Georgia Educational Researcher

Volume 14 | Issue 1 Article 4

7-2-2017

An Examination of the Testing and Spacing Effects in a Middle Grades Social Studies Classroom

Mary C. Liming

Joshua Cuevas

An Examination of the Testing and Spacing Effects in a Middle Grades Social Studies Classroom

Mary C. Liming
University of North Georgia

Joshua Cuevas
University of North Georgia

Abstract: This study investigates the relation between review spacing and question format on student retention. Participants in an 8th grade Georgia Studies class reviewed previously learned material either in one sitting (massed review) or in multiple sessions (spaced review). Following the review, each participant answered questions either in multiple choice or short answer format. Subsequent to answering the questions, all students received feedback. One week following the completion of the reviews, students were given a post test. One month after the post test, students were given a final test. Pre-, post- and final tests were identical and no treatment occurred between the posttest and the final test. Correlational analyses of review spacing and question format suggest that spaced review is positively related to success on the posttest. There was no such finding related to massed review or question format on the posttest. Additionally, neither review spacing or question format had any correlational effect on the final test. Results suggest that spaced review is beneficial, but that the benefit is lost over time.

Keywords: spacing effect, testing effect, short answer, multiple choice, middle grades

An Examination of the Testing and Spacing Effects in a Middle Grades Social Studies Classroom

Introduction

Teachers, researchers, and students have long been interested in memory and how to improve it. The advent of end-of-semester or end-of-course tests necessitate being able to recall lessons learned months previously. In history classes, it is usually not enough to remember broad trends – the movement from being a slave society to one pursuing equal rights for all citizens, for example – but rather students must be able to recall specifics such as dates or people who they may have been introduced to once and not seen again other than on tests. Where a math class makes frequent use of addition and multiplication facts, a history class may only touch on the leader of the Populist movement during one lesson, yet the student is expected to recall details of it months later on an end-of-course exam.

While there are many tricks, tips, and techniques for improving memory, this study will focus on two; the testing effect and the spacing effect. The spacing effect is the idea that "long-term memory is promoted when learning events are distributed in time rather than massed in immediate succession" (Gluckman, Vlach, & Sandhoffer, 2014). Researchers have been finding advantages in spacing lessons for years; however teachers may find it difficult to implement. Textbooks generally mass information by topic, and history by its very nature progresses linearly. The testing effect refers to the concept that testing on material prior to a final exam increases later retention more than simply studying the material (McDaniel, Roediger, and McDermott, 2007). Can the tests and quizzes which are inevitably part of every class be used to improve recall and boost improvement on exams held at a date removed from the original

DOI: 10.20429/ger.2017.140104

lesson? If so, this may help reduce anxiety over the number of tests involved in the school setting.

Testing Effect

Studies have indicated for years that short answer (SA) questions promote better retention than do multiple choice (MC) questions. In 1980, Gay compared MC quizzing to SA quizzing and their effects on final test scores with two small groups of undergraduate students (*N*=28). Students observed lectures and then were given either a MC quiz or a SA quiz covering the concepts from the lectures. A final exam of mixed format was given at the end of the term. Gay found that students performed equally on the MC section of the final exam but that the students who had taken the SA quiz did significantly better on the SA portion of the final (83.05% versus 63.07%). While this study seems to support the idea that SA testing results in greater retention than MC testing, it used a very small sample of older students. This brings into question its suitability for middle school students in general.

There is also some question as to the advantage of SA over MC. Further studies seem to indicate that the different types of questions measure different areas of memory and are affected differently by active learning techniques (Ozuru, Briner, Kurby, & McNamara, 2013). In this more recent study undergraduate students were asked to provide written explanations of passages they had just read. Positive examples were provided for the written explanations. Following the readings, students were given quizzes which contained SA and MC comprehension questions, followed by SA and MC prior knowledge questions. Upon scoring, it was determined that students who provided strong written explanations also performed strongly on the SA questions, while students with previous knowledge performed more strongly on MC questions, suggesting that the active learning involved with the written explanations provided a different foundation for learning than the scaffolding provided by prior knowledge. Written explanations involve the

active generation of relevant ideas and add to comprehension. MC questions reflect topic-specific knowledge that is not necessarily generalizable into short answers.

Ozuru et al. (2013) also found that even after answering SA questions correctly, students did not necessarily answer the corresponding MC question correctly. They proposed that without feedback to know if their SA answer was correct, students second guessed themselves into giving a different answer when presented with the MC version. Feedback is necessary to retaining accurate information. Other studies have scrutinized the necessity of feedback as well. Studies done by Kang, McDermott and Roediger (2007) indicate that immediate feedback given on an initial SA test leads to stronger long-term retention even if the final test is in MC format. However, if no feedback is given, performance is stronger if the initial quiz is MC. Extrapolating from Kang et al.'s conclusions, McDaniel, Roediger, and McDermott (2007) posit that the form of the initial test, combined with the presence or absence of feedback, will influence long term retention. With feedback after a SA initial test, long term retention is better supported through SA tests than MC tests.

