Fall 2016

The Effects of Repeated Global Self-Adapted Testing on Online Statistics Performance

Gita Sierra Hodell
San Jose State University

THE EFFECTS OF REPEATED GLOBAL SELF-ADAPTED TESTING ON ONLINE STATISTICS PERFORMANCE

A Thesis

Presented to

The Faculty of the Department of Psychology

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Gita Hodell

December 2016

The Designated Thesis Committee Approves the Thesis Titled

THE EFFECTS OF REPEATED GLOBAL SELF-ADAPTED TESTING ON ONLINE STATISTICS PERFORMANCE

by

Gita Hodell

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

SAN JOSÉ STATE UNIVERSITY

December 2016

Sean Laraway, Ph.D	Department of Psychology
Ronald Rogers, Ph.D	College of Social Sciences
Susan Snycerski, Ph.D	Department of Psychology

ABSTRACT

THE EFFECTS OF REPEATED GLOBAL SELF-ADAPTED TESTING ON ONLINE STATISTICS PERFORMANCE

by Gita Hodell

The purpose of this study was to evaluate the effectiveness of a computerized academic testing format. Centered on the motivating and stress-reducing aspects of personal control, a modified form of global self-adapted testing (GSAT) was explored to help students who are challenged by test anxiety or low academic motivation. Forty-two students completed multiple GSATs throughout one semester of college-level, online statistics. Of those students, 20 volunteered to complete an academic motivation questionnaire at the beginning of the semester. The relationships between scores on the motivation questionnaire, GSAT use characteristics, and statistics performance were analyzed. Students who used the GSATs correctly approached more challenging questions and performed better on exams than did students who used the GSATs incorrectly. However, the class that experienced the GSAT intervention did not differ significantly on exam scores when compared to a class that did not experience the GSAT intervention. We concluded that GSAT did not improve statistics performance. Confounds which could have limited the results of this study are discussed.

ACKNOWLEDGMENTS

I would like to express appreciation for Drs. Pamela Stacks and David Bruck, whose combined efforts made electronic thesis submission and review possible at SJSU. Thanks are due to Cheryl Cowan, the Graduate Studies Associate, and former Thesis Coordinator Alena Filip, in the Office of Graduate Studies. Cheryl and Alena's involvement in graduate research is a valuable resource to all graduate students who seek help and information. I would also like to thank the anonymous thesis reviewers whose efforts are essential for maintaining excellent standards in graduate studies.

Thank you to my thesis advisor, Dr. Sean Laraway, and my esteemed committee members, Drs. Ron Rogers and Susan Snycerski. The three of you have set an example of openness to new ideas and academic inquisitiveness. You inspired me to never stop asking questions and to trust myself. I also owe my gratitude to Dr. Clifton Oyamot for handing over enough responsibility to allow me to work freely within the class he was teaching. Finally, I would like to thank the undergraduate students of SJSU who made my thesis possible through their participation.

TABLE OF CONTENTS

List of Tables…… ix

List of Figures…….. x

List of Abbreviations……..... xi

INTRODUCTION……..... 1

Test Anxiety..... 2

Academic Motivation..... 5

Computer-Adapted Testing..... 7

Self-Adapted Testing..... 8

The Present Study..... 10

METHOD.....…….. 12

Participants……. 12

Materials.....…….. 13

Canvas..... 13

Udacity..... 13

GSAT..... 14

Measures..... 17

Academic Motivation and Test Anxiety..... 17

GSAT..... 19

Usage characteristics…….. 19

Statistics performance……. 19

Procedures..... 19

Summer 2014: Baseline..... 19
Fall 2015: GSAT condition..... 20
GSAT..... 20
RESULTS..... 21
Motivation and Test Anxiety..... 22
MSLQ..... 22
Usage Characteristics..... 24
Correct use..... 25
No change..... 25
Change..... 26
Incorrect use..... 26
Repeated levels..... 26
Attrition..... 26
Collapsed groups..... 27
Motivation by subgroup..... 28
Attempted difficulty..... 29
Statistics Performance..... 32
DISCUSSION..... 34
MSLQ..... 34
Usage Characteristics..... 37
Correct use..... 37
Incorrect use..... 38

Statistics Performance..... 40

Limitations..... 41

Future Directions and Conclusions..... 43

REFERENCES..... 45

Appendix A. IRB Approval Letter..... 47

Appendix B. GSAT Configuration Options..... 48

Appendix C. GSAT Written Instructions..... 49

Appendix D. Motivated Strategies for Learning Questionnaire..... 50

LIST OF TABLES

Table 1. Descriptive statistics of Su14 and Fa15 MSLQ scores..... 22

Table 2. Independent samples t-test for the Motivated Strategies for Learning Questionnaire..... 23

Table 3. Pearson correlation for Motivated Strategies for Learning Questionnaire and average exam score, both semesters..... 24

Table 4. Frequency distribution and percent of each usage characteristic..... 25

Table 5. Collapsed usage characteristics groups..... 28

Table 6. MSLQ scores broken down by use characteristic groups..... 29

LIST OF FIGURES

Figure 1. Correct usage attempted GSAT difficulty over time..... 30

Figure 2. Incorrect usage attempted GSAT difficulty over time..... 31

Figure 3. Total usage attempted GSAT difficulty over time..... 32

Figure 4. ANOVA of exam score by exam position and semester taken..... 33

LIST OF ABBREVIATIONS

CAT – Computer-adapted test

Fa15 – Fall 2015

GSAT – Global Self-adapted test

MSLQ – Motivated Strategies for Learning Questionnaire

SAT – Self-adapted test

SJSU – San Jose State University

Su14 – Summer 2014

Introduction

When teachers need an objective way to measure the extent to which their students have mastered a subject area, it is common to give students a fixed-item list of questions they must answer. The better their performance on the test, the higher their ability is presumed to be. Even though knowledge of the subject matter is the primary variable that is measured, there are other factors that can add measurement error to the final scores on tests. For example, examinees with test anxiety could adequately prepare for an exam, but when they arrive at the exam they may experience intrusive, deleterious feelings of worry that interfere with their ability to answer questions correctly even though they understand the material. In this case, a low score on the test would not necessarily reflect low ability or poor understanding of the material, but rather the affective state of the examinees (anxiety). In contrast, if examinees chose to watch movies instead of study for an exam, their subsequent low score on the exam would be a reflection of low effort. Had they studied, they could have achieved a score that was closer to their true ability.

It may be simple to say that that high test anxiety is not representative of the average student, or that ill-prepared students deserve to get lower scores on their exams. However, fixed-item tests are ubiquitously used as objective measures of ability, and the outcome of some of these tests can have potentially life-long consequences. In regard to academic trajectory, standardized testing is used as a gateway from one level of education to the next. Performance based assessment during the K-12 years is often used to group students together with high, average, or low academic ability. In high school, Scholastic

Aptitude Test scores are used to determine whether or not a student will attend college. After college, Graduate Record Examination scores are used to determine whether or not a student can continue on to graduate school. Test scores collected from standardized tests are the key to academic trajectories and career opportunities. The issue of fixed-item, paper-and-pencil tests negatively affecting underprepared or test-anxious students prompts the following questions. First, is a standardized, "one-size-fits-all" test the most accurate way to measure students' abilities and understanding of academic material? Second, if there are more accurate ways to test ability without disproportionately affecting certain groups of people, what would they be?

Ability has always been an important factor to measure. Musch and Bröder (1999) measured the amount of variance in academic test scores that was explained by ability alone. They found that ability was the largest contributor to variance in test scores in terms of the proportion of variance that it accounted for. They also observed that test anxiety was a somewhat smaller, yet still significant predictor, of test score variance. In addition, Wise and DeMars (2005) integrated 12 empirical studies and concluded that motivated examinees received performance scores an average of 0.59 standard deviations higher than did unmotivated examinees. The takeaway from these studies should be that some, but not all of standardized test score variance is explained by ability. The following sections will describe some of the factors that influence variance in test scores.

Test Anxiety

Test anxiety is associated with fear of failure (related with motivation to succeed), and personal awareness of physiological arousal (e.g. sweating, heart rate). The

dimensions of test anxiety, as described by Liebert and Morris (1967), are worry, and emotionality. Of the two, intense worry is considered to be the most predictive of poor test performance (Morris & Liebert, 1973). The presence of test anxiety can lead to intrusive thoughts that divide the attention of the examinee, causing a distraction (Sarason, 1984). According to Liebert and Morris (1967), worry reflects a sense of predicted failure. For instance, examinees that worry during a test do so because they feel they are not able to perform well.

Musch and Bröder (1999) surveyed undergraduate students from a statistics class about their study habits, test anxiety, and math skills, and performed a hierarchical multiple regression analysis to determine the amount of variance each factor accounted for in performance (on the final exam). The authors compared two models that explained low test performance; one model stated that lack of ability or skill contributed to poor exam performance, and the second model stated that anxiety causes interfering thoughts during the test. They found that math skill (measured by previous grades in math) contributed the most to accurately predicting performance on the test. Following math skill, test anxiety accounted for a significant portion of the variance. Study habits did not significantly predict performance. Therefore, math skill and test anxiety both contributed to performance on tests.

Test anxiety can be present in two different ways. Trait test anxiety is present whenever a test of any kind is being taken, and is independent of subject. However, state test anxiety is the heightened arousal that accompanies test taking in a specific subject area. For example, statistics students may experience statistics state test anxiety only

when they are taking a statistics test. The prevalence of statistics anxiety among American graduate students is reported at 80% (Ongwuebuzie, 1998). The dimensions of statistics anxiety are composed of six major factors: worth of statistics (the perception that statistics are relevant), interpretation anxiety (experienced while trying to interpret results), test/class anxiety (experienced while in class or taking tests), computational self-concept (self-perception of mathematical ability), fear of asking for help, and fear of statistics instructor (Baloglu, 2002; Cruise, Cash, Bolton, 1985). Some common antecedents of statistics anxiety are fewer mathematics courses taken, little knowledge of statistics, and poor achievement in mathematics (Roberts & Bilderback, 1980).

