exams to be included in the study. If a student failed to take a midterm exam, final exam items coming from that midterm exam would then be effectively novel to that student, and therefore not consistent with the way items were coded for novelty. Missing exams was most likely to occur in the

frequent class, where the students took more exams and therefore had more opportunity to miss exams. In the standard class, students were much less likely to miss an exam, as their (less frequent) longer exams constituted a larger percentage of their overall grade relative to the shorter exams in the frequent class.

Interactions. The trend of the interaction of testing frequency and post-test method confirms our predictions about the effect of test spacing. As shown in figure 3 and table 6, the frequent and standard classes did not differ significantly on questions post-tested as MC (approximately a 3% difference). However, the two classes greatly differed on post-test SA questions (approximately a 7% difference). This result suggests that spacing and frequent testing may benefit recall memory more than recognition. This finding is consistent with results from other studies where SA questions lead to a greater likelihood of being affected by training methods (e.g., Butler & Roediger, 2007; Kang, McDermott, & Roediger, 2007). It is intuitive that the more difficult a task is, the more that effective learning strategies make a difference in performance. The fact that frequent testing had little effect on MC question performance in our study implies that frequent testing may not yield large learning gains unless the response method is difficult (e.g., recall memory). This finding has serious implications for many disciplines and industries where recognition memory is not often used (e.g., performing arts and medical practice).

Other than the marginally significant interaction of testing frequency and post-test method, no other interactions approached significance. The lack of an interaction between testing frequency and pre-test method is surprising, as some research has found that frequently tested students perform better on novel material (Foss & Pirozzolo, 2014). However, this finding may

again be a product of the "intent to treat" analysis, and different results might obtain if we examined the impact of the treatment on the treated (Angrist, Imbens, & Rubin, 1999). The null effect of an interaction between testing frequency and quiz assignment suggests that taking practice quizzes had the same effect on both classes. This finding may highlight the limited role of optional practice quizzes and their hypothesized motivational impact. Finally, the lack of an interaction between testing frequency and student ability suggests that the benefit of frequent testing operates the same for all levels of students. The consistent finding in this study that student ability, quiz assignment, and item predictors do not significantly interact with testing frequency provides further evidence that the frequent testing advantage generalizes across many conditions. An understanding of the roles of all the factors in the spacing and testing effects is critical to effectively translating research into practice.

Implications

Most notably, the results of the present study provide evidence that college student performance can be improved by applying research in cognitive psychology. Evidence from this study shows that the effects of frequent testing are likely to generalize to all levels of students and across manipulations at the item level. The methods in the study provided a unique examination of the effectiveness of longstanding cognitive paradigms using an advanced analysis method that allowed subjects to vary in ability. In effect, the present study considered individual differences, which have been largely ignored by the literature on the testing and spacing effects. Our quasi-experimental design offers an appropriate balance between experimental control through statistical modeling and ecological validity.

We provide an estimate of the effect size of frequent testing, where instructors can expect to see approximately a 6% increase in performance using similar methods to those described in this study. The implications and scope of this study could be quite large. It is not unrealistic to believe that if frequent testing methods were used throughout college courses, and if each course netted students a 6% increase in performance, that substantial gains could be made relative to standard procedures over four years of instruction. It is also possible that by using frequent testing, more students would graduate, leading to a more prepared workforce and informed community.

Limitations

The sample size of this study presents a potential limitation. Typically EIRMs need many hundreds of subjects for confidence in parameter estimates. Another potential limitation is the number of official course withdrawals in the frequent class. As stated in the method section, 25 students (17%) in the frequent class officially withdrew from the course compared to only 4 students (4%) in the standard class. Although results from two ANOVAs (tables 1 and 2) provide evidence that differential attrition did not occur, the disproportionate exclusion rate could have influenced the results. We speculate that more feedback opportunities in the frequent class aided low performing students' judgments about their own performance, and influenced the higher dropout rate in the frequent class. For example, students in the frequent class had received feedback (score and verbal) from 6 short exams before the official withdrawal deadline, whereas the standard class only received feedback from 1 long exam. Although this difference only represents feedback from 12 more questions in the frequent class, more frequent testing may

allow students to more accurately detect a pattern of poor performance and lead to a higher dropout rate.

Another limitation of our design is our inability to randomly assign students to conditions. Although using covariates such as GPA, SAT, and ACT score helped control for the effects of self-selection into classes, random assignment would allow for a more balanced design and a stronger basis for causal inferences about study effects. We also were unable to address influential factors such as student attendance, study time, and method of study. Similarly, a potential limitation of this study is that frequency of testing was confounded with course section, and our study lacks replicates at the section level. This confounding and lack of replicates leads to two problems. First, it remains possible that the estimated effects reflect the influence of other factors related to these specific course sections that are unrelated to the difference between sections in testing frequency. Although neither section suffered from a specific disruptive student, or deviated in noticeable ways from other sections taught by this instructor as evidenced by course evaluations, or anecdotal observations about student attendance and engagement, the possibility of unobserved differences between sections cannot be ruled out. Secondly, the assignment of sections to conditions raises the issue of dependence across students that cannot be accounted for in the statistical model, because of the lack of replicates at the section level. Research in education has a long history of concern over ignoring the effects of clustering on standard errors in statistical models, and yet most applications of item response models in education ignore the clustering of students in test design and analysis. Although in this study the same professor taught both sections at similar times of day, it is possible that students in the same section are more similar to one another than students in different sections are to one another resulting in non-independence across observations, which would generally result in

underestimation of standard errors. Outside of taking great care to treat both classes similarly, we were unable to address this issue in our experimental design, or in the analysis.