McDermott, Agarwal, D'Antonio, Roediger and McDaniel (2014) performed an experiment with 141 7th grade students. Students took a pretest, were taught the lesson, took a post test, then a review quiz which happened the day before the final unit exam. The quizzes were administered via a projection screen, and immediate feedback was given after students entered their MC answers. After students answered SA questions, an ideal answer was displayed and read aloud. The final unit exam, which was a traditional paper and pencil test, included a mixture of SA and MC questions and included material which had not been included on the quizzes. Finally, a delayed exam was administered via the projection screen. All students improved as they progressed through the quizzes, which is to be expected since the pre-test was

DOI: 10.20429/ger.2017.140104

before the lesson. For SA questions on the final exams, having quizzing with feedback increased performance, even if the quiz was MC. These studies show that feedback is necessary to reinforce or correct understanding for any format quiz. It is most beneficial when initial testing is an SA format.

Since repetition of material is a widely accepted teaching strategy, it could be assumed that the benefits seen through retesting could simply be the result of multiple exposures to the material. This, however, does not appear to be the case. In another experiment described by McDermott et al. (2014), certain concepts were taught, but not quizzed. Instead, students "restudied" concepts by seeing the question stem and an ideal answer projected on the screen during the quiz. Quizzed information was retained better than restudied information. McDaniel et al. (2007) also examined restudying by offering their students SA tests, MC tests, or a restudy of the facts. Again, SA tests led to the best results on the final, followed by MC and restudying. SA quizzing produced the longest lasting retention, but even MC quizzing was better than restudying. In a third study they describe, groups of students took quizzes throughout a unit which combined MC, SA, and reread material. At the end of the unit, a MC unit exam was given. Again, a testing effect was evident, with student performance being stronger on material previously presented as MC or SA than performance on reread material. It should be mentioned that material which was re-read was retained better than material which was not re-examined at all, but quizzing produced stronger results. SA quizzing had the most benefit.

Roediger, Agarwal, McDaniel, and McDermott (2011) studied 143 6th grade social studies students. Students were more successful on items which had been MC on the pretest when taking the posttest as compared to items which were not on the pretest but were reread multiple times. In another portion of their study, Roediger et al. studied 132 6th grade social

studies students who were pretested (MC), taught the lesson, and encouraged to do online self-quizzing and games before the lesson exam and the final end of semester exam. The online program only included the material which had been included in the original pretest, but did not contain all of the material which was included in the final chapter tests or the end of semester exam. Overall, performance was stronger on material which had been included in the pretest and self-testing than on material which was included in the lesson but required individual study. These benefits crossed the format lines between MC, SA, and "free recall" despite the fact that the pretest was all MC. The conclusion of Roediger et al.'s three studies is that there appears to be a positive correlation between items which were pretested and items which were retained for both the posttest and the review test, indicating that quizzing and multiple testing is more beneficial than re-reading or individual review. Thus, the testing effect is not simply a result of repetition.

Finally, a look at the wording of questions. Do students perform better on retested material simply because they have seen the same questions before or are they able to produce results when the concepts are the same but the question wording has changed? McDermott, et al. (2014) examined this by changing the question wording between quizzes and final tests. Sixty 7th grade students took quizzes with immediate feedback on MC and SA questions and restudy material. Final paper and pencil tests were given 2 to 3 days after the review quiz. Students performed better on MC questions on the quiz but had the largest performance increase on SA questions. On the final exam, questions had different wording than had been presented on the quizzes, resulting in testing results which were similar to previous studies where wording was not changed. The best performance was on questions which had been SA on quizzes, followed

DOI: 10.20429/ger.2017.140104

by MC, then restudy, and finally those with no review. Changing the wording did not affect the testing effect.

Through the progression of research, we can see that the testing effect exists and is generalizable across age groups and between lab and classroom settings. Results seem to indicate that SA quizzes with feedback provide the best results on final exams regardless of the format of that final exam. However, there are indications that SA and MC questions simply test different types of memory, with SA questions building on active learning and MC questions tapping into previous knowledge. Additionally, question wording and format can change between quizzes and final tests without a decline in results, indicating that the benefits are not just from repetition.

The Spacing Effect

Rohrer and Pashler (2010) recently examined three different methods: testing, spacing, and interleaving. Rohrer and Pashler define the *spacing effect* as distributing study over multiple sessions rather than consolidating study into a single session in order to increase performance on a delayed final test. In regard to this, they cite multiple studies that point to the seemingly universal benefits of spacing instruction or review instead of massing it. They argue for research investigating the generalizability of spacing to more complex tasks since early experiments focused on vocabulary lists and were not done in classrooms. Rohrer and Pashler conclude that spacing invariably produces better results than massing instruction and that it is generalizable to other subjects and classroom settings of many ages.

An example of spacing utilized in a classroom is demonstrated by Sobel, Cepeda and Kapler (2010). They exposed a small sample of 5th graders to vocabulary words. Students were shown the words with their definitions and an example of the word being used in a sentence

while the teacher read all this out loud to them. They were asked to write the definition of the words from memory, were given and asked to re-read the correct definition, then re-wrote the definition from memory again and used the word in a sentence. Half of the class repeated the process immediately. The other half repeated the process one week later. Five weeks later, they were tested by being asked to write down the definitions from memory. The students whose learning was spaced had substantially higher retention levels (177% higher) than the massed group. Sobel, et al. conclude that teachers need to include review or time to relearn in their lesson plans several days after introduction of concepts. They also suggest using mini assessments to further space learning. They do express concern that spacing may not be appropriate for math and science material.