In general, test anxiety has been linked to achievement goal orientation (Yang & Taylor, 2013). The achievement goal framework (Elliot & McGregor, 2001) consists of mastery approach (learning for the sake of learning), mastery avoidance (learning out of fear of failure), performance approach (wanting to do better than others), and performance avoidance (fear of looking incompetent compared to others). Yang and Taylor (2013) surveyed graduate and undergraduate, online psychology students who conveyed their self-reported achievement goal orientation, sense of self-efficacy, and willingness to ask other students for help with homework. They found that both mastery and performance avoidance predicted higher amounts of test anxiety. Ultimately, the two achievement definitions associated with fear; fear of failure (mastery avoidance), or fear of appearing incompetent (performance avoidance), are also associated with test anxiety. Synthesis of these studies indicates that test anxiety, whether in respect to a subject like statistics, or just in general, is related to prediction of failure. The next section will focus

on previous literature regarding the relationship between prediction of failure, motivation, and task approach.

Academic motivation

Low-effort test taking may be a sign that examinees lack the motivation to finish a test using the best of their abilities. Some behaviors of low-effort test taking can include answering test items randomly, too quickly, or not at all. When the estimation of ability from a test is much lower than the true ability of the examinee, there is error in the test scores. There are several reasons an examinee might use low-effort on a test. For instance, unmotivated examinees might be prone to select items from the lowest possible level of difficulty to expend a minimal amount of effort. This type of behavior is explained in the expectancy value theory of achievement motivation (Pintrich, 2003; Wigfield and Eccles, 2000). The Wigfield and Eccles expectancy-value theory asserts that people make task decisions based on beliefs about their ability, the expected outcome, and the relative value of the task. People have beliefs about their ability in certain subjects, and they rely on them to predict an outcome of performing a task.

Weiner (1985) reasoned that anticipated success partially depends on prior experience, and context. Failure in a previous math class could be attributed to lack of ability (stable cause), or it could be attributed to lack of preparedness (unstable cause). If the cause of the failure is perceived to be stable, then an individual is more likely to predict failure for themselves in future math classes. Bandura's self-efficacy theory states that ingrained beliefs about ability can predict task approach as well as performance. Academic self-efficacy is often measured by asking students how confident they are they

can succeed on a task, such as completing a class, or earning a good grade. For example, mathematics students with a higher sense of self-efficacy also show higher intrinsic motivation and approach more challenging tasks (Bandura & Schunk, 1981). Conversely, low self-efficacy students may avoid challenging tasks due to the belief that they are unable to accomplish them.

When students choose between alternative task approaches, in addition to beliefs about ability and prediction of success, it is important to account for the subjective value of the task. Wolf, Smith, and Birnbaum (1995) investigated the difference in exam performance between 10th and 11th graders with high and low extrinsic motivation to succeed, and concluded that students with higher extrinsic motivation performed better on mentally taxing exam items than their less motivated counterparts. A state-wide high school graduation test was given to both groups of students from the same high school. The 10th graders' graduation from high school depended on their performance on the test, however the 11th graders, who had already taken the test in a previous year, did not have this consequence attached to their performance. Although the 11th graders could have outperformed the 10th graders, because they were older and had more experience with the test, the 10th graders ultimately achieved higher scores. This study illustrates the cost-benefit analysis that many students go through when deciding between task approach and task avoidance. The question of: can I do well, becomes will I do well, and is resolved by how important is it that I do well?

One thing to note about these studies is that they all conclude test performance is not based solely on the ability of the test taker. Instead, performance can be manipulated

by the affect of the test taker, their beliefs about their ability to do well, and the extent to which they value the outcome. With considerable evidence indicating detectable differences in the ability estimates of students who take tests with high anxiety or low motivation, one wonders if there could be a better way to approach academic examination. Could modern education create tests that are comparatively more motivating and less anxiety producing than traditional, fixed-item, paper and pencil tests?

Computer-Adapted Testing

Computer-adapted testing (CAT) is a form of computerized assessment that adjusts the difficulty of the items one at a time. Instead of a fixed-item format, the CAT remains fluid as it assembles itself. If an examinee were to answer a test item correctly, the following item would either match, or exceed the difficulty of the previous item. Conversely, if an item were answered incorrectly, the following item would present a lower level of difficulty. An algorithm selects test items both (a) at the time of the test taking, and (b) based on the skill level of the examinee. One of the problems with CAT is that the items must be completed sequentially, which means there is no opportunity for item review.

Stowell and Bennet (2010) gave undergraduate students in a psychology class one computerized exam online, and another paper-and-pencil exam in person. Before and after the exams, students filled out the Academic Emotions Questionnaire to measure test anxiety. They found that test anxiety was significantly negatively related with performance on the exams. They also found that the relationship between test anxiety and performance was weaker in the computerized exam condition. The participants who

reported a preference for online examination also had higher in-person test anxiety than those who reported a preference for paper-and-pencil exams.

Self-Adapted Testing

Early work by Rocklin and O'Donnell (1987) explored a style of computer based testing called self-adapted testing (SAT) in order to counteract test anxiety. SAT users are allowed to adjust the difficulty level of each item in the test according to their self-estimated ability. By infusing the exam process with a sense of personal control, Rocklin and O'Donnell demonstrated that students who used SAT outperformed other groups taking fixed-item, computerized tests. Participants in the study were college students (*M* age=20.6). Participants were told to complete the Test Anxiety Inventory, then complete a computerized verbal test. They were randomized into three groups, the easy test, the hard test, and the self-adapted test. The results demonstrated that participants with test anxiety in the self-adapted testing condition chose lower difficulty items at first, then gradually increased the difficulty of the items as the test continued. The self-adapted test also led to higher ability estimates for those with testing anxiety.

Rocklin (1994) suggested that examinees taking SATs achieve maximal performance by selecting items optimally matched not only to their skill levels but also to their current affective and motivational characteristics. Wise, Roos, Plake and Nebelsick-Gullet (2004) sampled both undergraduate and graduate students (*N*=377) on their preference for CAT or SAT. Participants were randomly assigned to take a computerized algebra test in a CAT format, a SAT format, or a choice format. Those in the choice category were instructed to choose between a CAT or a SAT before beginning the test.

The study revealed that participants with low test anxiety showed a preference for CAT, and participants with high test anxiety showed a strong preference for SAT. Also, among the participants who chose SAT there was a decrease in state anxiety from pre-test to post-test.

In a meta-analysis, Pitkin and Vispoel (2001) analyzed 19 studies that compared CAT ability estimates with SAT ability estimates. They found that SAT performance scores are 0.12 standard deviation units higher than CAT scores. There are a number of reasons why SAT users could out-perform CAT users. One reason is that test takers could choose to answer questions that are below their ability level, which would produce positive bias in ability estimation. Another reason is that post-test anxiety scores are lower for SAT test takers than for CAT test takers. Pitkin and Vispoel noted that post-test anxiety levels of SAT users are an average of 0.19 standard deviation units lower than those of CAT users. Some disadvantages of using SAT are that it takes longer to complete than other forms of testing (Vispoel, 1998; Vispoel & Coffman, 1994). It is also labor-intensive to develop question banks that are large enough to offer different options for each question in the SAT.

Assuming some examinees feel more worried about test taking than others, it is possible that personally controlling the difficulty of the test can reduce anxiety. According to the perceived control hypothesis, personal control, either real or perceived, reduces stress associated with aversive situations (Averill, 1973; Wise, 1994). By limiting anxiety caused by the fear of failure, personal control over the testing environment appears to promote self-efficacy in students by increasing their expectation

of success.

The Present Study

The focus of the present study was to determine whether or not a testing instrument targeting high-anxiety and low-motivation behavior could improve exam performance. A modified version of SAT was used in an online, undergraduate statistics course. The characteristics of the SAT we used differed in three crucial ways from the type of SAT used in Rocklin and O'Donnell (1978). First, difficulty adaptations were not possible at the item level. Instead, we split the test into two halves and made each half adaptable at the test level. For instance, students chose between difficulty options for the first half of the test, then halfway through the test they were told they could adjust the difficulty level for the second half of the test. Therefore, a more apt name for the testing instrument is global self-adapted testing (GSAT), as difficulty adjustments were applied globally rather than at the item level. The primary concern with adopting the GSAT strategy was that it could confuse the students who were not accustomed to taking tests in this fashion. The reasoning behind the decision to use GSAT, and not traditional SAT, came from the specific technological capabilities that were available to us at the time of this study.

Another way in which our GSAT method differed from traditional SAT was that as the testing level of difficulty decreased, the number of items within the test increased. By designing the GSAT this way, students who took the easy test were required to spend longer amounts of time engaged in the material than those who took the hard test. For instance, a less motivated, or test-anxious student might favor the easy path and avoid the

hard path for the GSAT while knowing in advance that the cost of taking the easy path was that they had to answer the highest number of questions. The purpose behind this design decision was to give students more practice where it mattered the most without punishing them for feeling less comfortable with the material. In contrast, the non-test anxious, highly motivated, and high-effort examinees would not only benefit from the challenge of choosing a medium or hard GSAT, they would also be rewarded with fewer test items to complete.

