8-27-2012

The relationship of quizzing and student success in a college level core statistics course

David Glavin

David M. Glavin

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Kristin Umland, Ph.D., Chairperson

Michael Sonksen, Ph.D.

Michael Nakamaye, Ph.D.

THE RELATIONSHIP OF QUIZZING AND STUDENT SUCCESS IN A COLLEGE LEVEL CORE STATISTICS COURSE

by

DAVID M. GLAVIN

B.S., BIOLOGICAL SCIENCES, UNIVERSITY OF VERMONT, 1991

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Statistics

The University of New Mexico
Albuquerque, New Mexico

July, 2012

DEDICATION

I would like to dedicate this paper to my family, especially my parents Thomas and Elizabeth, whose love, support, guidance and unshakable confidence in me were essential to my success in this endeavor. I would also like to thank my grandmother, Marie, who always had words of encouragement for me when things seemed impossible. All of you helped me get through each day of this journey. I can never thank you enough.

Finally I would like to thank Brandy DiMaggio, who helped me in so many ways during the development and critical thinking process for this project. I could not have hoped for a better friend and unconditional source of support during this experience.

ACKOWLEDGEMENTS

I would like to acknowledge my advisor Dr. Kristin Umland who assisted me in ways too numerous to count. If not for her taking a genuine interest in this project I may, very well, not have completed this degree program. I don't think she realized that taking me on as an advisee was not just limited to thesis advisement, but also required a degree in psychology and positive thinking. I am fortunate to have found a mentor who understands the effort and dedication it takes to succeed at the graduate level and who pushed me to perform at levels beyond what I thought I was capable of. Her knowledge, sincerity, insight and confidence in my abilities were critical to the completion of this research project.

Fares Qeadan, who taught me how to program and was influential in assisting me during the analysis of this research project. More importantly, Fares was like a brother to me during this project. The sun did rise that next day, my friend.

Dr. Michael Sonksen, for guidance with statistical theory as well as numerous discussions about college football, which distracted me from the seemingly overwhelming task I faced each day. I look forward to the next ND-OSU game.

Julianne Gearhart, for helping me start this journey and more importantly, believing in me.

Finally, I would like to thank all of my peers and the staff in the Mathematics and Statistics Department at UNM.

The Relationship of Quizzing and Student Success

In A College Level Core Statistics Course

by

David M Glavin

B.S., Biological Science, University of Vermont, 1991

M.S., Statistics, University of New Mexico, 2012

ABSTRACT

This study investigated the relationship between quizzing and student success in an introductory college level statistics course. Demographic and student performance data were collected from a 100-level introductory Statistics course at the University of New Mexico during the Fall 2011 semester. Two statistical models were developed to determine if quizzing is related to student success as measured by final letter grades and final exam scores. Predictive modeling to determine the relationship between quizzing and students' final exam scores using a Hierarchical Linear Model (HLM) found quizzing to be marginally significant (*p*-value = 0.0567). Probabilistic modeling using logistic regression to predict if a student passes the course with a grade of C or higher yielded an odds ratio of 6.013 (95% Wald CI: 2.030, 17.813) for students who were given periodic quizzes versus students who were not given quizzes, while holding all other

variables in the model constant. Results indicate that quizzing is positively associated with student performance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Many students who enter college are underprepared for college-level mathematics and statistics, and lack of student success in these core disciplines both before and during college is well documented (ACT Inc., 2010). There have been many efforts in the U.S. to improve both secondary and post-secondary education in mathematics and statistics; the No Child Left Behind Act of 2001 (NCLB) is a recent major policy and funding initiative that intends to increase student achievement in mathematics and other domains. One of the key aspects of NCLB policy is that it links funding for K-12 schools with student attainment in mathematics as measured by standardized test scores. Although there is evidence that well designed standardized tests are useful tools for assessing students' understanding of general concepts in mathematics, English, reading comprehension, etc., (Volante, L., 2004) one of the primary goals of educators is to maximize student development in critical thinking skills and mastering core concepts of all subjects studied. However, because the scores that students earn on these standardized tests have such a great impact on teachers and schools, critics of NCLB argue that secondary educators are "teaching to the test" (Volante, L., 2004) rather than focusing on methods for improving student learning, critical thinking skills, and subject matter retention. Although NCLB primarily impacts secondary education administrators, educators, students and their parents (Heath, S., 2002), the effects of NCLB are also felt at the university and college level. Increasingly, state and national policy makers as well as students and parents are asking whether the

money spent on higher education results in the student learning outcomes that students need to be successful in their careers. As a result, many colleges and universities are using standardized measures to gauge student learning.

Several core courses in mathematics and statistics taught at the University of New Mexico (UNM), including Introduction to Statistics (STAT 145), use a "standardized" testing system to assess student performance. The methodology is considered "standardized" in that the same mid-term and final exams are administered to all the students taking these courses regardless of instructor. Because instructors are judged, at least in part, on their students' passing rates, university instructors teaching these courses face the same challenge that their counterparts in secondary education encounter.

The resulting conflict of balancing the need to improve students' scores on standardized exams without compromising the depth of understanding of core concepts of subject matter leads to the question:

> What pedagogical methods should instructors employ that develop students' core content knowledge and foster critical thinking skills so that they can use what they've learned across disciplines?

The use of "low- or no-stakes testing" (i.e. quizzing) is one technique employed by instructors for assessment and as a method to improve learning and retention of course content (McDaniel, M.A., et al., 2011). The use of quizzes is a tool that not only can assist instructors to assess how well their class is grasping a particular concept, but

also assist the student in achieving academic success in the course with minimal impact to the overall final grade by requiring them to study quiz-related material.

As a first-year instructor of introductory statistics at UNM (STAT 145), the Principle Investigator was interested in developing methods of instruction that increase student comprehension and understanding of core competencies identified by the University. Due to the fact that STAT 145 is a core mathematics/statistics course, the Department of Mathematics and Statistics attempts to employ uniform assessment of student achievement. This is done by requiring all section instructors to administer three midterm exams, each worth 20 % of students' final grade plus a cumulative final exam worth 25% of the final grade. The remaining 15% is left to instructor discretion. Being new to teaching, the Investigator decided to weight all four exams equally (25% per exam) by dispersing the discretionary scoring to each of the three mid-term exams during the first semester of instruction in the Fall of 2010.

In Spring 2011 the Investigator used in-class quizzing to account for the fifteen percent discretionary scoring and noted a marked increase in students' pass rate (a grade of "C" or higher, with the passing score set by the instructor), as well as an increase of the minimum overall score required to pass the course (refer to Table 1.1) In particular, the score for receiving a grade of "C" in the course increased by 5% and the percent of students passing the course increased by 13% from the previous semester, when quizzes were not administered.

Table 1.1
Pass Rates and Minimum Passing Score
For Investigator's First Two Semester of Instruction

Semester	Lowest Score for "C"	Percent Pass
Fall 2010	60%	68%
Spring 2011	65%	81%

The Investigator felt this observation implied that quizzing may have positively impacted student success in STAT 145, while noting that simply becoming familiar with instructional methods and classroom management may have had some impact as well. To determine if the implementation of periodic quizzing was, in fact, influencing student success in STAT 145, the Investigator decided to conduct an observational study using students from the Fall 2011 semester at UNM to quantitatively assess if the use of quizzing in an introductory college level statistics course is related to student success.

BACKGROUND AND LITERATURE REVIEW

The education literature expoloring the relationship between quizzing and academic achievement has generally focused on primary and secondary education (Agarwal et al., 2010; Lloyd, 1995). In addition, with the recent increase in use of Computer/Student Response Systems (CRS/SRS), commonly referred to as "clickers", and online assessment tools such as WebCT or Blackboard Learning Systems, some

researchers have investigated whether the use of these devices to administer quizzes has a positive impact on academic performance (Morling et al., 2008; Urtel et al., 2006).

No studies were found in the literature to have been conducted on quizzing and academic success in large-scale, undergraduate statistics or mathematics courses. Research on periodic quizzing and academic performance has been conducted for disciplines other than math and statistics, such as physics and psychology (Roediger, H. & Karpicke, J. 2006). A recent study comparing SRS and WebCT administered quizzes was performed using nursing students enrolled in "a spring 2004 General, Organic, and Biochemistry course," however the sample size was relatively small (n = 41). A limitation of these studies is that multiple linear regression methods were used to evaluate the relationship between explanatory and response variables, which may not account for correlation across instructors.

Educational studies are increasingly utilizing hierarchical linear models (HLM) to account for the influence of random effects correlated with instructors (Raudenbush, S., 1988; Lee, et al., 1991). The advantage of using hierarchical linear models arises from adjusting for random effects associated with instructors, prior to analyzing the effects of the explanatory variable of interest, i.e., quizzing (Raudenbush, S., 1988; Lee & Bryck, 1989). The Investigator assumes there are differences between instructors such as teaching style, experience, "good" instructors versus "bad" instructors, student satisfaction with instructors (Lee, et al., 1991) and other innate qualities associated with instructors that cannot be controlled for in an observational study. By adjusting for

variation between instructors, the Investigator hoped to more clearly identify the association between quizzing and student achievement in STAT 145.

In this observational study, the Investigator employed two models to assess the effectiveness of quizzing and its impact on students' success in a large-scale (50 students or more) undergraduate introductory statistics course. As previously indicated, a predictive modeling method utilizing an HLM was developed to determine the relationship between administration of quizzes and students' final exam score while adjusting for other confounding variables. In addition, a probabilistic modeling method was used to examine the likelihood of passing STAT 145 with respect to quizzing while adjusting for other confounding variables. The Investigator employed a logistic regression model (logit) to predict the likelihood a student will pass the course given that quizzes were administered, where passing is defined as receiving a grade of "C" or higher.

CHAPTER 2: METHODS

This thesis primarily focuses on two models to determine student success in a college-level introductory statistics course contingent on whether the student received periodic quizzing during the course. Data were collected on all students who were enrolled in the course after the first three weeks of the semester. The main outcomes considered were: (1) whether or not the student passed the course based on their final course grade (a passing grade is considered as C or higher); and (2) the score (out of 100) received on the cumulative final exam.

The two models developed to assess student success were: (1) a mixed or hierarchical linear model (HLM) used to estimate the value of a quantitative response variable; and (2) a logistic regression (logit) model which is used to predict the probability of an outcome for a categorical variable. The response variable for the HLM was the final exam score attained by each student, and the response variable for the logit model was whether a student passed the course with a grade of C or higher or not.

DATA SOURCE

This observational study uses student data collected from 16 sections of an introductory statistics course (STAT 145) during the Fall semester of 2011 at The University of New Mexico, Albuquerque, New Mexico. Data was collected from two sources:

(1) Information from instructors who agreed to participate in the study regarding quizzing policies and students' final exam scores, and

(2) Demographic and academic information for each student enrolled in sections taught by instructors who agreed to participate in the study from the University of New Mexico's Data Warehouse (institutional database).