Kornell and Bjork (2008) have also shown concern about generalizing spacing. Specifically, they worried that spacing would inhibit inductive learning, or learning by observation. They took 120 undergraduates and exposed them to paintings by 12 artists displayed on a computer. The students received instruction as to characteristics of each artist's style and then were shown paintings. Six of the artist's works were massed – all of the sample paintings for a given artist were displayed consecutively before moving on to the next artist – and six of the artists' works were shown in mixed order (spaced). Students were tested by seeing a previously unseen painting by one of the artists and asked to name the artist. Students were able to correctly identify paintings by artists in the spaced group significantly more often than those in the massed group. Kornell and Bjork conducted another, smaller, experiment which did not require the students to remember names. After similar instruction with new paintings, students simply had to state if the painting was by a familiar artist or unfamiliar artist. They also had an "I don't know" option. Again, those in the spaced condition performed better than those

DOI: 10.20429/ger.2017.140104

in the massed condition. These were two very small studies with older students. However, they do show that spacing can be generalized into a new subject and that it does not conflict with inductive learning.

Gluckman, Vlach, and Sandhofer (2014) brought spacing out of the laboratory and into the classroom. They used younger students and did not focus on vocabulary, thereby examining the generalization of spacing beyond what previous studies had been able to conclude. They used a small sample (N=24) of 1st and 2nd graders. The children were taught about four biomes and food chains. One set ("massed") received all of the information on one day. One set ("clumped") received two lessons on two consecutive days. The final set ("spaced") received one lesson a day for four days. The final test was given one week after the final lesson. Students were tested on memory ("What is a biome?"), simple generalizations (picture of an animal, child asked to pick what that animal would eat from supplied pictures), and complex generalizations ("What would happen to animal *a* if all of animal *b* were removed?" The children were asked to indicate with arrows or an equal sign what would happen to animal *a*). The final test was on a biome the child did not receive instruction on, forcing the student to generalize knowledge to a new set of information.

As expected, there was a significant improvement for all students between pre and posttests (Gluckman, Vlach, & Sandhofer, 2014). The "spaced" group had significantly greater retention. The same was true for simple generalizations – "spaced" did better than "clumped" or "massed." For complex generalizations, there was not a significant increase in performance between pre- and post-test for the massed group, while the spaced group showed significantly greater learning than the massed group.

If spacing helps promote retention, what is the appropriate spacing? Pashler, Rohrer, Cepeda, and Carpenter (2007) attempted to answer that. Based on multiple studies, by comparing results and spacing timing, they developed a formula. If the time between study sessions is the interstudy interval ("ISI") and the time between the second study session and the final test is the retention interval ("RI"), then they concluded the ISI should be 10%-20% of the RI. However, they find that if the RI is 50 weeks, the ISI should be 3 weeks which does not fit their equation. They also claim that longer than ideal spacing is less harmful to long term memory than shorter spacing.

Cepeda, Vul, Rohrer, Wixted and Pashler (2008) also attempted to define the optimal spacing. Their highly complex study examined 26 combinations of ISI and RI and concluded that spacing effects are nonmonotonic (as did Pashler, et al. 2007). They explained that as the retention interval increases, so must the ISI, but that when ISI gets too long, the rate of forgetting increases quickly. Knowing the ISI depends on knowing the RI. Knowing how long the gap between studying should be depends on how long it is necessary to remember the material. The price of too long a study gap is lower than the price of too short a study gap. They state that ISI should be about 15%-20% of RI, but that the relationship is not linear and therefore difficult to definitively identify.

Carpenter, Pashler, and Cepeda's study (2009) combined the testing effect with the spacing effect. In this study, they took 75 8th grade US history students and assigned them by class to various groups. Groups had MC and SA test questions which were reviewed through feedback, restudied, or not reviewed. One group had review one week after the course of study, one had a review 16 weeks after, and a final test for each group took place 36 weeks after their review. The group which had the immediate review scored higher on the study questions than

DOI: 10.20429/ger.2017.140104

the delayed review. However, the final test items, which had been tested/reviewed, were retained significantly better than those of study alone or no review. The delayed review group did better overall, but the interaction was not significant. The study seems to indicate that testing with immediate feedback aids retention more than study alone and that delaying the review – spacing it out from the course work – also aids retention. Since the final was so far removed from the course work, overall retention levels were very low. The RI spacing would make this very difficult to replicate in a traditional classroom. However, this study successfully combined both the testing effect and the spacing effect, showing that material that is tested with immediate feedback and reviewed at an interval spaced between the original introduction and the final exam is better retained than other combinations.

Previous studies have established the existence of both a testing effect and a spacing effect. Researchers have examined test format as well as the necessity of feedback. They have considered the possibility that the effect is simply a benefit of repetition and shown that repeated testing with feedback strengthens retention across concepts and does not rely on repetitive wording between tests. Studies of the spacing effect have demonstrated that spreading study time across multiple sessions leads to higher retention than one massed session and have attempted to determine the correct amount of space between initial exposure, restudy, and testing for maximum benefit. Many of these studies, however, used small sample sizes, were done with older students, or done in a laboratory setting, leading to questions of generalizability to a middle school classroom environment.