Finally, a major difference lies in the type of data being collected. Past studies have focused on comparing ability scores between two groups, one with SAT and one without. Unlike those studies, we were not interested in the short-term benefits of GSAT, or even how high the ability estimates were on the GSAT. Rather, our study attempted to observe longitudinal effects over the course of a full semester of repeating a GSAT once per week. To chart general behaviors in student engagement with the GSAT, the difficulty levels that the students chose were recorded. Lastly, after the course was over and the students had a chance to engage with the GSAT repeatedly, the ability scores distributed across three midterm exams were recorded**.** The primary question of this study was: Can an online class that uses repeated GSAT over one semester achieve higher performance scores in statistics than a class without GSAT? A secondary question we investigated was: What insights can be gained by using repeated GSAT to track task approach and avoidance behaviors such as high-effort and low-effort test taking?

Method

Participants

Participants for this study included 55 undergraduate students recruited from two different semesters of introductory statistics, an online course hosted by San José State University (SJSU). All students were non-statistics majors. The two semesters we used were Summer 2014 (Su14), and Fall 2015 (Fa15). Students who registered for the class but did not complete any GSATs or cumulative exams (n=5) were excluded from data collection. Of the participants who were included, 38 volunteered to complete a pre-class demographic questionnaire. The ages of the participants ranged from 18 - 31 (M=20.68, SD=3.21, $Range$ =13). There were 11 males and 27 females. The racial/ethnic makeup of the sample was 12 Hispanic/Latino, 12 Caucasian, 10 Asian, 2 African American, and 2 Native American.

During the first week of instruction, students were informed via email that they were invited to participate in a research study by providing self-reported information. A number of surveys were available through the class website such as demography, digital literacy, academic motivation, and statistics anxiety. Students who volunteered to provide information indicated their consent by completing an online consent form. In exchange for their participation, they were offered extra credit. Those who did not volunteer to fill out the surveys were offered an alternative extra credit assignment of equal value. In addition to demographic and motivation data, we collected secondary class performance data from both the Su14, and the Fa15 classes. The data came from archived student records that had been stored in the online learning management system.

The identities of all participants were removed by the course instructor before any data were analyzed. This study was approved by the SJSU Institutional Review Board (see appendix A).

Materials

Canvas. Canvas is an online learning management system (LMS) that is used by SJSU as well as other universities (https://www.canvaslms.com/). Students at SJSU can log in to the Canvas website and access course materials for multiple classes in which they are enrolled. For in-person classes, professors may choose to use Canvas as a way to supplement their course. However, online classes at SJSU use Canvas as the central platform to distribute lesson content, assignments, quizzes, and exams. The researchers in this study customized the computerized GSATs by using the software available in Canvas.

Udacity. Udacity is a website that hosts educational courses in video format (https://www.udacity.com/). Dr. Ron Rogers (Associate Dean of the College of Social Sciences) and Dr. Sean Laraway (Associate Professor in the Department of Psychology and statistics instructor) of SJSU collaborated with Udacity to create a series of online videos that comprised an elementary statistics course. Between 8 and 30 short video lessons were grouped together under 15 different subject headings. Each subject heading composed one module of statistics information, such as "Comparing means using t-test and one-way ANOVA." There were also periodic learning checkpoints embedded in the videos, where students could pause the video lesson and answer quiz questions about

what they had just learned. Completing the learning checkpoints was optional, and was intended to provide an active engagement component to the videos.

The Su14 class required both the Canvas and Udacity websites in order to access the video lessons. However, before the Fa15 class begun all the lesson videos were imported from Udacity into Canvas in order to consolidate the material onto a single learning platform. A notable difference between the courses was that in Su14, the videos hosted by Udacity had fully interactive learning checkpoints, such that students had to answer mini-quiz questions embedded within the current video in order to advance to the next video. When the lesson videos were ported into Canvas, they lost their interactive functionality. The researchers tried to replicate the interactive nature of the learning checkpoints by creating mini-quizzes within Canvas and accompanying each lesson video with a hyperlink to the quiz questions.

GSAT. The researchers generated 15 GSATs, one for every lesson module, to be completed as homework assignments. Each GSAT had two parts, part A and part B. Students could select from three difficulty levels (easy, medium, or hard), once before they began part A, and again before they began part B. Essentially, students chose a level of difficulty for their homework assignment, and then had the opportunity to change their decision once they were halfway through. The easy level quizzes were composed of 10 items, the medium levels contained five items, and the hard levels contained two items. Therefore, six individual quizzes made up of 34 total items were generated for each GSAT. Overall, 510 quiz items were added to the question database (see Appendix B). To generate the high volume of quiz items, the researchers consulted published statistics

textbooks and instructor manuals with statistics question banks (Carlson & Winquist, 2014; Hendricks, Walls & Heiman, 2004; Howell, 2011; Wilson, 2005). The quiz items were not copied verbatim, but their narrative format was used as a guideline to generate similar questions.

Since the GSAT was intended to include an easy, medium, and hard version of the same quiz, one criterion used to determine item difficulty was that many of the items collected from the published question banks came pre-categorized into recommended levels of difficulty. However, in some cases the item difficulty was not pre-identified, so we had to use our best judgment. To do this we adhered to subjective criteria, such that an easy item could contain one statistical concept of low complexity. For example, "To calculate a z-score from a raw score, what three pieces of information do you need?" was used as an easy GSAT item from the "standardized scores " module. The z-score formula was displayed next to this question, and the response options were in multiple-choice format. For this question, students had to identify the variables in the formula, but they did not have to do any calculation. A medium level question could contain two statistical concepts of moderate complexity, such as the item "Put these steps in order for calculating the z-score for a sample mean" from the "sampling distributions" module. This question did not display the z-score formula. Instead, the students had to place the steps in the right order using a drop down menu. First, students either had to search for the z-score formula, or use their memory. Second, they had to identify the order of operations without doing any calculation. A question could be classified as hard if it contained more than two statistical concepts of high complexity. A representative hard

item is “Two samples, each with n = 6 subjects, produce a pooled variance of 20. What is the SEM for the sample mean difference?” from the “Independent samples t-tests” module. This question was presented as a fill in the blank response type, and had no visual or memory aids to go with it. Students had to look at notes or use their memory, then perform the calculation correctly.

To incentivize students to complete weekly GSATs a point system was created. We chose to make the GSATs compulsory instead of optional, and therefore a small portion of the overall grade in the course was dependent on completing the GSATs. We wanted students to factor certain costs and benefits for taking each level of difficulty into their decision making. Therefore, all GSATs were worth the same number of points (2.5 for part A, 2.5 for part B, 5 points total), but each level of difficulty contained a different number of items. The easy GSAT always had ten questions. Students were encouraged to take the easy GSAT if they were uncomfortable with the material from the module, or felt like they needed more practice (see Appendix C). The benefit of taking the easy GSAT was the simplicity of the questions, yet the cost was that they had to answer a relatively large number of questions. In contrast, the medium GSAT always contained five questions. Students were encouraged to take the medium level of difficulty as a challenge if they did well on the easy GSAT, or a relief if they did poorly on the hard GSAT. The benefit of taking the medium level was potentially more efficient use of time due to the fewer number of items, however the cost was that the items were slightly more mentally taxing than those of the easy level. Finally, students were encouraged to take the hard GSAT, which had only two items, if they wanted to challenge themselves.

Similar to the medium level, the benefit of taking the hard test was that fewer items needed to be answered, but the cost was that those items were more mentally taxing.

Measures

Academic motivation and test anxiety. The Motivated Strategies for Learning Questionnaire (MLSQ; Pintrich & De Groot, 1990) was used to measure student academic motivation and test anxiety in the Su14 and Fa15 semesters (see Appendix D). The MSLQ is a self-report survey that addresses beliefs that motivate students to learn, and habits that students employ when they are learning. The independent variables were the five dimensions of the MSLQ. The dependent variable was the extent to which students agreed with each of the statements in the questionnaire. All items had Likert-style response categories. The original MSLQ had a 1 to 7 rating scale (1 = Not at all true of me, 7 = Very true of me), however a technical issue with our computerized version of the MSLQ reduced the number of response categories from 7 to 5, such that 1 = Not at all true of me, and 5 = Very true of me. It is important to note this difference between the questionnaire used in this study and the original.

The MSLQ has two scales, the *motivation* scale, and the *learning strategies* scale. The motivation scale has three dimensions, *value components*, *expectancy components*, and *affective components*. The value components dimension is concerned with beliefs about the worth of completing academic tasks, and the type of motivation that students use to engage with the tasks (intrinsic vs. extrinsic). An example from the value components scale is, “In a class like this, I prefer material that arouses my curiosity, even if it’s difficult to learn” (intrinsic motivation).

The expectancy dimension is concerned with student beliefs about their performance capabilities, and the extent to which they take responsibility for their own learning progress. A representative item from this sub-scale is "I'm confident that I can do an excellent job on the assignments and tests in this course" (self-efficacy). The third dimension of the motivation scale is affective components. Affective components measure test anxiety. A representative item from this subscale is "When I take tests I think of the consequences of failing." (test anxiety). Together, the three dimensions of the motivation scale measure beliefs and affect.

The second scale in the MSLQ is called the learning strategies scale. The first dimension of learning strategies is called Cognitive and Metacognitive Learning Strategies. This dimension is concerned with whether or not students use strategies such as rehearsal, organization, critical thinking, and self-regulation. A representative item is "When I study for this class I practice saying the material to myself over and over" (rehearsal).

The second dimension is called Resource Management, and is concerned with how students utilize their individual environments while they study, and how useful their study habits are. For example, the item "I find it hard to stick to a regular study schedule" (reverse scored). The MSLQ was scored by generating an average score for each of the five dimensions. Together, two dimensions of learning strategies measure the behaviors that students employ while they are learning.

GSAT.