Instructors (a mixture of graduate students, part-time instructors and lecturers), who taught Stat 145 during the Fall semester of 2011 were given a 30-minute presentation by the Investigator describing the study prior to commencement of instruction for the Fall 2011 term. Instructors who agreed to participate in the study submitted signed consent forms, as required by UNM's Internal Review Board (IRB), to the Investigator by August 26, 2011. Data was only collected from students whose instructor agreed to participate in the study.

Data was collected in two phases. Demographic and academic performance data, not specific to Stat 145, was collected from the UNM Data Warehouse following the third week of the Fall 2011 semester, the last date a student could withdraw from the course without a grade and the official census date for enrollment figures at UNM. Table 2.1 provides student information obtained from the UNM Data Warehouse.

Table 2.1
Explanatory Variables from UNM Data Warehouse

Variable	Type	Explanation
Gender	Explanatory	Male/Female
Ethnicity	Explanatory	Hispanic
		White
		Other (Asian, Black, Native American, etc.)
GPA	Explanatory	College GPA
GPA Credits	Explanatory	GPA Credits Earned at UNM
Credits Attempted	Explanatory	GPA Credits Attempted at UNM
Credits Earned	Explanatory	All College-Level Credits Earned
School GPA	Explanatory	High School GPA
ACT Math	Explanatory	ACT Math Score
SAT Math	Explanatory	SAT Math Score

At the end of the semester, instructors participating in the study were asked to complete a questionnaire to determine which instructors used quizzing during the term, how the quizzing was employed such as frequency, style (web-based, written), and weight of quizzes given in computing the final grade for the course (see Appendix A, *Post-Semester Questionnaire*). In addition, instructors were asked to submit a spreadsheet that included each student's final exam score and final letter grade in the course. Table 2.2 lists a summary of the information submitted by instructors.

Table 2.2
Instructor Submitted Data

Variable	Type	Explanation
Quiz/No Quiz	Explanatory	Received Quizzes
		Did not Receive Quizzes
Frequency	Explanatory	Number of Quizzes Administered
Time	Explanatory	Time of Section: AM/PM
Weight	Explanatory	Percent of Final Grade Each Quiz was Worth
Administered	Explanatory	How Quiz was Administered (written, web, clicker, take-home)
Final Exam Score	Response	Final Exam Score (out of 100)
Final Letter Grade	Response	Final Letter Grade Student Received in the Course

To maintain student anonymity and compliance with the Family Educational Rights and Privacy Act (FERPA), each student was issued a unique Research Identification Number (RIDN) prior to submission of data to the Investigator. Data obtained from the UNM Data Warehouse and submitted by each section instructor was merged using the RIDN.

Ten instructors agreed to participate in the study. Of these instructors, three taught multiple sections. The data set included 905 students from 16 sections taught by the 10 instructors to represent the population of students enrolled in Stat 145 during the Fall 2011 term. Of these observations, 35 were removed from the analysis because no final exam grade and no final grade were reported. Fifty-seven observations were removed because these students received a grade of "W" (withdraw), "WP" (withdraw pass), "WF" (withdraw fail) and did not take the final exam, indicating they did not complete the course. Of the remaining observations, 117 individuals did not report their Ethnicity; therefore these observations were removed prior to analysis. A total of n = 696 of the original 905 submitted observations were used for analysis in the HLM and logit models. For the remainder of this report all references to the sample size refer to the number of valid records used for analysis (i.e., n = 696).

EXPLANATORY VARIABLES

Gender

The gender of each student was used as an independent variable in both models of the study. Of the 696 observations, 428 (61.49%) were female and 268 (38.51%) were male.

Ethnicity

The distribution of ethnicity for students in the survey as supplied by the UNM Data Warehouse is presented in Table 2.3. To avoid issues related to small cell size, it was decided to categorize the ethnicity explanatory variable into three groups as Hispanic 262 (37.64%), White 275 (39.51%) and "Other" 159 (22.85%) where the latter included the Asian, Black, Native American and other reported ethnicities. Imputation was not performed for non-reported ethnic backgrounds.

Table 2.3
Separation Table of Students by Ethnicity

			OTHER				
	White	**Hispanic**	Asian / Pacific Islander	Black / Af. Amer.	Native American	Other Reported Ethnicity	**Total**
COUNT	275	262	29	23	46	61	n = 696
PERCENT	39.51%	37.64	4.17%	3.31%	6.61%	8.76%	100%

College GPA

Each student's UNM Grade Point Average (GPA) at the time they entered the course was collected from the UNM Data Warehouse. The 149 (21.41%) missing values associated with these observations were deemed to be first semester freshmen, (individuals with no college GPA, no credits attempted and no credits earned; 59 observations, 8.48%) or transfer students, i.e. students with previous college experience at another institution (no college GPA and at least one college credit or at least one college credit attempted; 90 observations, 12.93%). For the purpose of this study, an indicator variable was created to account for the difference between individuals who were either first semester freshmen and/or transfer students (cited as "Frosh"), and those who had already attended classes at UNM (cited as "UC", or Upper Classmen).

GPA Credits / Credits Attempted / Credits Earned

The variables GPA Credits (college credits counting towards a students GPA), Credits Attempted (in college) and Credits Earned (college) were all highly correlated (see the Analysis section). Therefore it was determined to use only one of these explanatory variables for modeling. The Pearson Correlation Coefficients were calculated for each pair of continuous variables to determine which variable was most highly correlated with the continuous response variable, Final Exam Score. Credits Attempted was selected by the Investigator for the purposes of this study. 59 (8.48%) observations had missing values for all of these three variables and therefore were assumed to be associated with incoming freshmen. 87 (12.5%) of the observations had

no GPA Credits or Credits Attempted but did have at least one credit earned. These observations were determined to be transfer students with "Credits Earned" at another institution or possibly students who had received credit on an AP exam. The remaining three (0.43%) missing observations were deemed to be transfer students who did not have transferrable credits from another institution, therefore reporting zero "Credits Earned."

ACT/SAT Math Scores ("Test")

ACT and SAT Math scores were used as explanatory variables in both analytical models. Counts for students taking SAT, ACT, both or neither of the standardized exams are in Table 2.4.

Table 2.4
Separation Table of Students Taking Pre-College Standardized Tests

	ACT ONLY	SAT ONLY	BOTH	NONE	TOTAL
COUNT	513	38	75	70	n = 696
PERCENT	73.71%	5.46%	10.78%	10.05%	100%

If a student took both exams, only their ACT exam was used for analytical purposes. The ACT score was selected because a majority of the students in the sample (513 or 73.71%) only took the ACT as a pre-college entry exam yielding a total of 588 (84.48%) individuals with ACT scores. Thirty-eight students (5.46%) took only the SAT as their pre-college entry exam. The remaining 70 observations (10.05%) did not report a pre-

college exam score. It was assumed that these individuals were transfer students who were not required to submit these scores for admittance to the University.

All observations that had reported ACT/SAT scores (626, 89.94%) were then standardized to z-scores to allow for uniform comparison of ACT and SAT pre-college assessment exams. ACT scores were converted using $\mu = 21$ and $\sigma = 5.3$ as reported by 2010 National ACT Profile Report (The ACT, 2010). SAT scores were converted using $\mu = 516$ and $\sigma = 116$ as reported by the 2010 College Board Total Group Profile Report (The College Board, 2010).

Two methods were employed to impute missing values for standardized ACT/SAT scores. Multiple linear regression using quantitative variables from the students with reported ACT/SAT scores was initially employed. This method yielded an extremely weak coefficient of determination ($R^2 = 0.03$) making imputation of the missing values unreliable. As a result, the mean value of reported standardized test score was used to impute the 70 missing values.

Quiz

Instructors participating in the study were asked to indicate if they employed quizzing during the Fall semester of 2011. In addition, instructors were asked if quizzes counted toward the final overall grade for students within their section(s). Five instructors indicated that they administered quizzes during the Fall 2011 semester and all five of these instructors used quiz scores as part of the students' final overall grade in the course. Of the 696 observations that were used in the modeling process, 219

students (31.47%) received quizzing while the remaining 477 (68.53%) were not administered quizzes during the semester.

Time

The time of day each student took the course was collected from the instructor as part of the post-semester survey. The time variable was converted to a binary variable for the purposes of analysis, either AM or PM, depending on if the student took the class before 12:00 PM or 12:00 PM and after. 351 (50.43%) students were categorized as "AM" while 345 (49.57%) were categorized as "PM".

RESPONSE VARIABLES

Final Exam Score

Final exam scores were reported for each student whose instructor participated in the study. This continuous random variable was used as the response variable in the mixed model (HLM) to determine the relationship between quizzing and students' performance on the final exam. The final exam administered for Stat 145 in the Fall of 2011 was cumulative and therefore it is considered a reasonable measure of student aptitude with regards to course material. In addition, instructors do not grade their students' final exams individually; rather, all exams are pooled together and graded collectively by all instructors of Stat 145. Because all the exams are randomly graded by different instructors, grading bias is significantly reduced and therefore the variable is considered a good response measure for student aptitude in the course.

Final Grade

The final letter grade was reported for each student in the study. It was used as the response variable in the logit model as a measure of whether a student passed the course (a grade of C or higher) versus not passing the course. Students who withdrew before the end of the semester were not included in the final analysis because these individuals did not complete the course, and therefore did not have a final grade that could be used to determine if they, in fact, passed or failed the course.

UNUSED VARIABLES

Data associated with the following explanatory variables was collected, but was not used in either model for analysis: *Quiz Frequency, Quiz Weight, Quiz Type,* and *High School GPA.* The Investigator chose not to include frequency, weight or type of quiz variables during analysis of the data set because these variables varied significantly between instructors and no logical partitioning of these variables could be derived. An example is that some instructors used blended approaches to what they called a "quiz," employing both homework assignments and in-class quizzing as a method of quizzing. The Investigator decided to categorize any student whose instructor issued an in-class quiz that counted toward the final grade, regardless of employment method (i.e. paper-based or a combination of "clickers" and PowerPoint) as "quiz." Any student who did not receive in-class quizzing or was administered a quiz but the quiz score had no impact on the student's final grade was classified as "no quiz".

Students' High School GPA was included in the dataset supplied by the UNM Data Warehouse. There were four-hundred-seventy-two (472) missing values for this measure and therefore this variable was not used in model development or final analysis.

MODELS

Logistic Regression Model (Logit)

Because passing rates are typically a measure of interest to school administrators, a logistic model, using the SAS procedure PROC LOGISTIC (SAS Institute, 2012), was employed to determine the relationship between the administration of quizzes and students passing the course with a grade of C, or higher. This model is commonly used to predict the outcome of a categorical variable with a binary result, such as zero vs. one, using several explanatory variables, including the variable of interest (Downer, R.G. & Richardson, P.J., 2002) In this case, the response variable was "pass" vs. "not pass" and the variable of interest is "Quiz". Adjusted odds ratios and the 95% corresponding confidence intervals were used to estimate the underlying relationship between the explanatory variables and response.