Future Directions

Future research on the conditions that lead to better learning in the college classroom should extend to include a wide range of disciplines outside of psychology. Further examinations of these effects across various subjects will increase our knowledge by understanding the limits of learning methods. Researchers should continue to translate results found in laboratories to applied settings where learning improvements are most meaningful. Further research in this area should also explore the long-term effects of the testing and spacing effects. As reported by several studies, the true benefits of superior study and practice methods might be most observable on a long-term scale (Karpicke & Roediger, 2006; Roediger & Karpicke, 2006b).

Conclusion

This paper provides evidence from a quasi-experiment with high ecological validity that frequent testing can significantly improve college student performance. Specifically, frequent testing was found to have the largest benefit on short answer questions. Taking practice quizzes had little effect on performance on the final exam. The EIRM approach used in this study allowed for us to examine the effects of testing and testing method for novel and familiar items while controlling for other factors that might affect student ability of item difficulty. We conclude that findings from laboratory studies generalize to the college setting and can be used to improve adult learning in a real, semester-long college course. We further conclude that the

benefits of frequent testing can be generalized across levels of student ability and other experimental manipulations, such as item novelty.

References

Allen, M. J. & Yen, W. M. (2002). Strong true-score theories and latent-trait models. In Allen & Yen, *Introduction to Measurement Theory,* Long Grove: Waveland Press Inc., pp. 239-273.

Angrist, J., Imbens, G. and Rubin, D. 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444–55.

Arnold, K. M., & McDermott, K. B. (2012). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *3*(39), 940-945.

Bacon, F. (2000). *Novum organum* (L. Jardine & M. Silverthorne, Trans.). Cambridge, England: Cambridge University Press. (Original work published 1620)

Bjork, E. L., & Bjork, R. A. (2010). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Gernsbacher, M. A., Pew, R. W., & Pomerantz, J. R. (eds.) *Psychology and the real world: Essays illustrating fundamental contributions to society.* New York: Worth Publishers.

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 & A. Shimamura (eds.), *Metacognition: Knowing about knowing,* (pp. 185-

205). Cambridge, MA: MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Learning and Behavior*, *9*, 567-572.

Bjork, R. A., & Linn, M. C. (2006, March). The Science of Learning and the Learning of Science: Introducing Desirable Difficulties. *American Psychological Society Observer*, *19*, 29-39.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a sim- ulated classroom setting. *European Journal of Cognitive Psychology*, 19, 514–527.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354-380.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

De Boeck, P., & Wilson, M. (2004a). A framework for item response models. In De Boeck, P., & Wilson, M. (eds.), *Explanatory Item Response Models: A generalized linear and nonlinear approach,* New York: Springer, pp. 3-41.

De Boeck, P., & Wilson, M. (2004b). Descriptive and explanatory item response models. In De Boeck, P., & Wilson, M. (eds.), *Explanatory Item Response Models: A generalized linear and nonlinear approach,* New York: Springer, pp. 43-74.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychology*, *43*, 627-634.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1,* 309–330. doi:10.1007/ BF01320097

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*(1): 4-58.

Ebbinghaus, H. (1885). Memory (translated by H. A. Ruger and C. E. Bussenius). New York, Teachers College, 1913. Paperback ed., New York, Dover, 1964.

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

Foss, D. J. & Pirozzolo, J. W. (2014, November). Improving undergraduate performance: Testing the "testing effect" in a college methods course. *Poster presented at the Annual Meeting of the Psychonomic Society*. Long Beach, CA.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

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. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163.

Karpicke, J. D., & Roediger, H. L., III (2006). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151-162.

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

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the enemy of induction? *Psychological Science*, *19*, 585-592.

Kulesz, P. A., (2014). The effects of reader characteristics, text features, and comprehension processes on reading comprehension (unpublished doctoral dissertation). University of Houston, Houston, Tx.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In Gruneberg, M. M., Morris, P. E., & Sykes, R. N. (eds.), *Practical Aspects of Memory*, New York, Academic Press, pp. 625-632.

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition, 3*, 171-176.

Little, J. (2010). Optimizing multiple-choice tests as learning events (Unpublished doctoral dissertation). University of California- Los Angeles, Los Angeles.

McDaniel, M. A., & Fisher, R. P. (1991). Test and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192-201.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III. (2013). Quizzing in middle school science: Successful transfer performance on classroom

exams. *Applied Cognitive Psychology*, 27, 360–372.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology*: Applied, *20*, 3–21.

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE 8*(11): e79774.

Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2004). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, *30*, 96-116.

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

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

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241-249.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289-318.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Szpunar, K. K., Kahn, N. Y., & Schacter, D. L. (2013). Interpolated tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, USA, 110,* 6313-6317.

Toppino, T. C., & Cohen, M. S. (2009). The Testing Effect and the Retention Interval: Questions and Answers. *Experimental Psychology*, *56*(4): 252-257.

Toppino, T. C., & DiGeorge, W. (1984). The spacing effect in free recall emerges with development. *Memory and Cognition*, *12*, 118-122.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: The effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465-478.

Wichawut, C. (1972). Encoding variability and the effect of spacing of repetitions in continuous recognition memory. (Air Force Office of Scientific Research: AD- 754 960). Springfield: National Technical Information Service.

Wiklund-Hornqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10-16.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Lui, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*, 214-221.