Usage characteristics. To measure how Fa15 students engaged with GSATs over a long period of time, we took into account the length of time spent in the semester, and the levels of difficulty attempted week by week. Over the course of the 15 week semester, we were interested in whether or not students would approach a challenge by taking harder GSATs over time, or if they would avoid the challenge by taking the easiest levels possible. Note that actual performance on the GSATs was not measured, only the level of difficulty attempted.

Statistics performance. To measure overall statistics performance longitudinally, the independent variable was the serial position of the exam (first, second, or third), and the total score of each exam was the dependent variable. To measure statistics performance by semester, the cumulative exam scores from the Su14 (non-GSAT) semester were compared to the Fa15 (GSAT) semester.

Procedures

Summer 2014: Baseline. The Su14 online statistics class lasted 10 weeks during the summer session of 2014. During the first week of class, students completed the MSLQ on a voluntary basis in exchange for extra credit. Dr. Sean Laraway was one of the original content creators for this course. He appeared as the instructor in the Udacity lesson videos and interacted with the students by holding weekly online question and answer sessions. In addition to these chat sessions, two graduate students with experience in elementary statistics were available to assist the students with questions. All students were encouraged to initiate discussions and ask questions at any time

throughout the course by posting in Piazza, an online class communication platform (https://piazza.com). All course videos were accessed through Udacity, and all quiz and exam content was accessed through Canvas. Homework quizzes were assigned once per week, and every three weeks there was a cumulative exam.

Fall 2015 (Fa15): GSAT condition. For our quasi-experimental condition we used the same online class from a different semester. During the first week of class, students completed the MSLQ on a voluntary basis in exchange for extra credit. Fa15 was a 16-week course instructed by Dr. Clifton Oyamot. Because the lesson videos were already made by Dr. Laraway, we were able to control for how the lesson content was taught during both semesters. However, Dr. Oyamot was responsible for keeping online communications with the students available. To do this, he offered impromptu statistics coaching sessions through Piazza, and encouraged all students to post questions and discussions. Unlike the Su14 semester, Fa15 did not employ any graduate student statistics coaches. The lesson videos had been imported to Canvas, so Fa15 students only had to log into one website to access videos, homework, and exams. The link to the Udacity website was still active and provided to the students, although logging into Udacity was neither necessary nor required to view the lesson videos. Similar to the Su14 semester, weekly fixed-item homework quizzes were assigned, and cumulative exams were assigned on week 5, 9, and 13.

GSAT. In addition to weekly fixed-item quizzes, students in Fa15 were asked to complete weekly GSATs after they finished watching the lesson videos corresponding to the module for that week. Written instructions were provided to help the students

navigate the GSAT. The GSAT corresponding to module 1 was called "Self-adapted Quiz 1A/1B," module 2 was "Self-adapted Quiz 2A/2B," and so on. Both parts A and B had to be completed consecutively. Once the data collection was completed, student performance information was collected from the online grade book. Personally identifying information about the students was removed by Dr. Oyamot and replaced with numeric codes before the data was given to the researchers. The levels of difficulty completed in the GSATs were given numeric codes, such that 2 indicated the easy level, 4 indicated the medium level, and 6 indicated the hard level. Therefore, smaller average numbers reflected easier levels attempted and larger average numbers reflected harder levels attempted. The mean of quizzes A and B was used to indicate an overall score for each GSAT. The number coding system was not used to indicate student performance on the quizzes, but rather the difficulty level that was attempted. Once all the GSAT information was coded, the data were entered into the SPSS statistical software program. The data from the Su14 and Fa15 demographic surveys, MLSQ, and cumulative exam scores were then compiled into a master file and prepared for analysis.

Results

First, we measured the MSLQ scores of the Su14 and Fa15 semesters, with special interest in the motivation and test anxiety components, to identify any covariates before conducting any other analyses. The linearity of the relationships between the five dimensions of the MSLQ (value, expectancy, affect, cognitive/metacognitive strategies, and resource management) and average statistics performance was also evaluated. Second, we charted the attempted level of difficulty for each GSAT. Weekly averages

and preferences for certain test taking behaviors were noted. Finally, to test the hypothesis that prolonged use of GSATs over a full semester would have a positive effect on statistics performance, cumulative exam scores from semesters Su14 and Fa15 were compared. We used a 2 x 3 mixed ANOVA design with one between subjects factor (semester: Su14, Fa15) and one within subjects factor (serial position of exam: first, second, third). The dependent variable was the total percentage of correct answers on each of the three exams. For all hypothesis tests, the level of significance was set to .05.

Motivation and Test anxiety

MSLQ. Motivation, test anxiety, and learning strategies were measured by the MSLQ to measure any differences between Su14 and Fa15. If differences existed, it would be important to treat them as covariates in subsequent analyses. Scores for the MSLQ were calculated by taking the average of each dimension.

Table 1

Descriptive Statistics of Su14 and Fa15 MSLQ Scores

	N	*M*	*SD*	*Skew*
Summer 2014	16	3.44	0.45	-0.05
Fall 2015	20	3.67	0.47	-0.36

Note. MSLQ = Motivated Strategies for Learning Questionnaire

Before making direct comparisons between groups, descriptive statistics were calculated to ensure that the assumption of normality was not violated (see Table 1). The MSLQ scores for the Su14 (M=3.44, SD=0.45) and Fa15 (M=3.67, SD=0.47) classes were similar, and appeared to be normally distributed. Fa15 scores were slightly more positively biased (*skew* = -0.36) than the Su14 class (*skew* = 0.05), however the skewness did not exceed tolerable levels (*skew* > 1, or < -1). Therefore, we determined that the assumption of normality was not violated.

To identify potential differences between Su14 and Fa15 on any dimension of the MSLQ, multiple *t*-tests were computed by comparing each dimension (value components, expectancy components, affective components, cognitive/metacognitive strategies, and resource management) (see Table 2). Only one dimension revealed a statistically significant difference, expectancy components, $t(36)$=-2.01, p=.05). The results indicated that the Fa15 class identified more strongly with items about positive self-efficacy than the Su14 class. None of the other dimensions significantly differed between groups. From the *t*-test analyses, we concluded that Su14 and Fa15 had similar scores on the MSLQ, with the exception of the self-efficacy dimension.

Table 2

Independent Samples t-test for the Motivated Strategies for Learning Questionnaire

Subscale	*T*	*df*	*p*	*d*
Motivation scale				
Value	-1.74	36	.09	0.57
Expectancy	-2.01	36	.05*	0.66
Affect	-.55	36	.60	0.18
Learning strategies scale				
Cognitive/Metacognitive	-0.85	36	.40	

strategies				0.28
Resources	0.70	36	.49	0.23

Note. * Significant at the $p<.05$ level.

To determine the relationship between scores on the MSLQ and average statistics performance, we conducted a Pearson correlation analysis for each semester (see Table 3). The mean score from each MSLQ dimension was compared to the average exam score for the whole semester. No significant relationship was found between any MSLQ dimension and statistics performance (average exam score) for either the Su14, or Fa15 semesters. Since average statistics performance could not be predicted by scores on the MSLQ, none of the dimensions were treated as covariates in subsequent analyses.

Table 3

Pearson Correlation for Motivated Strategies for Learning Questionnaire and Average Exam Score, Both Semesters

	Summer 14 Average exam score		Fall 15 Average exam score	
Motivation scale	*r*	*p*	*r*	*p*
Value	-.12	.66	.34	.14
Expectancy	-.05	.85	.30	.19
Affect	.24	.37	.11	.64
Learning strategies scale				
Cognitive/Metacognitive strategies	.36	.17	.06	.80
Resources	.20	.45	.04	.87

Note. * Significant at the p<.05 level.

Usage Characteristics

To determine how individual students interacted with the GSATs, we wanted to first identify any noticeable patterns of behavior. Using the GSAT data, four overall patterns of behavior emerged: (a) correct use/no change, (b) correct use/change, (c) incorrect use/repeated levels, and (d) incorrect use/attrition (see Table 4).

Table 4

Frequency Distribution and Percent of Each Usage Characteristic

		Frequency	Percent
Correct			
	Only easy	189	28.7
	Only medium	46	7
	Only hard	38	6
	Changed levels	81	12.3
Incorrect			
	1	15	2
	2	3	0.5
	3	31	5
	4	27	4
	5	9	1
	6	40	6
	Attrition	180	27.3
	Total	659	100

Note. Incorrect 1 - 6 indicate the number of difficulty levels per GSAT that were completed using incorrect parameters.

Correct use.

No change. Correct use/no change indicated that the student completed both Self-adapted Quiz parts A and B and chose to use the same difficulty level for both parts. The data indicated that 42% of the total number of GSATs that were taken over the semester

were taken using the same levels of difficulty for both part A and B (*n*=274*).* Of the group of GSATs that were classified as no change, there were three possible scenarios. First, the GSAT could have been taken using only the easy quizzes for part A and B. Second, the students could have used only the medium quizzes, and third, they could have used only the hard quizzes. Out of these three scenarios, taking the GSAT without changing difficulty levels using the easiest level was the most popular option (*n*=189). The second most popular choice was to take only the medium level (*n*=46), and the third most popular choice was to take only the hard level (*n*=38).

Change. The second example of GSAT behavior was correct use/change, which meant that the student completed both Self-adapted Quiz parts A and B and chose to use different difficulty levels for each part. The data indicated that 12.5% of the total number of GSATs were taken by changing the level of difficulty from part A to part B (*n*=81). Among the GSATs classified as correct use, changing levels of difficulty was the second most popular scenario.

Incorrect use.

***Repeated Levels*.** Out of the total number of GSATs, 18.6% were completed by students selecting from one to six levels of difficulty across both parts of the test (*n*=125). The most popular way to repeat levels was to take every level possible within a single GSAT, which was six (*n*=40). The second most popular scenario was to take three levels (*n*=31), followed by four (*n*=27) and five (*n*=9). Also included in this group are GSATs that were only halfway completed (one level of difficulty, *n*=15), and GSATs with two levels of difficulty completed within part A or part B only (*n*=3).