Model selection was performed by initially fitting a full model with the following explanatory variables included: Quiz, Time, Gender, Ethnicity, GPA (college), Test (the standardized ACT/SAT score), and an indicator variable for new students to UNM (freshmen and/or transfers) versus students who had previously completed coursework at UNM. Non-significant variables were dropped using $\alpha = 0.05$.

Two methods were employed to determine the strength of the model. The first was the Receiver-Operator Curve (ROC), which yields the percent of values correctly predicted by the final model when compared to the observed values included in the data set. In the case of this study, the ROC indicates the percent of students the final model correctly predicts will pass or fail the class when compared to the actual, observed value for each student. A cross-validation technique was used as the second measure of the logit model strength. This method uses a Monte-Carlo approach to estimate the cross-validation error. For each iteration of the cross-validation process, a training set is randomly selected from the data to develop the model. This model is then used to predict the remaining, "unused" observations, called the test set, and then compares predicted responses (pass or fail) to the actual responses of the test set.

Hierarchical Linear Model (HLM) / Mixed Model

A mixed or hierarchical linear model (HLM), using the SAS procedure PROC MIXED, was employed to determine the relationship between the administration of quizzes and students' final exam score. This model is often used in educational studies because it takes into account the variation between instructors. The adjustment for standard errors associated with the random effects attributed to instructors yields a more accurate measure of variability between students, nested within an instructor's class (Raudenbush, S. & Bryck, A.S., 1986). In this study, the HLM should yield a better estimate of quizzing and its' impact on students' final exam score than a multiple

regression model. We note the restricted maximum likelihood (REML) was used to estimate the parameters in the model.

To maintain valid assumptions in modeling, the Investigator applied a power transformation to the response variable in the HLM utilizing the Box-Cox method. Interactions between variables were analyzed; however the interaction terms were not deemed statistically significant therefore these were not included in the final model.

CHAPTER 3: ANALYSIS

Two models were developed to analyze this data set (*n*=696), a multiple logistic regression model (logit) using the binary dependent variable "pass/fail" for the course, and a hierarchical linear regression model (HLM) using the students' final exam score as the response variable.

LOGISTIC REGRESSION MODEL

The general logistic regression model is:

$$[1] \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \boldsymbol{X}_{i1} + \cdots + \beta_j \boldsymbol{X}_{ij} \;;\; y_i \mid \pi_i \overset{ind}{\sim} Bern\,(\pi_i)$$

Where y_i = 1 if student *i* passed the course and zero otherwise, and π_i = *P* (student *i* passed). The X_j's are the explanatory variable values (i.e., quiz, gender, etc) for student *i* and the β_j's are the parameters of interest, where increasing values of the parameters increase the log odds.

Because logistic regression is not a linear function, there are no assumptions for normality or equal variance of the independent variables included in the model. The Investigator used Wald χ^2 tests to evaluate the likelihood a student would pass the course based on administration of quizzes, a standardized math aptitude test score, college GPA entering the course, as well as ethnicity and gender. In addition, a χ^2 goodness-of-fit test to determine model adequacy was employed. The SAS procedure PROC LOGISTIC was employed allowing the Investigator to evaluate the relationship

between quiz administration and student success in the course. Finally, a Monte-Carlo Cross-validation method for logistic regression models was employed using the SAS macro-procedure CVLR (Cross-Validation for Logistic Regression).

MODEL SELECTION *– LOGIT*

In the logistic regression model, the effect of quizzing and the likelihood of a student passing the course were analyzed given that quizzes were administered in that section. Standardized math test scores (SAT or ACT), the time the class was taken (i.e., before or after noon), gender, current college GPA, and ethnicity were included in the model to adjust for confounding variables.

Separation tables were generated to determine the frequency of observations for the categorical variables included in the model (refer to Tables 3.1 and 3.2). Relatively small cell counts for (i.e., $\leq$ 10) associated with the indicator variable "Time" were noted (refer to Table 3.2). The Investigator retained this variable to account for potential confounding effects, although some argue this variable should be dropped for small cell counts (Peduzzi et al., 1996). As Table 3.1 indicates, we observed 647 "Pass's" (or 'yes's) and 49 "No Pass's" (or 'no's).

Table 3.1
Response Profile for Logistic Model

Ordered Value	PASS	Total Frequency
1	NO PASS	49
2	PASS	647

Table 3.2
Separation Table for Factors

Obs.	Quiz	Gender	Ethnicity	Time	Count	Percent
1	N	F	H	AM	37	5.3161
2	N	F	H	PM	82	11.7816
3	N	F	O	AM	32	4.5977
4	N	F	O	PM	41	5.8908
5	N	F	W	AM	44	6.3218
6	N	F	W	PM	71	10.2011
7	N	M	H	AM	30	4.3103
8	N	M	H	PM	34	4.8851
9	N	M	O	AM	10	1.4368
10	N	M	O	PM	24	3.4483
11	N	M	W	AM	20	2.8736
12	N	M	W	PM	52	7.4713
13	Y	F	H	AM	35	5.0287
14	Y	F	H	PM	6	0.8621
15	Y	F	O	AM	21	3.0172
16	Y	F	O	PM	8	1.1494
17	Y	F	W	AM	44	6.3218
18	Y	F	W	PM	7	1.0057
19	Y	M	H	AM	29	4.1667
20	Y	M	H	PM	9	1.2931
21	Y	M	O	AM	17	2.4425
22	Y	M	O	PM	6	0.8621
23	Y	M	W	AM	32	4.5977
24	Y	M	W	PM	5	0.7184

A multiple linear regression model using the variables Final Exam Score, Gender and Credits Earned was employed to impute the value of "Test" for 131 missing values. This analysis did not yield a suitable model to impute these scores so the average of reported "Test" scores ($\bar{x}$ = 0.17) was used as a means to impute this data (Greenless, J.S., et al., 1982).

To account for the difference between students "new" to the University and upper-class students, this variable was converted to an indicator variable. The reference for this variable is cited as "UC", or upper-classmen while the alternative is cited as "Frosh" for freshman/new students.

It should be noted that PROC LOGISTIC was used on the two subsets of data with missing values and compared to the model using imputed data. Since there were no significant differences in the output between the models with missing values and the model using imputed data, the model using the imputed data was selected as the final reduced model for the purposes of analysis.

RESULTS - *LOGIT*

Analysis of the model effects (explanatory variables) indicate that the effect of quizzing (Quiz) is significant (*p*-value = 0.0012) as were the standardized test scores (Test, *p*-value = 0.0002). Gender, Time, GPA (college) and Ethnicity were found not to be significant at $\alpha = 0.05$ but were left in the model to account for potential confounding effects (Tables 3.3 and 3.4).

Table 3.3
Type III Analysis of Model Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Quiz	1	10.4811	0.0012
Time	1	0.0012	0.9718
GPA	1	0.0072	0.9325
Test	1	13.8279	0.0002
Gender	1	1.6065	0.2050
Ethnicity	2	0.1245	0.9397

Table 3.4
χ^2 -test Analysis of Maximum Likelihood Estimators

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.1391	0.3064	104.9366	<.0001
QUIZ	Quiz	1	0.8969	0.2771	10.4811	0.0012
TIME	AM	1	-0.00569	0.1612	0.0012	0.9718
GPA	Frosh	1	0.0179	0.2111	0.0072	0.9325
TEST		1	0.9103	0.2448	13.8279	0.0002
Gender	F	1	0.2011	0.1587	1.6065	0.2050
race	H	1	-0.0537	0.2092	0.0659	0.7974
race	O	1	-0.0170	0.2435	0.0049	0.9442

The final model derived from the observed data is:

[2] $$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

where the parameters and variables are defined in Table 3.5.

Table 3.5
Logistic Regression Model Parameters and Variables

Parameter	Variable	Definition	Reference for Indicator
β_o		Intercept	
β_1	X_1	Quiz	
β_2	X_2	Time (AM)	PM
β_3	X_3	GPA (Frosh)	UC
β_4	X_4	Gender (F)	Male
β_5	X_5	Test	
β_6	X_6	Ethnicity (H)	Ethnicity (W)
β_7	X_7	Ethnicity (O)	Ethnicity (W)

Testing the hypothesis H_o: $\beta_i = 0$, $\forall i \neq 0$ versus H_1: at least one $\beta_i \neq 0$, the likelihood ratio goodness-of-fit test statistics has a *p*-value < 0.0001, indicating that there are no gross deficiencies with the model (Table 3.6).

Table 3.6
Logistic Regression Model Likelihood Ratio Test

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	34.7064	7	<.0001
Score	30.6267	7	<.0001
Wald	26.8969	7	0.0003

The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Table 3.7 gives a summary of the odds ratios point estimators for each of the explanatory variables included in the final logit model.

Table 3.7
Odds Ratio Estimates and 95% Confidence Intervals

Effect	Point Estimate	95% Wald Confidence Limits	
QUIZ Quiz vs. No Quiz	6.013	2.030	17.813
TIME AM vs. PM	0.989	0.526	1.860
GPA Frosh vs. UC	1.036	0.453	2.371
TEST	2.485	1.538	4.015
Gender F vs. M	1.495	0.803	2.785
race H vs. W	0.883	0.439	1.774
race O vs. W	0.916	0.405	2.072

In the case of quizzing, the primary focus of this study, students whose instructor employed in-class quizzes were 6.013 times more likely to pass the course than students who receive no quizzing with a 95% confidence interval of (2.030, 17.813), given that all other variables in the model are kept constant. The lower bound of the confidence interval is particularly interesting when one takes into account the very small *p*-value of 0.0012 associated with quizzing in the model. The results suggest that, at a minimum, students in this sample whose instructors gave quizzes in Stat 145 were more than twice as likely to pass the course than students whose instructors did not give quizzes, conditional on all other variables remaining constant.

Also of interest is the strength of a student's mathematical ability, as measured by the standardized pre-college math aptitude exam (Test). A student is 2.485 times more likely to pass the course (95% CI: (1.538, 4.015); *p*-value = 0.0011) for each increase in their standardized test score of 0.9103 points (equivalent to an increase of 105.59 points on the SAT or 4.8 points on the ACT) with all other variables in the model are kept constant. This indicates that students with increased math aptitude are more likely to pass the course versus students with presumably lower math aptitude, based on pre-college standardized testing.

We now consider the variable GPA that compares new students to students who have had prior coursework at the University of New Mexico. There appears to be no advantage for students with previous coursework performed at the University according to the odds ratio for this data set. For the case of this study, the odds of a student with no prior experience at the University (Frosh) is almost equivalent to Upper Classmen with an odds-ratio of 1.036 for new students (95% CI: 0.453, 2.371).