Research Question

This current study is designed to examine the generalizability of the testing effect to a middle school social studies classroom. A quick count indicates that there are about 50 historic

figures that a student in an 8th grade Georgia Studies Class needs to remember (GeorgiaStandards.org). Does frequent quizzing help students remember? If the quiz is accompanied with feedback, it might. Another idea is that writing helps retention by forcing the student to recall facts rather than just recognize them. If that is the case, short answer quizzes may be more beneficial than multiple choice tests. This study will compare two instructional techniques. Does frequent quizzing help with retention, and are short answer questions more helpful than multiple choice?

The current study will also examine the spacing effect in a middle school classroom. The idea of "boot camps," "intensive instruction," and "immersion" learning force a large amount of information on students in a short amount of time. But as efficient as that is with time, how does it affect memory? Previous studies seem to show that spacing learning and review help with the memory retrieval process more than processing all of the information at once. This study will compare short review sessions spaced over several days to one long review session. Do students who "space" their review out do better on assessments than students who "mass" their review? And which format lends itself to longer retention?

Method

Participants

This study was conducted at a suburban middle school in Fulton County, Georgia. Fulton County is a sprawling county which includes the city of Atlanta. The population in the county grew approximately 8.2% between 2010 and 2014 (United States Census Bureau, 2014). In 2014, 46.7% percent of the population identified as White, with the primary minorities being "Black or African American alone" at roughly 44.3%, "Hispanic or Latino" at 7.6% and "Asian alone" at 6.7% ("Hispanics" can be any race and are therefore reported multiple times). The

DOI: 10.20429/ger.2017.140104

median household income in the county between 2009 and 2013 was $56,857, somewhat above the state median of $49, 179. However, during this time, 17.6% of the population of Fulton County lived below the poverty line.

According to USASchoolInfo.com, total enrollment at this middle school was 1332 at the time of the study. Its population was 58.5% White, with the primary minorities being Asian at 19.1% and Black at 13.7%. The rest of the population identified as Hispanic (4.8%), Multiple Race, and Pacific Islander. According to schooldigger.com, 9% of students were eligible for free or reduced lunch in 2014, a drop from 9.9% in 2013.

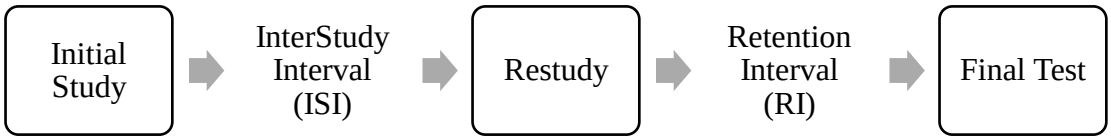
According to Schooldigger.com, the school is ranked 8th of all middle schools in the state based on average standard test score across grades and is 7th out of 9 middle schools in the district based on that same scoring system. Again, according to Schooldigger.com, 8th grade 2014 CRCT passing rates ranged from 100% for reading to 96.6% for science. Social studies passing rates were 97.6%. CRCT pass rates for the school were higher than the district and state test scores in all categories. The ranking of social studies pass rates as the second lowest category is mirrored throughout the county.

The participants in the study were four preformed classes totaling 76 on-level 8th-grade social studies students. In Georgia, 8th grade social studies is a study of Georgia geography, history, economy, and government referred to as Georgia Studies. The body of participants was approximately 33% male, 66% female. Two of them had 504 accommodations involving providing extra time if necessary. Each pre-formed class of students was randomly selected for one of the four conditions of the study. Two of the classes received their review material in a "massed" format and half of them in a "spaced" format. Within each format, one class received multiple choice (MC) questions with their review and one class received short answer (SA)

questions. Thus, the four research groups were spaced/multiple choice (n = 20), spaced/short answer (n = 22), massed/multiple choice (n = 15) and massed/short answer (n = 17).

Materials

As this study was intended to collect information regarding students' retention and review of first semester lessons, it focused on the four units covered before winter break: The Prehistoric Period - which includes general geography as well as early inhabitants of Georgia, Exploration and Colonization, Statehood, and the structure of state government (Georgia Performance Standards for Social Studies). In relation to the standard wording in regard to the spacing effect, students already had the "initial study" when the subjects were introduced. The interstudy interval (ISI) was the period between when the lessons were finished and this study. The readings and quizzes represent the "Restudy" portion of the flow chart below. This experiment represented the restudy, retention interval ("RI"), and final test portion of a spacing effect study.



Students were presented with four review modules created to align with the curriculum map presented on the Georgia Standards website. Each review module was designed to include a reading followed by a quiz and took approximately 10 minutes to complete. Text came from the Teacher's Notes section of the Georgia Standards website and was simplified and streamlined so as to take the typical 8th grade student approximately 5 minutes to read. Here is a sample of the text:

> The state of Georgia is divided into five geographic regions. The Blue Ridge Region is in the northeastern portion of the state. Dahlonega, the site of

DOI: 10.20429/ger.2017.140104

America's first Gold Rush is in this region. The Blue Ridge receives the most rain in the state and is the starting point of most of Georgia's rivers.

The Valley and Ridge Region was traditionally a mining region, with the valleys being used for farming. This region was a major battle ground during the Civil War and is a major transportation route between Georgia and Tennessee.

Quizzes were of either MC or SA format. Each format of the quiz tested the same concepts; questions were simply re-worded to fit the format:

1. How many geographic regions are in the state of Georgia?
    a. 2
    b. 4
    c. 5
    d. 7

1. How many geographic regions are in the state of Georgia?

Subsequent to turning in the quiz, students received printed versions of the correct answers to the questions from the quiz. This served as feedback on those concepts.