Attrition. The fourth example of GSAT behavior was incorrect use/attrition, which meant that a student did not complete the GSAT assignment in a given week. The data indicated that 27.3% of the total number of GSATs were left blank (n=180). One note about the attrition group is that it represented the second most frequent scenario when taking into account total GSAT use. Out of all the GSAT characteristics combined between correct and incorrect use, correct use/no change (easy only) was the most popular, and incorrect use/attrition was the second most popular.

Increased rates of attrition were significantly strongly related to later points in the semester, such that more GSATs were left incomplete at the end of the semester than at the beginning of the semester, r=.88, p<.001. Due to the high frequency of attrition found in this sample, the relationships between attrition and length of semester for the Su14 and Fa15 classes were compared. Since Su14 did not experience GSATs in the lesson program, the weekly assigned, fixed-item quizzes were used as a comparison. Results indicated that the Fa15 class had a stronger relationship between attrition and length of semester (r=.91, p<.001) than did the Su14 class (r=.64, p=.01). In other words, students opted out of the weekly quizzes and the GSATs at a higher rate in Fa15 than students who opted out of weekly quizzes in Su14.

Collapsed groups. To facilitate interpretability, we collapsed the four usage characteristics into three groups for some of the analyses. These groups were (a) correct, (b) repeated levels, and (c) attrition (see Table 5). Slightly over half of the total GSATs were completed correctly (54.7%) by students either changing or not changing levels of difficulty between parts A and B. Following correct usage, the second largest group

(attrition) was composed of GSATs that were not taken by students, earning a score of zero (27.3%). Finally, the repeated levels group was composed of GSATs that had between one and six levels completed across parts A and B. This group constituted approximately one fifth of total GSAT behaviors.

Table 5

Collapsed usage characteristic groups

	Frequency	Percent
Correct	354	54.7
Repeated	125	19
Attrition	180	27.3
Total	659	100

Note. Numerical values indicate numbers of GSATs taken, not number of students who took the GSAT.

Motivation by subgroup. To further distinguish between groups of GSAT use characteristics we calculated the mean dimension scores from the MSLQ for three types of users, correct (change and no change, n=16), repeated levels (n=5), and attrition (n=4). To place individual students into these groups, two criteria had to be met. First, students had to demonstrate a given behavior for at least 8 out of 15 GSATs (i.e. correct users had to complete at least eight GSATs correctly, and attrition users failed to complete GSATs at least eight times). Second, students had to have completed the MSLQ (see Table 6).

When comparing the three subgroups, results indicated that the repeated levels group had consistently higher motivation scores on three out of the five dimensions (*value components: M*=3.69, *SD*=0.09, *d*=0.88, *expectancy components*: *M*=4.29, *SD*=0.18, *d*=1.82, *affective components: M*=3.90, *SD*=0.42, *d*=0.36) than the correct group (*value components: M*=3.41, *SD*=0.44, *expectancy components*: *M*=3.63, *SD*=0.48, *affective components: M*=3.69, *SD*=0.72). Therefore, repeated levels users valued statistics more, had higher self-efficacy and higher test anxiety on average than correct users. These results should be interpreted with caution due to the small sample sizes used in this analysis. It should also be noted that none of the students who qualified for the attrition group completed the MSLQ, so we were unable to analyze motivation data for this group.

Table 6

MSLQ scores broken down by use characteristics subgroups

	Correct	Repeated	Attrition	*d*
n	16	5	4	
Value	3.41 (0.44)	3.69 (0.09)	..	0.88
Expectancy	3.63 (0.48)	4.29 (0.18)	..	1.82
Affect	3.69 (0.72)	3.90 (0.42)	..	0.36
Cognitive/ Metacognitive	3.65 (0.62)	3.25 (0.53)	..	0.69
Resources	3.42 (0.80)	3.25 (0.71)	..	0.22

Note. MSLQ scores are represented as M(SD). Two dots (..)=missing data.

Attempted difficulty. To evaluate the longitudinal relationship between the correct use of GSAT and level of difficulty attempted, a Pearson correlation was conducted by analyzing the weekly average difficulty across 15 weeks. There was a significant, moderately strong, positive relationship between the serial position in the semester and the level of difficulty attempted $r(36)=.70$, $p=.004$, such that later times in the semester related to harder levels of attempted difficulty (see Figure 1). In other words, students who took the GSAT correctly began the semester by taking easier GSATs, and as the semester progressed they attempted more difficult levels.

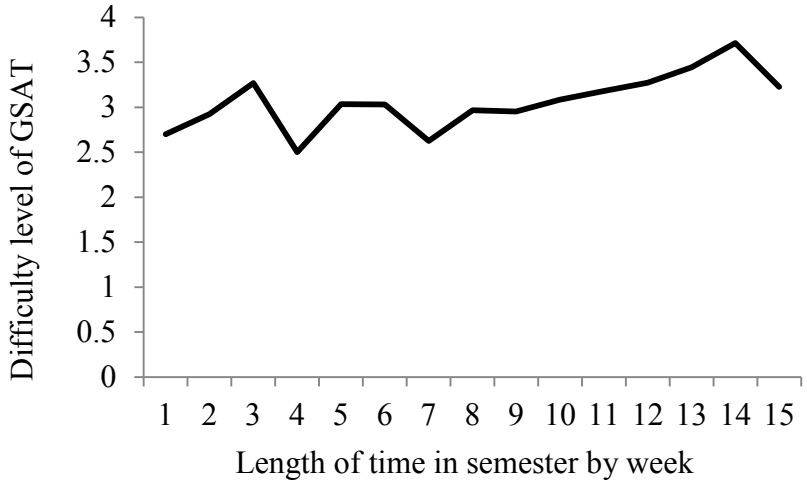


Figure 1. Correct usage attempted GSAT difficulty over time. This figure illustrates the level of difficulty attempted by correct users. Higher numbers on the y-axis indicate more difficulty.

To evaluate the relationship between incorrect use/repeated levels and level of difficulty attempted, we calculated the Pearson correlation coefficient of the within group average difficulty level across 15 weeks. There was a moderately weak, negative

relationship between serial position in the semester and the attempted difficulty levels, such that later times in the semester were associated with lower levels of difficulty attempted, but this relationship was not significant $r(15)=-.28$, $p=.31$ (see Figure 2).

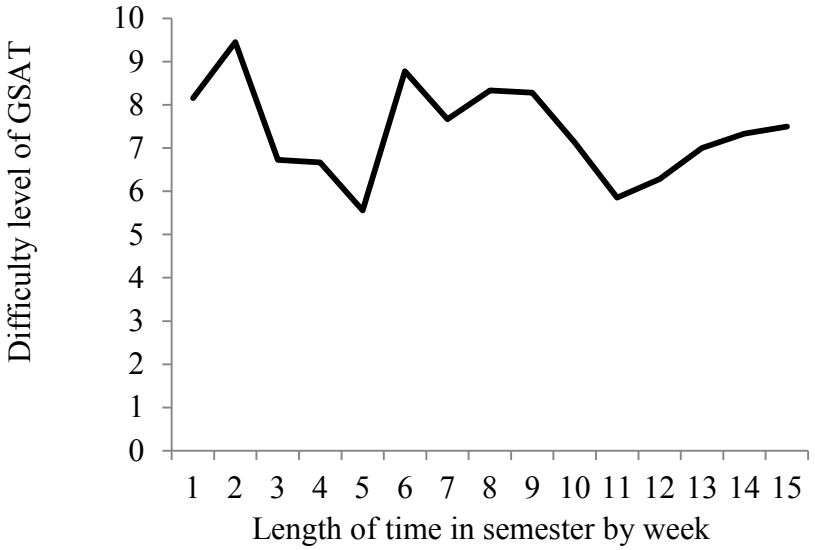


Figure 2. Incorrect usage/repeated levels attempted GSAT difficulty over time. This figure illustrates the level of difficulty attempted by repeated level users. Higher numbers on the y-axis indicate more difficulty.

To evaluate the overall longitudinal relationship (both incorrect and correct use cases) between using the GSAT and level of difficulty attempted, a third Pearson correlation was computed. Results indicated a significant, strong negative correlation between the GSATs serial position in the semester and the level of difficulty attempted, $r(43) = -.87$, $p < .001$. Overall, the entire class tended to take easier GSATs as the semester progressed (see Figure 3).

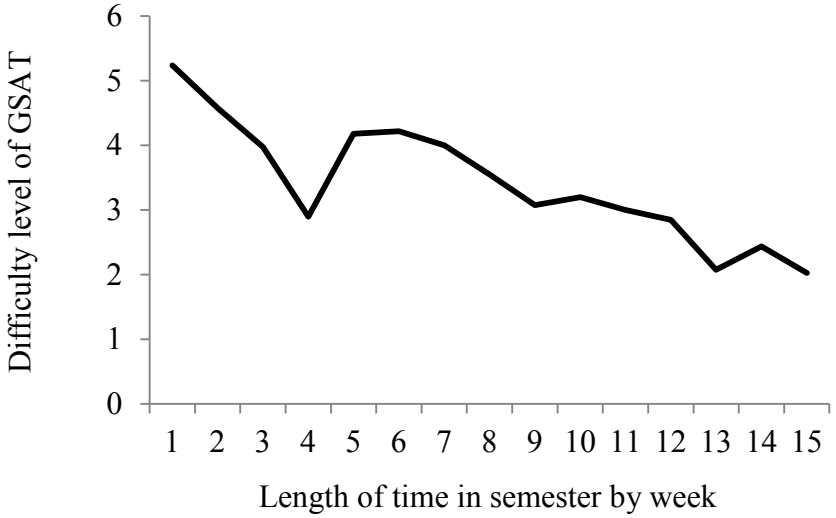


Figure 3. Total usage attempted GSAT difficulty over time. This figure illustrates the level of difficulty attempted by all users. Higher numbers on the y-axis indicate more difficulty.