Odds ratio analysis of the categorical variables ethnicity and gender yield interesting, but less compelling results. The logit model *suggests* Females are 1.495 more likely to pass the course (95% CI: 0.803, 2.785) relative to Males, with all other variables kept constant. The model also *suggests* that students of Hispanic or Other ethnic backgrounds are less likely, on average, to pass the course as compared to White students. In the case of Hispanic versus White students, the odds ratio is 0.883 (95% CI: 0.439, 1.774). The variable "Other" that includes students who reported their ethnicity as Asian, Black, Native American and Other, yielded an odds ratio of 0.916 (95% CI:

0.405, 2.072) when compared to White students and keeping all other variables constant. However, these results all include an odds ratio of 1.0 within each confidence interval indicating that the odds-ratios for gender and ethnicity are not statistically significant.

MODEL STRENGTH *–LOGIT*

Receiver-Operator Characteristic (ROC)

To validate the strength or predictive power of the derived logit model, analysis of the Receiver-Operator Characteristic (ROC) Curve and the associated area under this curve (c) was performed. The ROC curve provides a method of mapping predicted binary outcomes, in the case of this study "pass" versus "not pass", based on the derived model against observed values in the data set. The ROC curve plots the proportion of true positive rates against false positive rates. The true positive rate (TPR) yields the number of predicted positive results (i.e. "pass") that correctly match the observed result. The false positive rate (FPR) yields the number of predicted positive results that did not correctly match the observed result, in other words, when the model predicted a result of "pass" for a particular observation, but the actual observed value was "not pass." For this model, the proportion of the TPR (predicted "pass" to observed "pass"), or the percent concordant, was 73.9% versus the FPR (predicted "pass" to observed "not pass"), or the percent discordant, was 25.9% yielding an area under the ROC curve, c = .740. This statistic implies that the model correctly predicts a student's performance in the class ("pass" versus "not pass"), based on the explanatory

variables included in the model, approximately 74% of the time, which is an indication of strong predictive power (refer to TABLE 3.8 and Figure 3.1).

Table 3.8
Concordant/Discordant Values

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.9	Somers' D	0.480
Percent Discordant	25.9	Gamma	0.481
Percent Tied	0.2	Tau-a	0.063
Pairs	31703	c	0.740

Figure 3.1
Receiver-Operator Curve

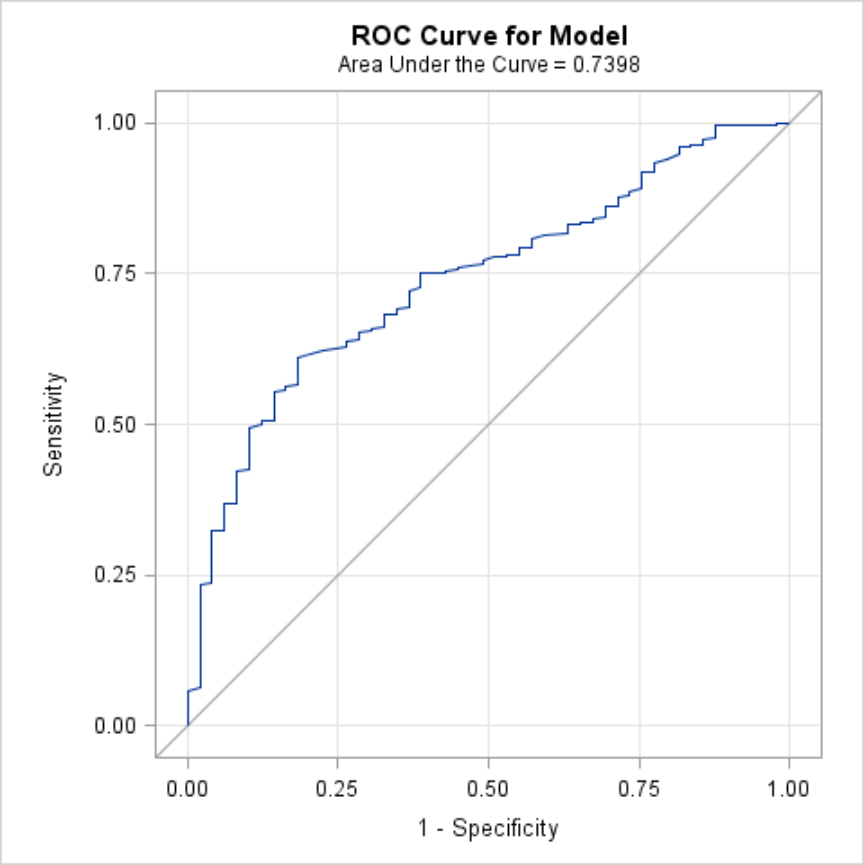
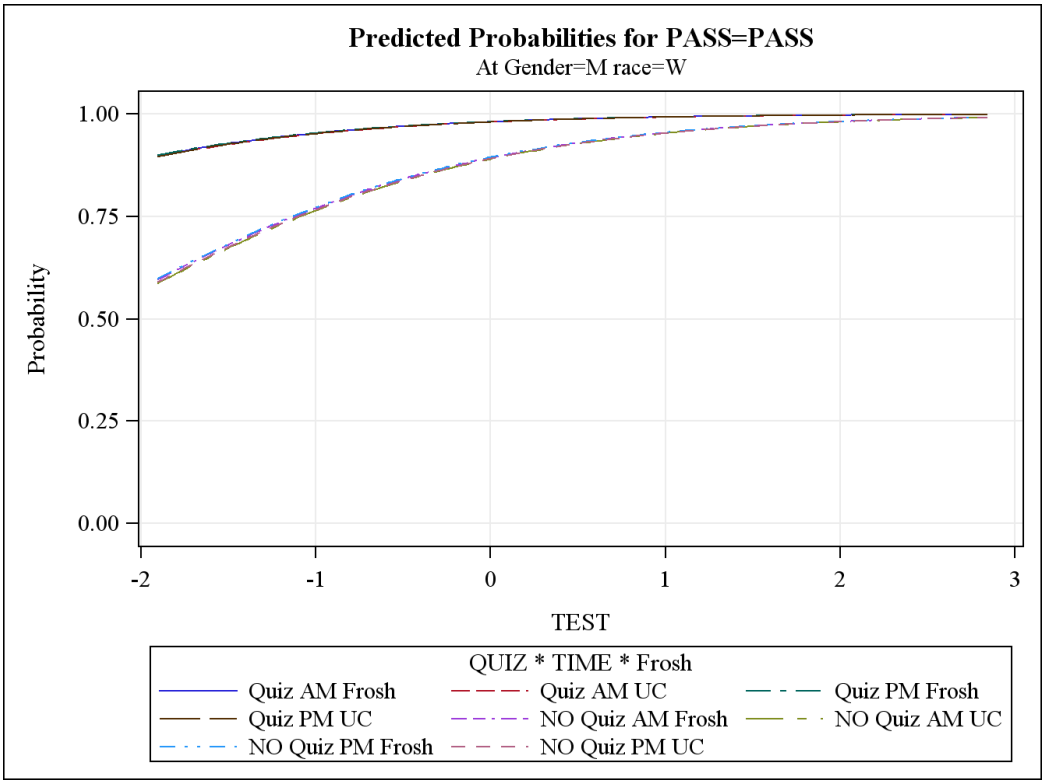


Figure 3.2 plots the probability a student will pass the course (with 95% Confidence Limits) against their standardized pre-college math aptitude score, Test (at gender = Male and ethnicity = White) for Quiz (vs. No Quiz), Time (AM/PM) and GPA (Frosh/UC). Consistent with the derived model, as individuals' standardized test score increases, the probability that they will pass the course increases, regardless of the time of day or if they were new students (Frosh) versus upperclassmen (UC). Of considerable interest is the difference in the likelihood of passing the course for students with low math aptitude scores. These students benefit significantly from quizzing. The lower curves in the graph are associated with students who did not receive quizzing while the upper curves yield probabilities for students who received quizzing. For students with standardized math aptitude scores approximately two standard deviations below the mean, the probability of passing the course increases at least 20% (P (pass | no quizzing) $\approx$ .65 vs. P (pass | quizzing) $\approx$ .87).

Figure 3.2

Probability to Pass Given: Quiz, Time, GPA by Test Score



CROSS-VALIDATION *- LOGIT*

Cross-validation of the data set was performed to estimate the general performance of the model's ability to correctly predict if a student will pass or not pass the course (Refaeilzadeh, P., et al., 2007). This method of cross-validation removes a subset of observations, called a testing set, and uses the remaining observations to create the model, referred to as the training set. The subsequent model is then used to predict results for the testing set and then compares the predicted results to the observed results in the testing subset. For this study the SAS macro CVLR, using a

Monte-Carlo selection method to create the training set, was employed to perform the cross-validation procedure. Thirty percent (30%) of the data was used as the testing set while the remaining 70% of the data was used for the logistic regression. The cross-validation method was employed 1,000 times, each time yielding the percent correctly classified. The CVLR procedure then averages the percent correctly classified for all iterations of estimable models. The results of this procedure were compelling, yielding a mean percent correctly classified of 92.92% with a Monte Carlo standard error of 0.05%.

LIMITATIONS *– LOGIT*

Because there were several cases where values were not reported, imputation of observations for several missing variables was employed. Initially, the Investigator attempted to use available quantitative variables and employ multivariate linear regression techniques to impute missing values. These models (used to impute standardized SAT/ACT math scores ("Test")) were not robust; therefore the mean of observed scores was used to impute these values. (Greenless, J.S., et al., 1982) Using this method of imputation introduces bias into the model because we assume the observations with missing standardized math scores have an average standardized score of approximately 0.1712. In addition, the process of multiple imputation can lead to inaccurate parameter estimates, standard errors and hypothesis tests (Little, R.J.A. & Rubin, D.B., 1987)**.**

As cited earlier, some argue that small cell counts for categorical variables can lead to inaccurate parameter estimates. According to Peduzzi (1996) the minimum number of events per variable (EVP) is 20. "Logistic regression is a large sample method. A rule of thumb is that there should be at least 10 'yes's and 10 'no's, and preferably 20, for each predictor variable" (Peduzzi, P., et al. 1996). The Investigator felt that because this was an observational study the influence of confounding variables outweighs this school of thought and decided to retain this factor in spite of small cell counts.

ANALYSIS *– Hierarchical Linear Model (HLM) / Mixed Model*

A mixed model, sometimes referred to as an HLM, was used to describe the relationship between administration of quizzes in the course and students' final exam score (FES), a continuous response variable, while adjusting for other variables. The general form of the mixed model is:

$$[3] \qquad Y = X\beta + Z\tau + \varepsilon,$$

In this model, Y is an (n x 1) vector of response variables where X is an (n x p) design matrix of predictors (including a column of 1s, n = 696); β is a p-dimensional vector of fixed-effects parameters, Z is an (m x 1) vector of instructors (m is the number of instructors), and τ is a (1 x m) vector of random effects. It is assumed that $\tau \sim N(0, \sigma_\tau^2 I)_p$, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)_n$ and $\operatorname{cov}(\tau, \varepsilon) = 0$, where I_k is an (k x k) identity matrix.