1. There are five (5) geographic regions in Georgia.

A pre- and posttest was teacher-created and consisted of 24 questions: 12 multiple choice and 12 short answer. Eight questions were on material which was not included in the quizzes but was included in the review modules and served as a control for the testing effect. The remaining 16 questions were material included in the quizzes. Samples of the modules, quizzes, feedback and tests can be found in Appendix A through Appendix E.

Procedures

Before beginning the study, IRB approval was sought. Then, students were given a consent form containing a letter of explanation addressed to parents/guardians. The letter explained the purpose of the study and indicated that reviewing could ultimately help boost their

Milestone score. Students were informed that at any time they could ask that their data not be used but that they would be taking the tests and reviews with their class regardless of inclusion in the study. Parents were asked to sign the form and return it if they had issues with their student's data being included. No parents indicated that they objected.

Upon returning from winter break, students took a pre-test to assess prior knowledge. As stated above, this test consisted of 24 questions, 12 multiple choice, 12 short answer, and revolved around lessons to which the students were originally exposed during the first semester.

Students then began with the review modules. The massed groups began their review modules 14 days after taking the pre-test. The spaced groups started 13 days after taking the pre-test. Each of the four classes involved fell into one of the following groups:

- Massed review with multiple choice quizzing
- Massed review with short answer quizzing
- Spaced review with multiple choice quizzing
- Spaced review with short answer quizzing

The "spaced" groups received one module and its quiz each day for four consecutive days. The "massed" groups received all four modules and their quizzes on one day. The spaced groups began taking the modules on the same day as the massed group took all four of theirs but continued to take one module per day for four days total. Each review module was designed take approximately 10 minutes including the quiz. Students read the module, turned it in, and then received their quiz. Quizzes ranged from five to seven questions – either all MC or all SA as determined by their group. Students handed in their quiz and received the answer sheet which served as feedback.

DOI: 10.20429/ger.2017.140104

Eight days after finishing the review modules, students re-took the 24 question pretest. This test was to examine the testing effect. We examined how students with MC quizzes performed on the posttest as opposed to students with SA quizzes. Additionally, we compared how all previously quizzed material was retained in comparison to un-quizzed but reviewed material. One month after the posttest, students took the test a final time. This final test assessed the spacing effect on long term retention. It compared the retention levels of MC quizzes versus SA quizzes. The pre, post, and final test were all the same so students took the test three times, just at different intervals.

Results

Of the 76 students originally included in the study, several transferred out of the school. Additionally, students who did not complete either the pretest or the posttest were not included. This left 55 remaining in the study; 9 in the short answer/massed group, 18 in short answer/spaced, 13 in multiple choice/massed, and 15 in multiple choice/spaced.

An ANCOVA analysis was completed to compare the effect of question type on the result of the posttest compared to the pretest. With the pretest serving as covariate, the posttest as the dependent variable and multiple choice vs. short answer question type as the independent variable, the result was insignificant, $F(1,52) = .321$, $p = .573$, indicating that the type of question used in the review played no part in the increase of performance between the pretest and the posttest. A second ANCOVA was run with the pretest as the covariate and the final test scores as the dependent variable relative to question type showed similarly insignificant results: $F(1,48) = .158$, $p = .693$. This appears to indicate that the type of question used when reviewing played little role in the ultimate performance on the following tests.

Corresponding ANCOVAs were performed to look at the effect of review spacing on test performance. The first ANCOVA included the pretest as the covariate, the posttest as the dependent variable, and the massed vs spaced review as the independent variable. The result was significant, $F(1, 52) = 4.497$, $p = .039$. However, when the ANCOVA comparing the final test to the pretest with the spacing as the independent variable was performed, the results were insignificant, $F(1,52) = .020$, $p = .889$. Means and standard deviations for the spacing effect with pretest as the covariate can be found in Table 1. Table 2 shows the means and standard deviations for the spacing effect between the pretest and the final test. The results of this ANCOVA were insignificant, $F(1, 48) = 3.313$, $p = .075$. The sample size for this was slightly smaller due to four students missing the final. This does not meet the threshold for significance but may indicate that if all of the pretested students had taken the final test, results may have been significant.

Table 1
Posttest results for the spaced and massed review with pretest results as covariate

Group	Estimated Marginal Means	Std. Error	N
Massed	58..48	2.47	22
Spaced	65.307	2.01	33

Table 2
Final test results for the spaced and massed review with pretest results as covariate

Group	Estimated Marginal Means	Std. Error	N
Massed	65.26	2.58	20
Spaced	71.34	2.06	31

Finally, the scores were analyzed using a 2 x 2 (Review: Spaced vs Massed x Questions: Multiple Choice vs. Short Answer) ANOVA. The interaction between review style and question format was not significant, $F(1, 50) = 1.691$, $p = .199$.