Statistics performance

To assess the relationship between individual use characteristics (correct vs. incorrect) and average statistics performance (average percent correct on three exams), descriptive statistics and an independent samples *t*-test were calculated. The *t*-test indicated that average exam scores were higher for the correct group (M=76% SD=15%) than for the incorrect group (M=64%, SD=17%), $t(37)=2.30$, $p=.03$, $d=0.78$. We also compared characteristic subgroups (correct, repeated levels, attrition) and found that students who used the GSATs correctly over 50% of the time earned higher average exam scores (n=27, M=77%, SD=16%, $d=0.80$) than students who repeated levels over 50% of the time (n=5, M=64%, SD=16%), and students who did not complete any GSATs over 50% of the time (n=4, M=57%, SD=31%, $d=0.78$).

To test the hypothesis that prolonged use of GSATs can contribute to higher statistics performance, an analysis of variance between two groups (Su14 and Fa15) was conducted (see Figure 4). There was a significant main effect of serial position of exam, $F(2,183)=8.16$, $p < .001$, such that the later the exam was in the semester, the lower the exam score was. However, the main effect of semester was non-significant, $F(1,183)=0.59$, $p=.44$. The interaction effect was also non-significant, $F(1, 183)=0.24$, $p=.79$, indicating that the serial position of exam effect was the same for both semesters. In other words, the GSAT intervention did not have any noticeable positive or negative effects on cumulative exam performance at the $p=.05$ level.

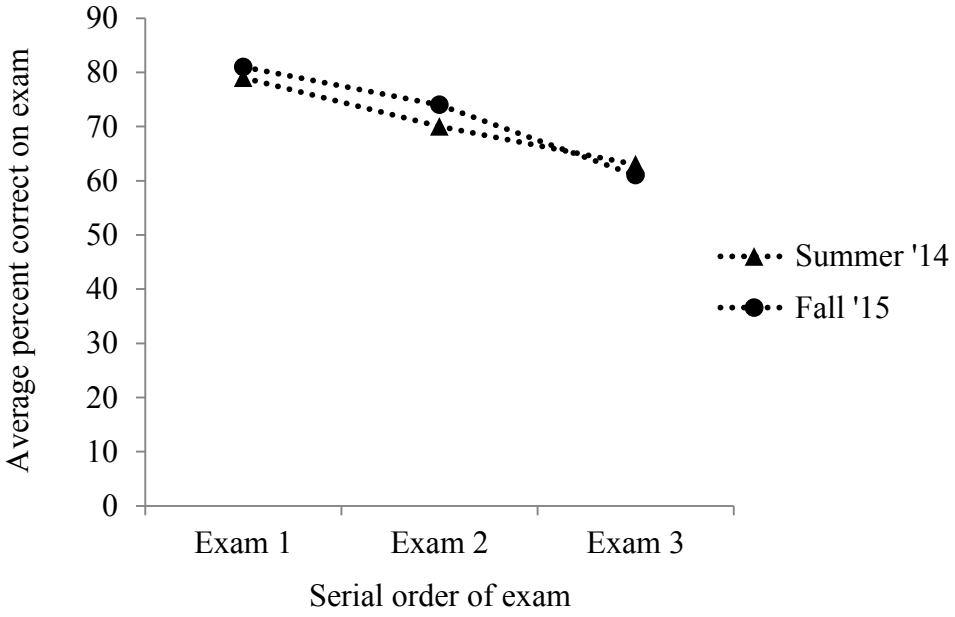


Figure 4. ANOVA of exam score by exam position and semester taken. This figure illustrates the average exam scores of Su14 and Fa15 for each of three exams.

However, to determine whether or not the correct usage of GSATs resulted in numerically higher exam scores than the baseline condition, descriptive statistics were computed. In Fall 2015, the students who took GSATs correctly more than 50% of the time earned higher exam scores on average (M=77%, SD=15%, d=0.46) than did students in Summer 2014 (M=72%, SD=12%).

There were two GSAT behaviors that characterized the correct usage group, (a) no change, and (b) change. Exam performance was compared between the no change and change groups to determine whether or not one group performed better than the other. First, individual students were grouped according to the criterion that they demonstrated a given behavior for at least 8 out of 15 GSATs (i.e. no change users completed at least eight GSATs without changing levels between part A and B, change users completed at least eight GSATs by choosing different levels for part A and B).

Descriptive statistics were calculated for the no change and change groups. Results indicated that the no change group earned higher scores on average (n=19, M=77%, SD=15%, d=0.22) than the change group (n=3, M=70%, SD=18%). Note that these results should be interpreted with caution due to small sample size comparisons.

Discussion

MSLQ

Since the premise of using the GSAT intervention was that students differ in levels of academic motivation and test anxiety, it was important to measure these variables, and identify potential nuisance variables. For example, if one semester indicated significantly higher test anxiety than the other, it would be important to

investigate the level of test anxiety as a potential covariate. When the two semesters were compared, there were no significant differences between them on any dimension (value, expectancy, affective, cognitive/metacognitive strategies, and resource management) of the MSLQ except for one, expectancy components. However, none of the dimensions emerged as significant predictors of average statistics performance. The MSLQ was essentially not correlated with statistics performance, so we decided not to include expectancy components as a covariate. It is possible that small sample size could have affected the power of these results.

Once different usage characteristics were established, scores on the MSLQ dimensions were used to compare the correct, repeated levels, and attrition subgroups within the Fall 2015 class. From the data we had available, it appeared that the repeated levels group experienced higher levels of valuing statistics, belief in personal statistics ability, and were more test-anxious than the correct group. In addition, the repeated levels group indicated lower use of cognitive/metacognitive strategies (i.e. self-regulation, rehearsal, etc.), and lower resource management skills (i.e. studying in a quiet place, or having a designated time to study each day). Furthermore, it was impossible to compare either group to the attrition group because no one who qualified for the attrition group volunteered to complete the pre-class MSLQ for extra credit.

If taken at face value, these comparisons suggest that even though the repeated levels group did not follow the instructions for the GSATs, which was technically incorrect usage, these students could have been the most motivated and high-effort students in the class. This interpretation is further supported by the tendency of repeated

levels users to complete more than what was required of them. It may be the case that these students wanted more practice with the material, so they chose to practice more repetitively which is a strategy for learning.

Furthermore, the correct users group demonstrated higher levels of cognitive/metacognitive strategy use and resource management compared to the repeated levels group. Since completing part A and part B of the GSAT only once is more efficient than repeating each part multiple times, it is possible that the correct users group was composed of individual students who prioritized maximizing benefit while minimizing cost. Another source of evidence for this interpretation is the fact that by far the most popular GSAT scenario for the entire class was correct use/no change/only easy. In other words, these students may have felt that the quickest way to get the most points was to take only the easy GSATs and not to bother with adaptation. Finally, the attrition group provided no motivation data whatsoever. One could extrapolate that this group had the lowest motivation of the entire class. A hallmark of the attrition group was that they received the lowest exam scores, which could also be indicative of low motivation (low effort).

It is extremely important in this particular study that the interpretations of the MSLQ dimensions and their reflections on different subgroups are taken lightly. To analyze within group differences with any degree of certainty is impossible due to the sample size we used. It is possible that the characteristics we found in this study could be used as a reference for a future study. It would be interesting to see whether or not the same conclusions can be replicated.

Usage Characteristics

Correct use. The group designated correct/change, or the group of GSATs that were taken correctly by students who consciously adapted the levels of difficulty within the test, demonstrated an interesting trend. Although this scenario was the least likely to occur, we observed when students chose this behavior that the attempted difficulty levels within GSATs increased over time. This trend is further underscored by the fact that it runs counter to the overall trend of the entire class. When analyzed together, the class decreased attempted difficulty levels over time. These results suggest possible qualitative differences between the students who took the tests correctly versus incorrectly.

Another observation we made about the correct users was that the majority of correct/no change users preferred to take GSATs from the easiest level of difficulty. From the outset, one of the goals of including easy tests was to avoid negatively punishing those who needed extra help with the material. The same concepts from all the hard questions were reflected in the easy questions, they were simply subdivided into smaller, somewhat more repetitive tasks. The repetitive, prolonged engagement with the easy tests may have given the students enough practice without causing so much pressure that the task was avoided altogether.

By grouping no change, and change users together under the heading of correct users, an assumption is being made about the level of personal control that these students used to complete the GSATs. On one hand, if students decided to change levels of difficulty between part A and part B, then we might think of them as truly adapting the test, while students who kept the same level for both parts were merely taking a fixed

difficulty test. On the other hand, there were many users in the no change category that varied the level of GSAT difficulty by week, such that one week they might have taken part A/easy, part B/easy, and the following week they took part A/hard, part B/hard. In cases such as these, no change users were actually exercising the adaptive capabilities of the GSAT. Therefore, we assumed that both no change and change methods of taking the GSATs were both correct.

Finally, the group of students who used the GSATs correctly for the majority of the semester earned the highest cumulative exam scores in the class. One way to interpret this finding is that correct GSAT usage might have contributed to higher exam scores. However, an alternative explanation is that students with higher ability in statistics self-selected themselves into the correct use group. This could have occurred because high-ability, high-performing students are better at completing tasks or reading instructions. Without further analyses, this issue is left an open question.