It is worth noting that usually there is no interest in comparisons among the levels of random effects. Rather, there is interest in studying the variability of these

effects or in controlling for that variation so that we can derive reliable conclusions about fixed effects.

The following explanatory variables were used to fit the full model to estimate the Final Exam Score (FES) for students whose instructors participated in the study: Quiz, Test, Credits Attempted, GPA (college), Gender, Ethnicity and Time. The Investigator used a *Z*-test for Covariance Parameter Estimates (H_0: $\sigma_\tau^2 = 0$ vs. H_1: $\sigma_\tau^2 > 0$) to determine if the random effects due to instructors are significant.

All analyses were performed using type III sum of squares to account for unequal sample sizes for evaluating parameters associated with the model's fixed effects. Least Square Means (LSM) were used to construct 95% confidence intervals for predicted final exam scores given students' were (or were not) administered quizzes during the course. SAS procedures PROC CORR and PROC SGSCATTER (SAS Institute, 2012) were used to calculate Pearson Correlation Coefficients and to construct Correlation Scatter Plot matrices respectively.

To adjust for possible confounding effects due to the variables Test, Credits Attempted, GPA (college), Gender, Ethnicity and Time, a regression analysis, taking into account random effects (using SAS PROC MIXED), was employed to examine if there were significant differences between students' quizzing status with respect to achievement (final exam scores). We note that PROC MIXED fits random effects models in order to accommodate several sources of variation instead of just one as stated in [1]. Finally, Tukey-Kramer's test was used to explore pair-wise comparisons between levels of the categorical variables used in the final model (see Methods).

MODEL SELECTION *- HLM*

Missing values for students' standardized pre-college math exam (Test) were imputed using the sample mean similar to the logit model multiple imputation method. A scatterplot diagram of the response and explanatory variables was generated to determine if any of the explanatory variables were highly correlated with each other, thus allowing the opportunity for variable reduction prior to fitting the full model (Figure 3.3). The scatterplot diagram and a Pearson Correlation Matrix (Table 3.9) show that GPA Credits, Credits Attempted and GPA Credits Earned were highly correlated with each other. Credits Attempted was the most highly correlated of these variables with the Final Exam Score. To reduce potential multicollinearity, the Investigator chose to use only Credits Attempted for the purposes of fitting the full model.

Figure 3.3
Scatterplot Diagram for Model Response and Explanatory Variables

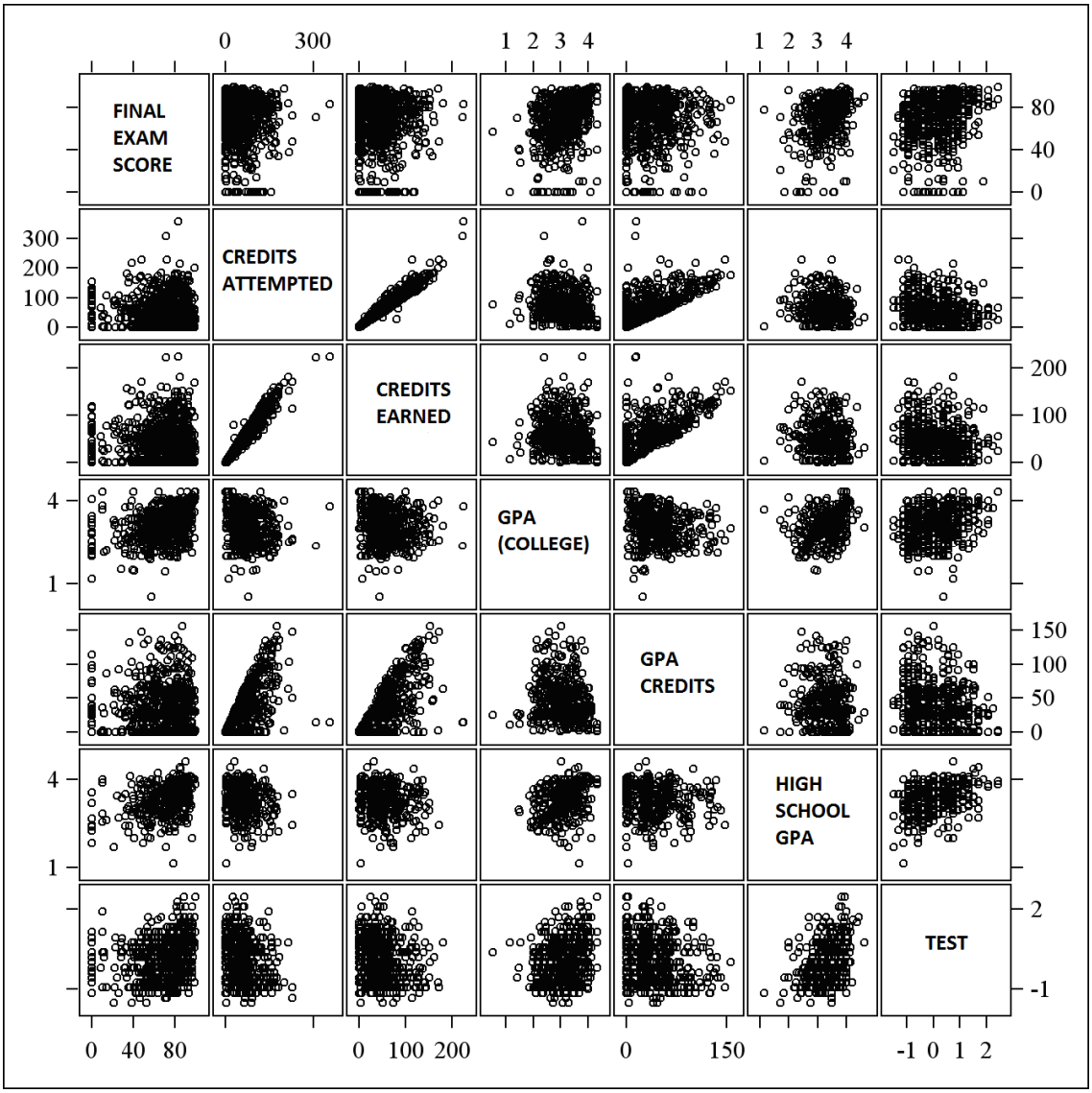


Table 3.9
Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0 Number of Observations							
	Final Exam Score	**Credits Attempted**	**Credits Earned**	**GPA (college)**	**GPA Credits**	**H.S. GPA**	**Test**
Final Exam Score	1.00000 818	-0.05514 0.1151 818	-0.01997 0.5685 818	0.33797 <.0001 610	-0.01893 0.5887 818	0.34317 <.0001 364	0.22920 <.0001 695
Credits Attempted	-0.05514 0.1151 818	1.00000 818	0.97341 <.0001 818	-0.23736 <.0001 610	0.71537 <.0001 818	-0.17125 0.0010 364	-0.26765 <.0001 695
Credits Earned	-0.01997 0.5685 818	0.97341 <.0001 818	1.00000 818	-0.15288 0.0002 610	0.73575 <.0001 818	-0.12110 0.0208 364	-0.24522 <.0001 695
GPA (college)	0.33797 <.0001 610	-0.23736 <.0001 610	-0.15288 0.0002 610	1.00000 610	-0.22638 <.0001 610	0.35102 <.0001 340	0.22665 <.0001 539
GPA Credits	-0.01893 0.5887 818	0.71537 <.0001 818	0.73575 <.0001 818	-0.22638 <.0001 610	1.00000 818	-0.07201 0.1704 364	-0.25744 <.0001 695
H.S. GPA	0.34317 <.0001 364	-0.17125 0.0010 364	-0.12110 0.0208 364	0.35102 <.0001 340	-0.07201 0.1704 364	1.00000 364	0.36916 <.0001 349
Test	0.22920 <.0001 695	-0.26765 <.0001 695	-0.24522 <.0001 695	0.22665 <.0001 539	-0.25744 <.0001 695	0.36916 <.0001 349	1.00000 695

As in the logit model, the researcher used an indicator variable to account for the difference between students “new” to the University and upper-class students. The reference for this variable is cited as “UC”, or upper-classmen, while the alternative is cited as “Frosh” for freshman/new students.

After fitting the full model and a review of diagnostics was completed, it was determined that the assumption of constant variance was violated. A Box-Cox procedure was employed to determine the best power transformation for the response

variable, yielding $\lambda = 1.7$. It was determined that a square transformation on the response variable, Final Exam Score (FES), corrected the non-constant variance issue.

Estimating conditional variance components associated with the random (instructor) effects was performed. The total error associated with the model is $\sigma_\tau^2 + \sigma_\varepsilon^2 = 5{,}644{,}336$. The variance component estimate associated with instructor effects is $\sigma_\tau^2 = 357{,}063$ while the value of the estimate for fixed effects is $\sigma_\varepsilon^2 = 5{,}287{,}243$. The interclass correlation (ICC) yields the proportion of the total variance between instructors:

[4] $\text{ICC} = \sigma_\tau^2 / (\sigma_\tau^2 + \sigma_\varepsilon^2) = 0.0633.$

This value indicates that approximately 6.33% of the total variation in the model is associated with instructors. The remaining variation is associated with the residuals of the fixed effects retained in the model.

Covariance parameter estimates using the χ^2- test for the instructor random effects terms (τ_i) were significant at α = 0.05. The SAS output yields a *p*-value = 0.0579, however, "testing the significance of variance components is a nonstandard problem since the null hypothesis (i.e., random effects has zero variance) is on the boundary of the parameter space of the alternative hypothesis" (Van Dongen, S., et al., 1999). Van Dongen suggests that dividing the *p*-value by two yields a corrected level of significance, in this case a *p*-value = 0.02895 (Refer to Table 3.10).

Table 3.10
Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	*Z* Value	Pr > *Z*
INSTRUCTOR	357063	227036	1.57	0.0579
Residual	5287243	286670	18.44	<.0001

The final reduced model is:

$$[5] \quad y_i^2 = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \tau_1 z_{1i} + \tau_2 z_{2i} + \tau_3 z_{3i} + \tau_4 z_{4i} + \tau_5 z_{5i} + \tau_6 z_{6i} + \tau_7 z_{7i} + \tau_8 z_{8i} + \tau_9 z_{9i} + \tau_{10} z_{10i} + \varepsilon_i$$

for $i = 1, 2, \ldots, 696$

The HLM parameters and variables are defined in Table 3.11 as follows:

Table 3.11
HLM Parameters and Variables

Parameter	Variable	Definition	Reference for Indicator
β_1	X_1	Test	
β_2	X_2	Time (AM)	PM
β_3	X_3	Gender (F)	Male
β_4	X_4	Ethnicity (H)	White
β_5	X_5	Ethnicity (O)	White
β_6	X_6	Quiz	No Quiz
β_7	X_7	GPA (Frosh)	UC
τ_1	Z_1	Instructor A	
τ_2	Z_2	Instructor B	
τ_3	Z_3	Instructor C	
τ_4	Z_4	Instructor D	
τ_5	Z_5	Instructor E	
τ_6	Z_6	Instructor F	
τ_7	Z_7	Instructor G	
τ_8	Z_8	Instructor H	
τ_9	Z_9	Instructor I	
τ_{10}	Z_{10}	Instructor J	

RESULTS *- HLM*

The primary variable of interest, Quiz, was marginally significant (*p*-value = 0.0567). The variables TEST (*p*-value < 0.001) and Gender (*p*-value = 0.012) were both found to be significant at α = 0.05. To account for confounding effects the variables Time, Ethnicity and GPA were retained in the final model, although all parameter estimates associated with these variables were deemed not statistically significant (refer to Table 3.12).