DOI: 10.20429/ger.2017.140104

Discussion

The purpose of this study was to examine the benefits of different types of questions in a testing format with respect to recall. It was also designed to investigate the effects of massed review before tests compared to spaced review. Given the extensive literature available on both of these topics, the expected results were that students who spaced their review over time would show better retention than students who massed their review, and that students who had taken interim quizzes with forced recall/short answer questions would perform better on tests than those who had quizzed with recognition/multiple choice formatted quizzes. Additionally, students who had the premium combination of spaced review with short answer quizzing were expected to show better retention between the posttest and the final test than any of the other three possible combinations. This, however, was not the case. The only significant result was the Spaced Review group when comparing pretest to posttest. The fact that there was no appreciable gain between the posttest and final is to be expected as there was no further treatment following the posttest. In fact, when one examines the adjusted means, the spaced group made much more progress than the massed during the reviews, which reflect in the posttest scores.

Many previous studies, for instance Godbole, Delaney and Verkoeijen (2014), involved lists of words rather than text. The words were either seen one time by the test taker, or seen multiple times. If seen multiple times, the words were either seen in close repetition (massed review) or spaced apart (spaced review). Perhaps a crucial difference between procedures such as Godbole et al's and those done for this study is the source of material. Instead of learning random lists of words, the students in the current study were expected to remember facts from history. Instead of remembering a single word, these students were expected to extract bits and pieces of the content from their memory in order to answer questions.

Smith and Karpicke (2013) examined the value of forced recall questions over recognition questions and found that with sufficient success on the practice, the question format played very little role in effective recall. This study had many similarities to the current study. In both studies, the students used educational texts. Both studies involved mixed format tests and immediate feedback. Both studies showed limited differences between multiple choice and short answer questions and their effect on recall. Smith and Karpicke also spent considerable time comparing results within the test. They broke down how students performed on multiple choice questions and short answer questions within the tests rather than just as groups. Perhaps significant interaction would be indicated if tests had been run on how the short answer quiz groups produced on short answer test questions. Smith and Karpicke's study seemed to indicate that interim successes played a bigger role in overall success. Perhaps if the quizzes in the current study were more general or the text was shorter, resulting in more successful quizzes, the posttest and final test scores would have been higher as well.

Limitations

The motivation of the students themselves limited the results of the current study. This study required a pretest, four review modules, four quizzes, a posttest and a final test. With presentation of each element of the study, the students were quick to ascertain if a grade was being attached to their product. When they received a negative answer, their interest in the material dropped noticeably. The students with spaced review grew irritated each day they were faced with another review module and quiz. The students with the massed review had trouble maintaining concentration through all four modules and quizzes. While each quiz was followed with feedback in the form of sentences containing the correct answers and students made a show of reviewing the feedback, there did not seem to be any retention of correct answers. When faced

DOI: 10.20429/ger.2017.140104

with a project that had no accountability, they simply did not give it time or attention. To get an honest indication of the effect of the review methods or the question format, consequences or rewards for performance may be required.

Additionally, the sample size in each group was very small. The groups varied from 15 to 22 students. Eliminating students who missed the pretest, posttest, or final brought the sample size to 51. No account was made for which students may have missed any of the review modules. Therefore, "Christmas Tree" answers or other low scores had a larger effect on outcomes. Through random selection, the class of students which had the highest grade point average ("GPA") outside of the study was assigned to the least desirable combination of massed study and multiple choice quizzing. This was also one of the smallest classes with only 15 students (13 were included in the study). The class with the lowest GPA outside of this study was assigned to one of the more desirable combinations of spaced review with multiple choice quizzing. This was also one of the biggest groups with 20 students (15 of them were included in the study). The difference in self-efficacy and GPA between these two groups of students most likely had an effect in the results.

Implications

While this study did not produce many significant interactions, the one which was produced indicated that having spaced review was the most beneficial to students. This directly corresponds with the findings of several reports (Gluckman, Vlach, & Sandhoffer, 2014; Rohrer & Pashler, 2010; Sobel, Cepeda & Kapler, 2010). Further research should be conducted on the impact of question type. Ideally, this research should be done with larger sample sizes and with consequences or rewards for results. Additionally, the subject matter should be text rather than lists of words or paintings in order to make the study more generalizable to generic social studies

or history classrooms. It is counter-intuitive to think that seeing material multiple times does not aid retention, but more research is needed to find the ideal number of times and the spacing of the repetitions. Finally, the idea presented by Smith and Karpicke (2013) that question type in the review is not as important as success during the review merits further study. The implications are that easy review is more beneficial than challenging review, which would contradict the idea of short answer quizzing being more beneficial than multiple choice quizzing.

Conclusion

Typical middle school students require re-exposure to material several times in order retain and recall material. The question of ideal spacing between these exposures is something which still needs to be established and warrants further study. Additionally, the type of exposure and the self-efficacy of the students may also play into retention. This study looked at the spacing and type of question for the review. However, all the reviews involved reading text. Perhaps presenting different formats for the reviews such as videos, photos, timelines, or interactive projects would have proved more effective. While the reviews for the "spaced" groups were slightly more incremental than those used for the "massed" groups, they were still condensed into a relatively short time period. Perhaps having weekly reviews rather than daily reviews would have further increased retention. Finally, developing a way to keep the students involved in the activities may be essential in helping students realize the potential benefits of the spacing and testing effects.

One last issue that may be an obstacle for students and teachers is the preconceived idea of future success. Students who perceive themselves as being "bad at social studies" have a much harder time engaging with the material and therefore retaining any of it, which becomes a self-fulfilling prophesy. These students may show no effects of either treatment as they have simply

DOI: 10.20429/ger.2017.140104

given up. Finding ways to re-assure and re-engage students to keep them invested in the process is likely a prerequisite to the success of such interventions.