Incorrect use. The most unexpected finding was the inordinately high number of GSATs that were taken incorrectly. Roughly half, or 46.3%, of the GSATs were not taken according to the parameters of the instructions we provided, or the theoretical construct we were using. Unfortunately, a flaw in the software configuration of the GSATs allowed for a considerably large group of students to take the GSATs repeatedly (incorrect, repeated levels). In a typical computerized test, the test maker can use a configuration option that disallows students to retake a test after they have already taken it. The student is essentially locked out of the test, and whatever score they got becomes their final score. However, in the GSAT design used for this study there was no

configuration option within Canvas to lock out a student from a test that they had not already taken. Therefore, if a student wanted to take Self-adapted Quiz 1A on the easy level, there was no way to prevent them from also taking Quiz 1A on the medium level and the hard level as well. For this reason, the instructions for the GSAT were written specifically to illustrate the concept that only one difficulty level for part A, and only one level for part B was necessary. It is unclear if these students did not understand the instructions given to them, or they simply chose to ignore them.

It was unclear how to analyze this group of data, because on one hand, there was no way to measure any trends according to the criterion we had established (i.e. having personal control over the levels of difficulty). On the other hand, if so many GSATs were taken with three or more levels of difficulty, then by our definition, this group demonstrated very high effort test taking behavior.

Regarding attrition, it is fairly common for the rate of homework completion to decrease as students get further into the semester. As the number of incomplete GSATs increased, the weekly average level of difficulty decreased for the entire class. This can partially explain why the level of attempted difficulty was significantly negatively related with time spent in the semester. A possible reason for the high rate of attrition could be that although the GSATs were compulsory, they were ultimately low-stakes assignments. They were worth a fraction of the points that some other homework assignments were worth; therefore it may have been easy for some students to justify not taking them. Another explanation that must be considered is the overall usability of the GSATs. If students became frustrated because the GSATs were difficult to use, they may have

chosen to opt out of the assignments. Since post-class data was not collected, that is a question that must be saved for a future study.

Another point to consider comes from the comparison of Fa15 attrition and Su14 attrition. We found that Fa15 students were far more likely to opt out of taking both the GSATs and the weekly quizzes than the Su14 students. For this reason, it is possible that the GSATs were not solely responsible for the drop in completed homework towards the end of the semester. Rather, it seems like the Su14 class could have been more resistant to attrition due to some qualitative difference. For example, summer online enrollment at SJSU often sees higher numbers of students who either (a) need to repeat the class after failing it once before, or (b) need the class to fulfill a graduation requirement. Therefore, summer students could potentially be more risk averse than fall students.

Statistics performance

One of the primary questions that we wanted to answer with this study was whether or not employing a GSAT program longitudinally would be beneficial to statistics performance, measured by exam performance. In our comparison of the Su14 (non-GSAT), and the Fa15 (GSAT) conditions, we were unable to detect any differences in statistics performance. Therefore, we failed to reject our null hypothesis. This result does not support a strong argument in favor of using GSAT, however it does show that our GSAT program was at least as good as a more traditional online class curriculum.

Given the amount of variability in the way Fa15 students engaged with the GSATs and the high amount of attrition, inferential statistics and hypothesis tests are probably of limited usefulness in this study. However, we were able to determine

through descriptive statistics that correct GSAT users earned the highest exam scores of the class during the Fa15 semester, and also earned slightly higher exam scores than the Su14 semester. Although the reliability of our results is questionable, we consider this result encouraging enough to warrant further exploration in this avenue of research.

Limitations

By choosing to use Canvas as the sole platform for course content delivery, the GSAT design was affected by the constraints of that particular system. For example, there was no way to make the GSATs adaptable at the item level. True SATs offer examinees difficulty options each time they answer a question. Canvas was incapable of applying pattern logic to each item, therefore we created a workaround to make the quizzes adaptable at the quiz level (part A and B). Students were given the opportunity to make their difficulty choice only twice instead of multiple times. There are positive and negative aspects to this approach. One positive aspect is that unlike CATs or SATs, item review was possible. In our GSAT, examinees could review their answers to quiz items as many times as they wanted before submission. One negative consequence of dividing each GSAT into two parts was that there may have been some confusion about how the quiz was supposed to work. There was no formal feedback system to indicate whether or not the students were using the quizzes correctly. There could have been errors in GSAT use due to misunderstanding the instructions on the syllabus. Furthermore, there was no systematic way to prevent errors. In order to reduce errors, we would have implemented item logic, making it impossible to take more than one level of difficulty from either part A or B.

Another limitation to this study was the lack of independent validation of the items used in the GSATs. Were the hard quizzes really twice as hard as the medium quizzes? In the data, they were worth twice as much, but that does not indicate whether or not students attempting harder difficulty GSATs were better at the subject material. In addition, there is no way to know if the students who attempted the easy GSATs learned an equivalent amount of to that of the hard GSATs. That was an assumption we made. For a future study, it would be important to validate the quiz content more vigorously by consulting multiple subject matter experts.

There were quite a few differences between the two courses that were compared, so it is difficult to interpret the results of this study with respect to any single intervention. For one thing, a statistical analysis of variance has more power if the subjects are evenly distributed across different groups. In our sample, the Fa15 group (n=38) outnumbered the Su14 group (n=16) by a considerable amount. Furthermore, the likelihood of unexplainable variance increases when sample sizes are small. For a future study, we would recommend that larger samples be studied with greater parity between sample sizes. One way to do this would be to repeat the study and collect data over multiple semesters.

From a usability perspective, the computer graphical interfaces were not the same between courses. In Su14, students had to access two different websites in order to watch the lesson videos and complete homework. However, Su14 had comparatively less homework to complete. In contrast, Fa15 had the simplicity of using a single learning management system for the entire course, yet their workload was higher due to the

introduction of the GSATs. We did not measure how labor intensive it was to access the videos or complete the homework. It is possible that heuristic difficulty and frustration could have influenced how students used the GSATs.

Future Directions and Conclusions

Going forward, it would be useful to evaluate the desirability of self-adapted testing among online students. One way to accomplish this would be to issue a post-class questionnaire asking students to identify their opinions of the testing environment. Another way to approach the question of desirability could be to measure how much students value an adaptable testing environment. By making the self-adapted quizzes optional, we could gain insight on the motivating factors behind organic adoption of an SAT program. In contrast, we could keep the SATs as a compulsory part of the course, but make them worth more points towards the overall grade.

Another potentially fruitful area of research may require interdisciplinary work between psychologists, educators, and software developers. In the future, software functionality may expand such that the configuration options for online courses become more sophisticated. Today, item logic within Canvas is not possible, but the technology does exist, and has existed for some time. There could be great potential for fully adaptable online courses, but reaching a better understanding of adaptability will take effort and open-mindedness. With such high demand for fully online, or hybrid university courses, education technology could become an important focus for research.

Based on this study alone, we cannot recommend the Repeated GSAT program to online university professors or lecturers. The amount of time that it would cost someone

to build a fully adaptable question bank to allow students to choose from multiple levels of difficulty is simply prohibitive. Using the GSAT program did not raise statistics students' scores to a satisfactory enough level to justify implementing the specific procedure that we used. However, if similar research could eventually verify the usefulness of SAT as a performance improving tool, then the time consuming nature of building the question banks would be less prohibitive. The reason is that computerized tests, whether they are used in an online class or a face-to-face class, are fully scalable and automated. After the initial investment of time is made to create the tests, the tests require very little monitoring. In reality, classes with 25 students or 25,000 students could all use this testing strategy with the click of a button. Creation of tools, and improvements upon tools we already have, could serve future students in ways that were previously impossible.

References

Averill, J. (1973). Personal control over aversive stimuli and its relationship to stress. *Psychological Bulletin. 80*, 286-303.

Baloglu, M. (2002). Psychometric properties of the statistics anxiety rating scale. *Psychological Reports*, *90*, 315-325.

Bandura, A., Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*, 586-598.

Carlson, K. A., Winquist, J. R. (2014). An introduction to statistics: An active learning approach. Thousand Oaks, California: Sage Publications, Inc.

Elliot, A. J., & McGregor, H. A. (2001). A 2× 2 achievement goal framework. *Journal of personality and social psychology*, *80*, 501.

Hendricks, D. J., Walls, R. T., Heiman, G. W. (2004). Essential statistics for the behavioral sciences. Boston, Massachusetts: Houghton Mifflin Company.

Howell, D. C. (2011). Fundamental statistics for the behavioral sciences: Instructor's manual with test bank. Belmont, California: Wadsworth, Cengage Learning.

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, *20*, 975-978.

Morris, L. W., & Liebert, R. M. (1973). Effects of negative feedback, threat of shock, and level of trait anxiety on the arousal of two components of anxiety. *Journal of Counseling Psychology*, *20*, 321-326. doi:10.1037/h0034768

Musch, J., & Bröder, A. (1999). Test anxiety versus academic skills: A comparison of two alternative models for predicting performance in a statistics exam. *British Journal of Educational Psychology*, *69*, 105-116. doi:10.1348/000709999157608

Onwuegbuzie, A. J. (1998b) The dimensions of statistics anxiety: a comparison of prevalence rates among mid-southern university students. *Louisiana Educational Research Journal*, *23*, 23-40.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology. 95*, 667-686. doi: 10.103/0022-0663.95.4.667

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a Statistics Attitude Survey. *Educational And Psychological Measurement*, *40*, 235-238. doi:10.1177/001316448004000138

Rocklin, T. (1994). Self-adapted testing. *Applied Measurement in Education, 7*, 3-14.

Rocklin, T. & O'Donnell, A. M. (1987). Self-adapted testing: a performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. Journal Of *Personality And Social Psychology*, *46*, 929-938. doi:10.1037/0022-3514.46.4.929

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement. 35*, 328-347.

Vispoel, W. P., Coffman, D. D. (1994). Computerized-adaptive and self-adapted music-listening tests: psychometric features and motivational benefits. *Applied Measurement in Education, 7*, 25-51.