Table 3.12
Solution for Fixed Effects

Effect	Ethnicity	Quiz	Gender	Time	GPA	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept						4491.51	343.45	8	13.08	<.0001
Test						927.95	128.28	680	7.23	<.0001
Time				AM		-115.16	260.54	680	-0.44	0.6586
Gender			F			457.50	182.12	680	2.51	0.0122
Ethnicity	H					281.22	202.53	680	1.39	0.1654
Ethnicity	O					288.64	231.23	680	1.25	0.2123
Quiz		Quiz				860.55	450.83	680	1.91	0.0567
GPA					Frosh	-83.0814	228.13	680	-0.36	0.7158

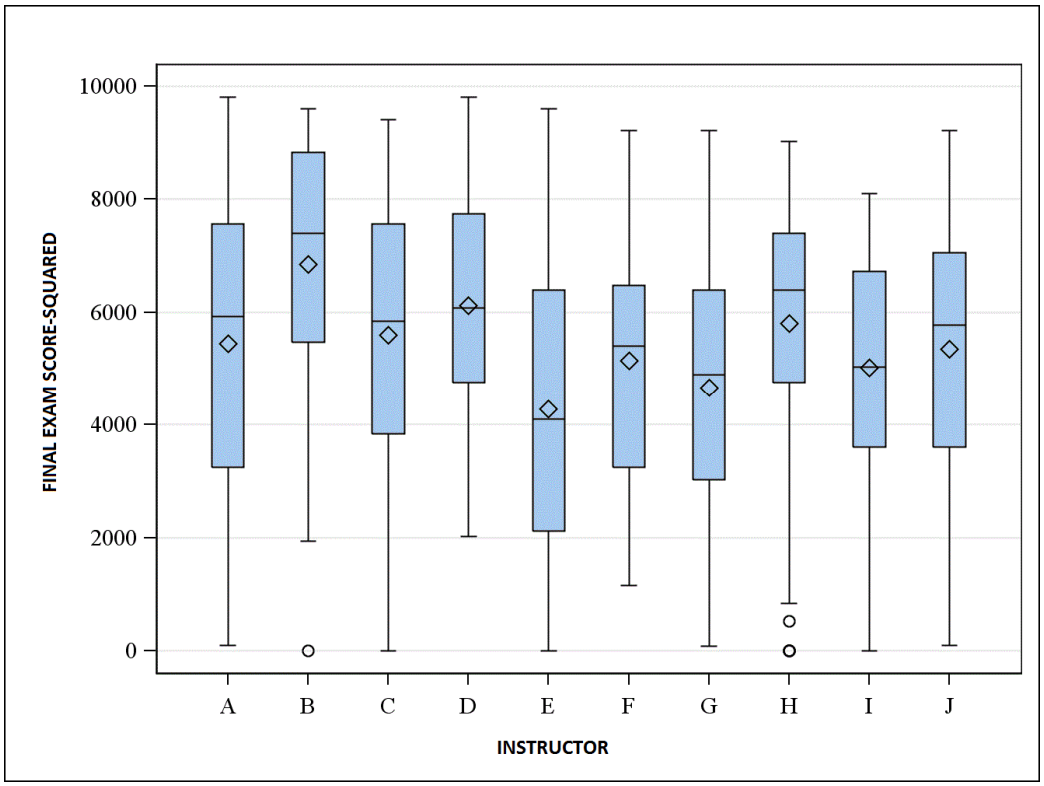
We note that the prediction for each student is the sum of the solution for fixed effects plus the solution for random effects found in Table 3.13. The solution for random effects reveals that the source of variation between instructors is due to instructors B, E and H which is illustrated in Figure 3.4.

Table 3.13
Solution for Random Effects

Effect	INSTRUCTOR	Estimate	Std Err Pred	DF	t Value	Pr > |t|
INSTRUCTOR	A	342.77	337.59	680	1.02	0.3103
INSTRUCTOR	B	810.95	382.59	680	2.12	0.0344
INSTRUCTOR	C	-197.22	375.54	680	-0.53	0.5996
INSTRUCTOR	D	228.17	391.21	680	0.58	0.5599
INSTRUCTOR	E	-613.49	329.02	680	-1.86	0.0627
INSTRUCTOR	F	-371.65	377.69	680	-0.98	0.3255
INSTRUCTOR	G	-316.00	370.03	680	-0.85	0.3934
INSTRUCTOR	H	696.41	344.25	680	2.02	0.0435
INSTRUCTOR	I	-109.69	374.50	680	-0.29	0.7697
INSTRUCTOR	J	-470.25	414.41	680	-1.13	0.2569

Figure 3.4 graphically displays the distribution of conditional residuals for each instructor who participated in the study. Analyzing the box plots for instructors B, E and H confirm the larger variation estimates cited in Table 3.13.

Figure 3.4
Distribution of Final Exam Score-Squared by Instructor



Means Comparisons

Means comparisons for levels within the following categorical variables were performed: Quiz, Ethnicity, Gender, Time and GPA. LS Means comparisons for the factors Ethnicity, Time and GPA were not found to be statistically significant. Plots and Tukey-Kramer Multiple Comparison output for these factors can be found in Appendix A. The computed LS Means are for the squared final exam score, the power transformation required to adjust for non-constant variance of the initially fitted model.

Tukey-Kramer tests the null hypothesis that means of the factor levels are equal versus the alternative that at least one of the factor level means is different. The test also assumes observations are independent and variance across observations is constant. The test performs multiple comparisons simultaneously, which controls the overall probability of a Type-1 error, α (the probability of rejecting the null hypothesis when the null hypothesis is true).

Quiz – LS Means Comparison

A Tukey-Kramer *t*-test was performed for the factor Quiz which has two levels: Quiz and No Quiz as defined in the Methods section. Estimates for the LS Means for the squared final exam scores and associated confidence intervals are given in Table 3.14 for each factor level. The actual mean estimates and 95% confidence intervals for final exam scores (out of 100) are: Quiz (LS Mean Est. = 76.26; 95% CI: 71.91, 80.38); No Quiz (LS Mean Est. = 70.39; 95% CI: 65.95, 74.57).

The results of the Tukey-Kramer adjusted *t*-test (Table 3.15) indicate that the difference between administering quizzes and not administering quizzes is marginally significant (p-value = 0.0567). Figure 3.5 further illustrates that there appears to be a difference between students who received quizzes versus students who did not receive periodic quizzing.

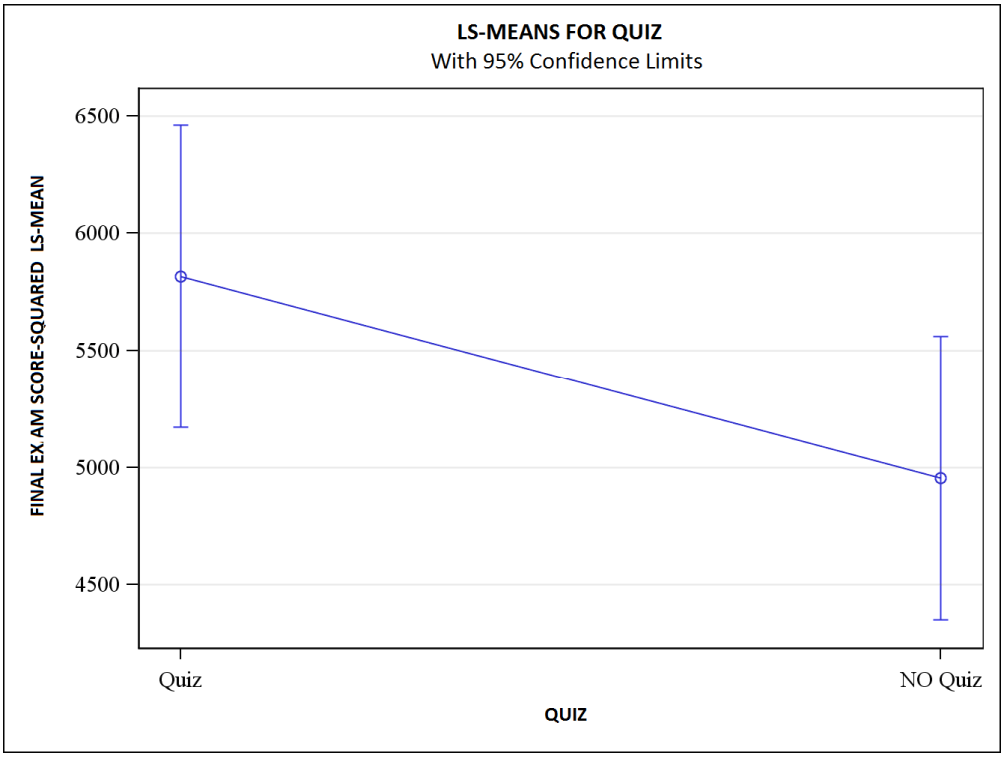
Table 3.14
Quiz Least Squares Means

Quiz	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Quiz	5815.73	328.45	680	17.71	<.0001	0.05	5170.83	6460.63
NO Quiz	4955.16	308.44	680	16.07	<.0001	0.05	4349.56	5560.77

Table 3.15
Tukey-Kramer Adjusted t-test for Quiz

Differences of Quiz Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
Quiz	Quiz	Estimate	Std. Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Quiz	NO Quiz	860.57	450.89	680	1.91	0.0567	0.0567	0.05	-24.7354	1745.87	-24.7356	1745.87

Figure 3.5
LS Means Comparison of Final Exam Score-Squared by Quiz



Gender – LS Means Comparison

The Tukey-Kramer *t*-test for the factor Gender as associated with the response variable "Final Exam Score" indicates that there is a significant difference when comparing females and males (*p*-value = 0.0122). Root adjusted LS Mean values for final exam scores and associated confidence intervals are as follows: Females: {LS Mean = 74.93, (71.80, 77.93)}; Males: {LS Mean Est. = 71.81, (68.29, 75.17)}. See Tables 3.16 and 3.17. Figure 3.6 graphically displays the difference in LS Means between males and females, with the scale being final exam score squared.