References

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology,* 23(6), 760-771. Doi:10.1002/acp. 1507

Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science,* 19(11), 1095-1102.

Gay, L. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement,* 17(1), 45-50.

Gluckman, M., Vlach, H., & Sandhofer, C. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. *Applied Cognitive Psychology*, (28), 266-273. doi:10.1002/acp.2997

Godbole, N., Delaney, P., & Verkoeijen, P. (2014). The spacing effect in immediate and delayed free recall. *Memory*, 22(5), 462-469

GPS by Grade Level, K-8. (2001). Retrieved July 27, 2015, from https://www.georgiastandards.org/standards/Pages/BrowseStandards/GPS_by_Grade_Level_K-8.aspz

Kang, S., McDermott, K., & Roediger, H. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4/5), 528-558.

Kornell, N., Bjork, R. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science,* 19(6), 585-592.

DOI: 10.20429/ger.2017.140104

McDaniel, M. A., Roediger, H. I., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14(2), 200-206. doi:10.3758/BF03194052

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*, 20(1), 3-21.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Expérimentale*, 67(3), 215-227. doi:10.1037/a0032918

Pashler, H., Rohrer, D., Cepeda, N.J., Carpenter, S.K., (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic: Bulletin & Review,* 14(2), 187-193.

Roediger, H. I., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal Of Experimental Psychology*: *Applied*, 17(4), 382-395.

Rohrer, D., Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher,* 39(5), 406-412. Doi:10.3102/0013189X10374770

Webb Bridge Middle School. (n.d.). Retrieved April 24, 2016, from http://www.schooldigger.com/go/GA/schools/0228001627/school.aspx

Smith, M., Karpicke, J. (2013). Retrival practice with short-answer, multiple-choice, and hybrid tests. *Memory,* 22(7), 784-802.

Social Studies Grade Eight Teacher Notes. (2012, October 16). Retrieved July 27, 2015, from https://www.georgiastandards.org/Frameworks/GSO Frameworks/Grade-Eight-Teacher-Notes.pdf

Sobel, H., Cepeda, N., & Kapler, I. (2010). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, (25), 763-767.

United States Census Bureau. (2015, May 29). Retrieved July 27, 2015, from http://quickfacts.census.gov/qfd/states/13/13117.html

Webb Bridge Middle School Alpharetta, GA Enrollment & Demographics. (n.d.). Retrieved April 24, 2016, from http://www.usaschoolinfo.com/school/webb-bridge-middle-school-alpharetta-georgia.25077/enrollment

DOI: 10.20429/ger.2017.140104

Appendix A

Sample Review Module

Review Module #1
Geography, Native American Cultures, and European Contact
Geography

The state of Georgia is divided into five geographic regions. The Blue Ridge Region is in the northeastern portion of the state. Dahlonega, the site of America's first Gold Rush is in this region. The Blue Ridge receives the most rain in the state and is the starting point of most of Georgia's rivers.

The Valley and Ridge Region was traditionally a mining region, with the valleys being used for farming. This region was a major battle ground during the Civil War and is a major transportation route between Georgia and Tennessee.

The Appalachian Plateau is in the Northwestern corner of the state and is the smallest region. It is sometimes called the TAG region.

The Piedmont Region is in the middle of the state and is the most populous of the five regions. Almost ½ of Georgia's population lives in the region. Many of Georgia's most important cities are located in the region including Athens, Atlanta, Augusta, and Macon.

The Coastal Plain Region is the largest region and makes up 3/5s of the state. The Inner Coastal Plain is the agricultural heartland of the state. The Outer Coastal Plain is home to Georgia's oldest city, Savannah.

The Fall Line is a natural boundary between the Piedmont and the Coastal Plain. Waterfalls found on the fall line caused many of the rivers to be difficult to navigate, but did offer sources of water power for mills.

Native Americans

The first people to inhabit the land that is now Georgia were the Paleo Indians. Paleos were nomadic hunters and gatherers who followed large game like mastodons and giant bison. Paleo homes were made from animal skins and were easily moved from place to place. "Paleo" means "very old."

The second culture to live in this area was the Archaic Indians. They were also nomadic. They also invented the grooved axe, the atlatl, and pottery, and made hooks and nets for fishing. The word "Archaic" means "old."

The third prehistoric culture here were the Woodland Indians. These were the first to use the bow and arrow, and they used pottery for storage. The Woodland Indians are the first to rely on farming for food. Because they relied on farming, they began to live in small villages with homes made of wood. They depended on corn, and they were the first mound builders.

The final prehistoric Native American culture was the Mississippian Indians. This is the most "complex" prehistoric culture in Georgia. They were large scale farmers and mound builders. This was the first group to encounter Europeans.

European Contact

Hernando de Soto, the first European explorer in Georgia, was directly responsible for starving and killing a large number Native Americans in his quest for God, gold and glory (1539-1542). Diseases such as influenza and smallpox caused massive population losses and the end of the Mississippian culture.

After DeSoto's expedition, both the Spanish and the French explored this area. Both attempted to create colonies. The Spanish set up several missions on both the barrier islands as well as the interior of the state. The primary reason for establishing these missions was to convert the natives to Christianity.

Although the French did explore the southeast, their primary focus was on the fur trade, so they did not have much of a presence in Georgia.