Wigfield, A., Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25,* 68-81. doi: 10.1006/ceps.1999.1015.

Wilson, J. H. (2005). Essential statistics: Instructor's manual with tests. Upper Saddle River, New Jersey: Pearson, Prentice Hall.

Wise, S. L. (1994). Understanding self-adapted testing: the perceived control hypothesis. *Applied Measurement in Education, 7*, 15-24.

Wise, S. L., DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., Roos, L. L., Plake, B. S., Nebelsick-Gullet, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education, 7*, 81-91.

Wolf, L., Smith, J. K., Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. Applied Measurement in Education. 8, 341-351.

Appendix A

IRB Approval Letter

SJSU SAN JOSÉ STATE UNIVERSITY | Office of Research Division of Academic Affairs | San José State University One Washington Square San José, CA 95192-0025 | TEL: 408-924-2272 sjsu.edu/research

To: Gita Hodell

From: Pamela Stacks, Ph.D.
Associate Vice President
Office of Research

Pamela C Stacks

Date: December 17, 2015

The Human Subjects-Institutional Review Board has registered your study entitled:

"The Effects of Repeated Global Self-Adapted Testing on Online Academic Performance"

This registration, which provides exempt status under Exemption Category 4 of SJSU Policy S08-7, is contingent upon the subjects included in your research project being appropriately protected from risk. Specifically, protection of the anonymity of the subjects' identity with regard to all data that may be collected about the subjects from your secondary sources needs to be ensured.

This registration includes continued monitoring of your research by the Board to assure that the subjects are being adequately and properly protected from such risks. If at any time a subject becomes injured or complains of injury, you must notify Dr. Pamela Stacks, Ph.D. immediately. Injury includes but is not limited to bodily harm, psychological trauma, and release of potentially damaging personal information. This approval for the human subject's portion of your project is in effect for one year, and data collection beyond December 17, 2016 requires an extension request.

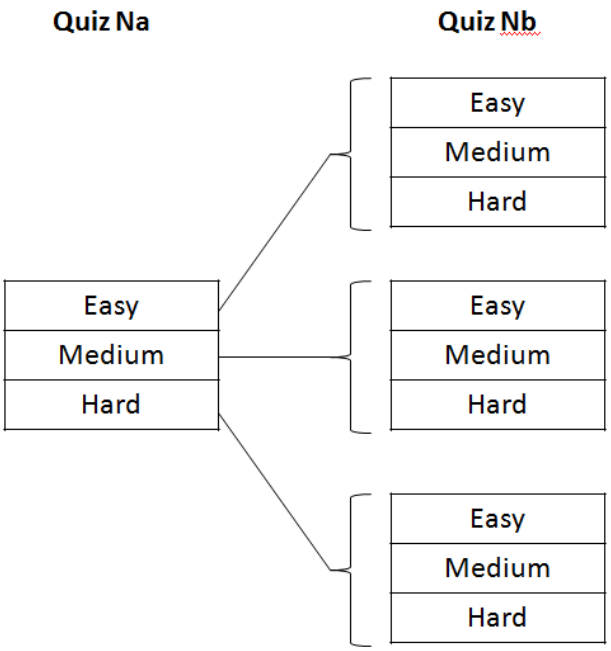
If you have any questions, please contact me at (408) 924-2427.

Protocol #: S15170

cc. Sean Laraway 0120

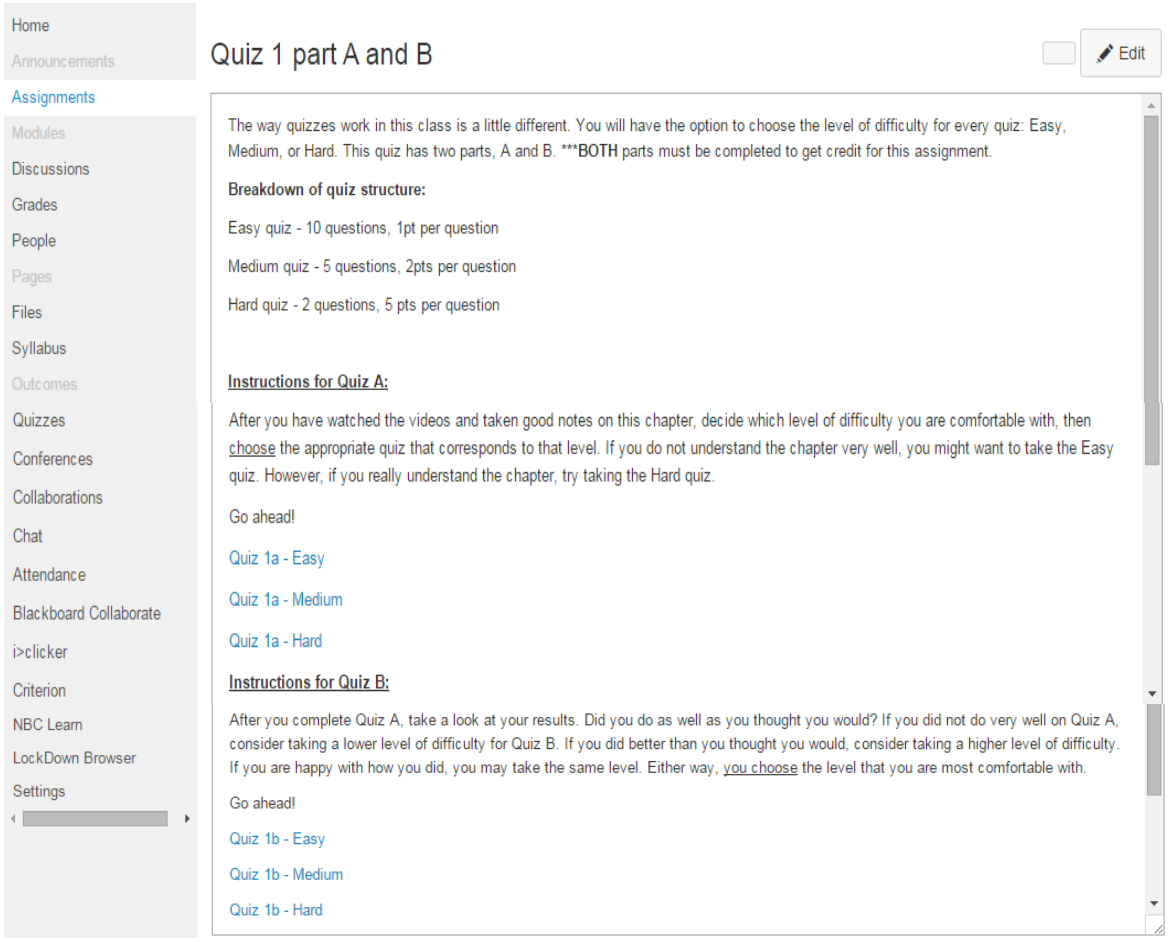
Appendix B

GSAT Configuration Options



Appendix C.

GSAT Written Instructions



Appendix D

Motivated Strategies for Learning Questionnaire

For the following items, please indicate how well each statement describes you as a student (where 1 = Not at all true of me; 3 = Somewhat true of me; 5 = Very true of me). There are no right or wrong responses - only different ones. Please respond to all of the items.

1. I prefer class work that is challenging so I can learn new things.
2. Compared with other students in this class I expect to do well.
3. I am so nervous during a test that I cannot remember facts I have learned.
4. It is important for me to learn what is being taught in this class.
5. I like what I am learning in this class.
6. I'm certain I can understand the ideas taught in this course.
7. I think I will be able to use what I learn in this class in other classes.
8. I expect to do very well in this class.
9. Compared with others in this class, I think I'm a good student.
10. I often choose paper topics I will learn something from even if they require more work.
11. I am sure I can do an excellent job on the problems and tasks assigned for this class.
12. I have an uneasy, upset feeling when I take a test.
13.1 think I will receive a good grade in this class.
14. Even when I do poorly on a test I try to learn from my mistakes.
15. I think that what I am learning in this class is useful for me to know.
16. My study skills are excellent compared with others in this class.
17. I think that what we are learning in this class is interesting.
18. Compared with other students in this class I think I know a great deal about the subject.
19. I know that I will be able to learn the material for this class.
20. I worry a great deal about tests.
21. Understanding this subject is important to me.
22. When I take a test I think about how poorly I am doing.
23. When I study for a test, I try to put together the information from class and from the book. 24. When I do homework, I try to remember what the teacher said in class so I can answer the questions correctly.
25. I ask myself questions to make sure I know the material I have been studying.
26. It is hard for me to decide what the main ideas are in what I read.
27. When work is hard I either give up or study only the easy parts.
28. When I study I put important ideas into my own words.
29. I always try to understand what the teacher is saying even if it doesn't make sense.
30. When I study for a test I try to remember as many facts as I can.
31. When studying, I copy my notes over to help me remember material.

32. I work on practice exercises and answer end of chapter questions even when I don't have to. 33. Even when study materials are dull and uninteresting, I keep working until I finish.
34. When I study for a test I practice saying the important facts over and over to myself.
35. Before I begin studying I think about the things I will need to do to learn.
36. I use what I have learned from old homework assignments and the textbook to do new assignments.
37. I often find that I have been reading for class but don't know what it is all about.
38. I find that when the teacher is talking I think of other things and don't really listen to what is being said.
39. When I am studying a topic, I try to make everything fit together.
40. When I'm reading I stop once in a while and go over what I have read.
41. When I read material for this class, I say the words over and over to myself to help me remember.
42. I outline the chapters in my book to help me study.
43.1 work hard to get a good grade even when I don't like a class.
44. When reading I try to connect the things I am reading about with what I already know.