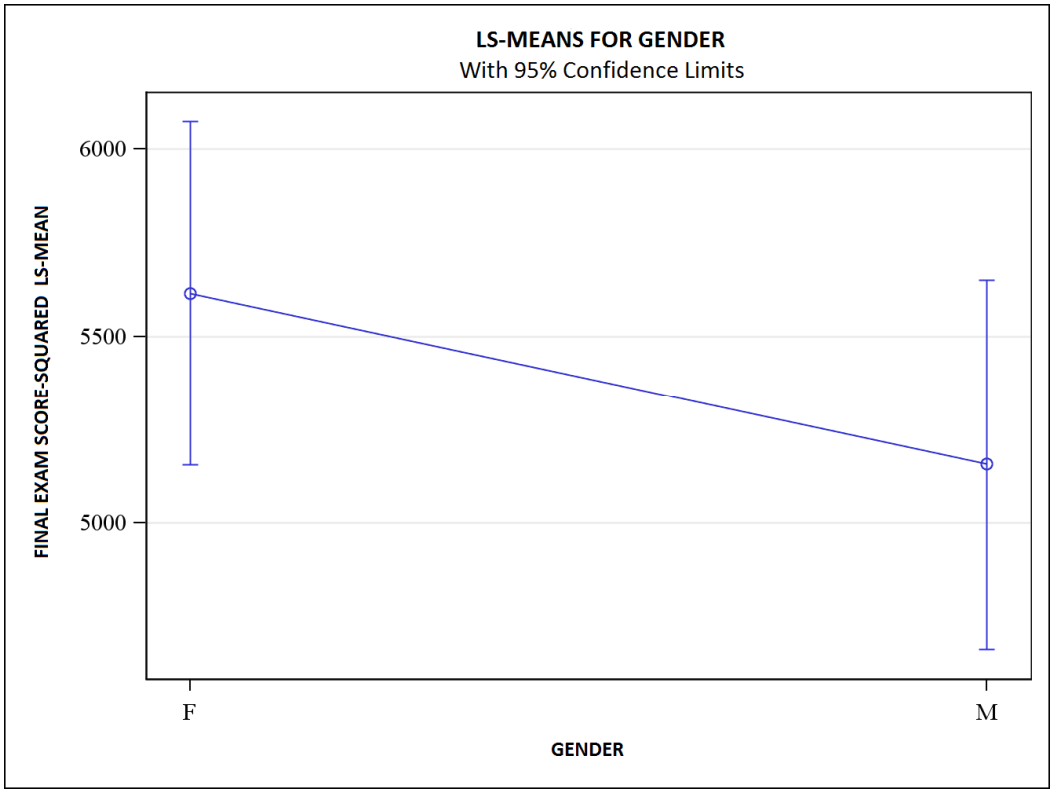
Table 3.16
Gender Least Squares Means

Gender	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
F	5614.20	233.92	680	24.00	<.0001	0.05	5154.91	6073.49
M	5156.70	251.45	680	20.51	<.0001	0.05	4662.98	5650.41

Table 3.17
Tukey-Kramer Adjusted t-test for Gender

Differences of Gender Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
Gender	Gender	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
F	M	457.50	182.12	680	2.51	0.0122	0.0122	0.05	99.9284	815.08	99.9283	815.08

Figure 3.6
LS Means Comparison of Final Exam Score-Squared by Gender



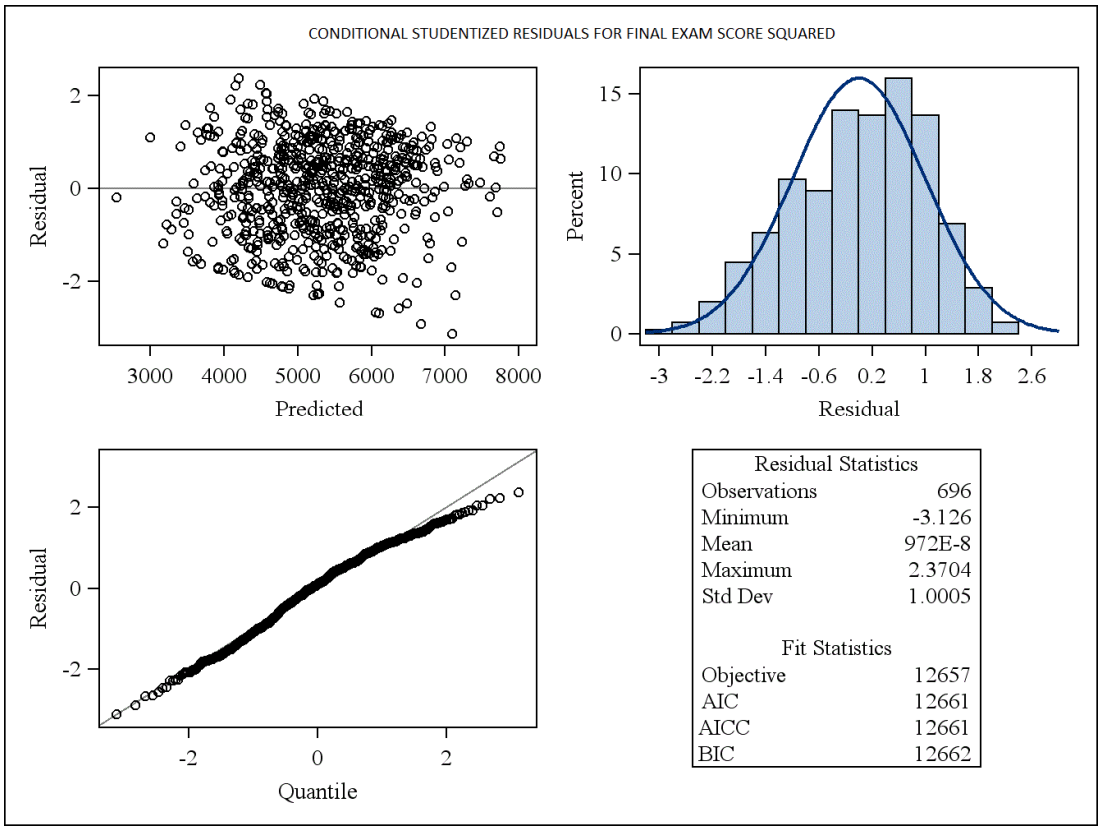
DIAGNOSTICS *- HLM*

The model assumptions for the HLM are non-constant variance of error terms, normal distribution of error terms and independence of error terms.

Figure 3.7 yields the diagnostic results for the model. The plot of predicted values versus the residuals indicates no violation of constant variance, with no apparent pattern to the scatter plot (e.g. random). Also the residuals are found to be roughly normally distributed with mean 0 and standard deviation of 1 and they fall primarily within 2 standard deviations of the center and are randomly scattered about zero, with few observations deviating from this trend. The Q-Q Plot indicates no evident deviation

from normality. Further, the histogram shows that the distribution of residuals is slightly left-skewed, however; this assumption could be relaxed due to homogeneity of variance. The interpretation is that the assumptions of normally distributed error terms and constant variance are not violated.

Figure 3.7
Diagnostic Plots and Residual Statistics for the HLM



LIMITATIONS *– HLM*

As with the logit model, due to the fact there were several cases where values were not reported, multiple imputation of missing observations for the variable "Test" was used. This method, as previously discussed, introduces bias by assuming that all students with missing standardized math test scores are "average" students. Furthermore, Little and Rubin (1987) argue that multiple imputation can lead to inaccurate estimates for parameters, standard errors and hypothesis tests.

Also, the assumption of normally distributed error terms for the random effects associated with instructors is hard to evaluate due to the small number of instructors in the sample. In particular, there were only 10 instructors included in this study, therefore, when performing diagnostics on the final model, we only considered the conditional studentized residuals which correspond to the students, not instructors.

CHAPTER 4: ANALYTICAL CONCLUSIONS

LOGISTIC REGRESSION MODEL

The purpose of the logistic regression model was to determine the relationship between quizzing status in an introductory college-level statistics course and the likelihood students pass the course (i.e., receive a grade of C or higher). Results suggest that a student who receives periodic quizzing, while holding all other variables constant, is over six times more likely to pass the course when compared with students who receive no quizzing. It also appears that, for each increase of 0.9103 points of the standardized math aptitude score, while holding all other variables constant, students are nearly 2.5 times more likely to pass the course.

In addition, it was shown that students who have low prior achievement differentially benefit from quizzing. Students with very low standardized math aptitude scores (approximately two standard deviations below the mean) are at least 20% more likely to pass the course if quizzes are administered by their instructor.

Non-significant variables were retained in the model to account for potentially confounding effects. These variables were gender, the time of day the class was taken (AM versus PM), ethnicity (white, Hispanic, other) and students new to the University (Frosh) versus upperclassmen (UC). All of these variables included a value of 1 in their respective confidence intervals for odds-ratio estimates. This indicates that there was no significant difference in the probability a student would pass the course based on these factors given the other factors remained constant.

The model is considered good based on the ROC area under the curve value (c = 0.740) (Jones, C.M. & Athanasiou, T., 2005), indicating that the model correctly predicts the observed value "pass" 74% of the time. The Monte-Carlo Cross Validation method was employed to further determine the model's strength at correctly predicting if a student passes or does not pass the course. The cross-validation procedure correctly classified pass or not pass at a rate of 93% for 1,000 iterations.

HIERARCHICAL LINEAR MODEL (HLM) / MIXED MODEL

A Hierarchical Linear Model (HLM) was fitted for the data to evaluate the relationship between quiz status and students' final exam score. The HLM was employed to account for variability associated with different instructors that could not be accounted for quantitatively. The model takes into account that students who have the same instructor are highly correlated. As a result, the model reduces the noise associated with variability due to instructors allowing us to more readily evaluate the fixed effects of quizzing on students' final exam score.

A power transformation was required to correct for non-constant variance after initially fitting the model. Results of the HLM analysis indicate a positive association between students who have instructors who administer quizzes and students' final exam scores, with the parameter estimate being marginally significant at $\alpha = 0.05$ (p-value = 0.0567). Random effects associated with instructors were considered significant (p-value = 0.02895) and accounted for approximately 6% of the variability within the model.

LS Means comparisons using a Tukey-Kramer *t*-test were performed for the variable "Quiz." The difference in LS Means was found to be significant at $\alpha = 0.05$ (p-value = 0.0567) with a mean final exam score of 76.26 out of 100 for students who received quizzing versus 70.39 out of 100 for students who did not receive quizzing.

Both the logistic regression model and the HLM suggest that quizzing has a positive association with students' performance in the introductory statistics course on which this observational study was based. However, observational studies always run the risk of finding correlations that mask real relationships because truly predictive variables are unobserved. For example, it is possible that instructors' use of quizzing is correlated with another unobserved variable that actually predicts performance in this introductory statistics course. For this reason, we cannot conclude that it was the quizzing *per se* that explains the phenomena we observe in the data. Due to the limitations of collecting and analyzing data in an observational study, an experiment, controlling for random effects due to students within classrooms and variability between instructors, would provide more robust results, allowing us to better determine the impact of quizzing in an introductory college level statistics course.