The English were interested in permanent colonization in North America due to mercantilism. In a mercantilist economy, the country sought to export more than it imported. Often, the "mother country" sought out colonies that could produce raw materials which would then be sent back for production. The colonies would then purchase the finished products. Other reasons for English settlement included "religious freedom" and the opportunity to begin "a new life." The first permanent English colony was Jamestown, Virginia, which was established in 1607.

https://www.georgiastandards.org/Frameworks/GSO%20Frameworks/Grade-Eight-Teacher-Notes.pdf

DOI: 10.20429/ger.2017.140104

Appendix B

Sample Multiple Choice

Multiple Choice Quiz – Review Module 1

Please pick the best response to the following questions.

1. How many geographic regions are in the state of Georgia?
    a. 2
    b. 4
    c. 5
    d. 7
2. Which geographic region is the largest?
    a. The Valley and Ridge Region
    b. The Piedmont
    c. The TAG
    d. The Coastal Plain
3. Which region contains the most population?
    a. The Valley and Ridge Region
    b. The Piedmont
    c. The TAG
    d. The Coastal Plain
4. Which Indians were the first mound builders?
    a. Woodland
    b. Paleo
    c. Archaic
    d. Mississippian
5. How did de Soto's expedition kill large numbers of Indians?
    a. By exposing them to disease
    b. By killing them in battle
    c. By forcing them to work at Spanish missions
    d. By forcing them to leave their land
6. Where did the Spanish build missions?
    a. The GA Mountains
    b. The Fall Line
    c. The Barrier Islands
    d. Atlanta

Appendix C

Sample Short Answer

Short Answer Quiz – Review Module 1

Please answer the following questions.

1. How many geographic regions are in the state of Georgia?
2. What geographic region is the largest (has the most land)?
3. Which region contains the most population?
4. Which Indians were the first mound builders?
5. How did de Soto's expedition kill large numbers of Indians?
6. Where did the Spanish build missions?

DOI: 10.20429/ger.2017.140104

Appendix D

Sample Feedback

Review Module 1 Answers

1. There are five (5) geographic regions in Georgia.
2. The Coastal Plain is the largest region.
3. The Piedmont is the region with the most population.
4. The Woodland Indians were the first mound builders.
5. De Soto's expedition brought 'flu and small pox to the Indians, killing them in large numbers.
6. The Spanish built missions on the barrier islands and the interior of the state.

Appendix E

Sample Test

Review Pre/Post Test

1. How many geographic regions are in the states of Georgia?
    a. 2
    b. 4
    c. 5
    d. 7

2. Which Indians were the first mound builders?

3. How did de Soto's expedition kill large numbers of Indians?
    a. By exposing them to disease
    b. By killing them in battle
    c. By forcing them to work at Spanish missions
    d. By forcing them to leave their land

4. Where did the Spanish build missions?

5. Who was the first European explorer in GA?
    a. Hernando De Soto
    b. John Rolfe
    c. James Oglethorpe
    d. Mary Musgrove

6. What is the name of the natural boundary between the Piedmont and the Coastal Plain which contains waterfalls that offer sources of water power for mills?

7. Who is considered to be the "founder" of Georgia?
    a. Chief Tomochichi
    b. Henry Ellis
    c. James Oglethorpe
    d. James Wright

8. Georgia was founded for these three reasons:

DOI: 10.20429/ger.2017.140104

9. The Salzburgers were
    a. Brought to Georgia based on their reputation for being the best soldiers
    b. Debtors
    c. Slave owners
    d. Peaceful and hardworking German speaking protestant refugees

10. Who allowed Oglethorpe to settle on Yamacraw Bluff?

11. The royal governor at the time of the American Revolution was
    a. James Wright
    b. James Oglethorpe
    c. Chief Tomochichi
    d. Mary Musgrove

12. The translator who helped Oglethorpe was

13. Georgia's *southern* border extended to this river after the French and Indian War:
    a. The St. Mary's River
    b. The Chattahoochee River
    c. The Mississippi River
    d. The St. Lawrence River

14. The Proclamation of 1763 forbade colonists from settling lands west of:

15. The Declaration of Independence was signed by these Georgians:
    a. Thomas Jefferson, George Washington, Benjamin Franklin
    b. Button Gwinnett, Lyman Hall, George Walton
    c. Austin Dabney, Elijah Clarke, Nancy Hart
    d. William Few, Abraham Baldwin

16. The Georgia patriot who is most well-known for capturing and killing several loyalist soldiers in her cabin is:

17. This battle raised the morale of Georgia patriots and gave them much needed supplies:
    a. Siege of Savannah
    b. Boston Tea Party
    c. Battle of Kettle Creek
    d. The Continental Congress

18. How many representatives did Georgia send to the First Continental Congress?

19. How many branches are there in Georgia's state government?
    a. 1
    b. 2
    c. 3
    d. 4

20. What is the name of Georgia's Legislative Branch?

21. This is the largest branch of Georgia's state government:
    a. The Executive Branch
    b. The Legislative Branch
    c. The Judicial Branch
    d. The Olive Branch

22. Who is sometimes called the "President of the Senate?"

23. This court is the highest in the state:
    a. The Superior Court
    b. The Juvenile Court
    c. The Supreme Court
    d. The Probate Court

24. A serious crime, which must receive at least one year in jail is called:

DOI: 10.20429/ger.2017.140104