CHAPTER 5: DISCUSSION

The two models used to analyze the effect of quizzing on student success were developed based on the response data collected by the Investigator: students' final exam scores and students' final letter grades. The Investigator explored the likelihood that a student would pass STAT 145 given that quizzes were administered by the instructor. A probabilistic method, employing a logistic regression model was used to model the binary response of either passing or not passing the course. The Investigator used a grade of "C" or higher to define "passing." Logit modeling is a commonly employed statistical regression method used for dichotomous categorical outcomes given a set of explanatory variables.

For the continuous random variable, final exam score, a predictive modeling method using a hierarchical linear model (HLM) was employed. The purpose of using an HLM versus a single-level multiple regression model is to account for variation between instructors. As Raudenbush and Bryk argue, ignoring this variation may lead to inaccurate estimates of the response variable in terms of its association with explanatory variables included in the model (Raudenbush, S. & Bryk, A.S., 1986).

LOGISTIC REGRESSION MODEL (LOGIT)

To analyze the probability a student passes STAT 145 at the University of New Mexico, the Investigator employed a logit model. Results of this probabilistic modeling method are consistent with the results of the HLM, that quizzing has a positive impact

on student achievement in the large undergraduate core statistics course investigated in this study. The parameter estimate for quizzing in the logit model (0.8969, *p*-value = 0.0012) indicates a positive association between quizzing and the likelihood a student passes the course with a grade of "C" or higher.

Odds-ratio results for quizzing indicate that a student is approximately 6 times more likely to pass the course given quizzing versus similar students who do not receive quizzing. It is important to note the lower bound of confidence interval associated with the odds-ratio estimate of 6.013 (95% CI: 2.030, 17.813), indicating that a student who receives quizzing is at least two times more likely to pass the course with the probability of making a Type I error less than 5%.

The logit model also indicates that math aptitude is an important factor when predicting how well a student will perform in STAT 145. Math aptitude, as measured by the standardized SAT/ACT score ("Test"), was shown to be significant within the model (*p*-value = 0.0002). The point estimate for the odds-ratio associated with the variable "Test" indicates that for each increase of 0.9103 points of the standardized math aptitude score students are nearly 2.5 times more likely to pass the course than students who receive a score equivalent to the standardized national average. Once again, this result is consistent with the HLM giving further evidence that quizzing positively affects student achievement in the course.

Of note is that gender did not appear to make a difference in terms of a student's likelihood to pass the course. The parameter estimate for gender was not significant (*p*-value = 0.2050). While the point estimate for the odds-ratio of gender

indicates that females are 1.5 times more likely to pass the course than males, closer inspection of the confidence interval associated with this estimate shows that the interval contains 1.0 (95% CI: 0.803, 2.785). The interpretation of this observation is that gender does not play a significant role when determining if a student will pass the course or not. As Christensen (1990) points out, "…if the odds ratio is one, the two sets of odds are equal"; the interpretation for gender is that it is not relevant if a student is male or female in terms of the likelihood of passing the class because the confidence interval contains one (i.e. equal likelihood to pass the course).

The Investigator felt it important to determine the predictive strength of the logit model by performing cross-validation using a Monte Carlo selection method to create training and testing sets. The results of the cross-validation (CV) were compelling such that the model correctly classified students as either passing or failing the course approximately 93% of the time for 1,000 simulated models generated from the data set. This method of cross-validation was selected as previous studies indicate that, although methods such as bootstrapping and jack-knife are good measures of model accuracy, multiple-fold CV (in this study 1,000-fold) tends to yield equivalently accurate results, especially for models with multiple categorical variables (Kohavi, R., 1995).

The second method the Investigator used to determine how well the logit model predicts if a student will pass the course was through analysis of a Receiver-Operator Curve and the associated area under the curve (AUC). This curve yields the percent concordant and discordant, as discussed in the analysis section. The AUC is denoted by "*c*" and is a measure of the percent concordant as determined by the model. For this

study *c* = 0.740, or correctly predicts a student passing or not passing the course 74% of the time. Jones and Athanasiou developed a "scale" to determine the accuracy of the AUC. "A fair test shows better than average accuracy, and has an AUC above 0.5. To demonstrate excellent accuracy, the AUC should be in the region of 0.97 or above. An AUC of 0.93 to 0.96 is very good; 0.75 to 0.92 is good. Less than 0.75 can still be reasonable but the test has obvious deficiencies." They go on to say "It is important to remember that the AUC must be interpreted according to the context of the individual analysis and that these guidelines are not absolute." (Jones, C. & Athanasiou, T., 2005) Park, Goo, and Jo state that "The closer AUC is to 1, the better the overall diagnostic performance of the test, and a test with an AUC value of 1 is one that is perfectly accurate…" (Park, S.H., et al., 2004). Of course an AUC of 0.5 is merely the same as flipping a coin when trying to predict if a student passes or does not pass the course. Park, Goo, and Jo go on to say, "A diagnostic test with an AUC value greater than 0.5 is, therefore, at least better than relying on pure chance, and has at least some ability to discriminate between subjects…" (Park, S.H., et al., 2004). The Investigator argues that, for an observational study with no experimental controls, an AUC of 0.74 is high enough to provide some evidence of the impact of quizzing on student success. This study suggests that further, more careful research into the effects of quizzing on student success is warranted.

HIERARCHICAL LINEAR MODEL (HLM) / MIXED MODEL

It was shown that there is a positive association between a student's final exam score and the administration of quizzes. For example, a white, male upperclassman, with an average standardized SAT/ACT score who takes the course in the afternoon is predicted to score 6% higher on the final exam if quizzes are administered. This is consistent with the Investigator's research hypothesis for this study; in other words, students' performance in an introductory college level statistics course can be improved by the implementation of periodic quizzing. A similar study to determine if online quizzing techniques (SRS and WEB CT based quizzes) affected student achievement using a randomized experimental design was performed by researchers at the Catholic University of America. The study, as cited in the introduction of this paper, involved nursing students taking a "General, Organic, and Biochemistry course". Their study showed that WebCT-based quizzes "have a significantly positive effect on student achievement on teacher written exams." In addition, their study yields a mean score on teacher-written exams of 89.87% (s = 12.25) versus 75.18% (s = 15.41) when no quizzing was employed (Bunce, D.M., et at., 2006). Their study used multiple linear regression for analysis but did not employ an HLM to account for random effects associated with instructors.

The HLM in this study found the variation between instructors to be significant (*p*-value = 0.02895). The ICC of 0.0633 indicates that variability between instructors accounts for 6.33% of the total variation in the model. Considering that data used to

construct the model utilized the 696 observations associated with students, the variability associated with instructors is important. By employing an HLM, the "noise" in the model is reduced through accounting for instructor random effects, thereby allowing the Investigator to more accurately estimate the true effect of quizzing on final exam scores (Raudenbush, S. & Bryk, A.S., 1986).

Another interesting but somewhat expected result of the HLM is that math aptitude, accounted for by a standardized SAT or ACT score for each student, was found to be significant (p-value < 0.0001). Research by Goldstein and High (1992) indicate that math aptitude has a positive association with achievement in college level business statistics, supporting the results of this study.

Consider comparing an "average" math student, one near the 50^{th}-percentile of standardized math scores, versus a "good" math student, one who is in the upper 15^{th}-percentile (or one standard deviation above the mean). If we observe a white, male, upperclassman that takes an afternoon course and is administered quizzes, the simple effect of being a "good" math student versus an "average" math student (as measured by SAT/ACT) results in an estimated 5.66% increase on the final exam score (84.9% versus 79.24%, respectively).

The last explanatory variable that was significant within the HLM is gender. Numerous studies have been conducted to determine if there are differences in the quality of students based on gender, often with conflicting results. In a study that focused on how men and women approached taking college level courses, it was determined that women, in general, were better students (Zusman M., et al., 2005).

However studies that compared men and women in terms of math aptitude have shown that men often times outperform women (Felson, R.B. & Trudeau, L., 1991).

According to the HLM and the Tukey-Kramer adjusted t-test for gender, results from this study indicate that women scored higher on the final exam in STAT 145. Comparing the gender response estimates for white, upperclassmen with an average math aptitude (50^{th}-percentile) who takes the course in the afternoon, females are predicted to score 3 percentage points higher on the final exam. Although this value does not account for a half letter grade increase (considered to be approximatley 5%), it does appear that female students at UNM tend to perform better in STAT 145 versus their male counterparts.

CONFOUNDING VARIABLES

In this paper, the final HLM and logit models retained several explanatory variables that were not significant but may have potentially confounding effects on the response had they not been kept in the model. The variables retained are associated with variation between students that the Investigator could not control for due to the fact that this is an observational study. Leaving these variables in the model is similar to accounting for instructor variation, however the variables retained are considered fixed in the context of the model(s). The factors retained for each model are cited in Table 5.1.

Table 5.1

Confounding Variables Retained

Variable	Explanation	Model Retained In
Time	AM/PM	HLM and Logit
Gender	Female/Male	Logit
Ethnicity	Hispanic/White/Other	HLM and Logit
GPA (college)	New Student (Frosh)/Upperclassman	HLM and Logit

By retaining factors that are statistically not significant but may still influence the response variable the Investigator argues that a more accurate estimate of the influence of quizzing on the response variable can be determined. Prentice (1976) points out that retaining variables (in a logit model) that are either known or assumed to influence the response variable "leads to a direct estimation of the odds ratio associated with the (response) and of the dependence of the odds ratio on other explanatory variables." Retaining non-significant demographic variables to adjust for confounding effects in regression models is not uncommon. In a study to determine the association of air pollution and lung function growth the authors retained gender and ethnicity (as well as other non-significant, but potentially confounding variables) in their model to adjust "for subject-specific covariates" (Gauderman W.J., et al., 2000).

Multiple studies on gender and student achievement have been performed, yielding some evidence that gender can influence student learning and material retention. In particular, studies have been performed to determine if the time of day

students take classes impacts their performance, (Klein, J., 2001; Morton, L.L. & Kershner, J.R., 1985; Morton, L.L. & Kershner, J.R., 1993). Results from these studies indicate that the time of day can influence learning and retention; however other dependencies such as the age of students also influenced if they were more successful in the morning versus the afternoon. Because these studies imply that time of day does influence student learning but there is no consensus with regards to which time of day is most influential, the Investigator chose to retain this variable to account for differences between students' time choices.

Similarly, ethnicity and GPA (college) were also retained in both models developed for this study. As discussed in the methods section of this paper, GPA was converted to an indicator variable with the factor "FROSH" defining the population of students "new" to university studies (no college GPA) versus upperclassmen (UC), or students with previous college experience. The Investigator recognized that these two populations are essentially different and felt it is necessary to include a variable accounting for such differences in both models. Alternatively the study could have retained the continuous random variable "GPA" and analyzed the subset of students with previous college experience, i.e., students with a college GPA at the time of the study. The Investigator recognizes that information is lost due to converting the continuous random variable to an indicator function however we do not expect this transformation to influence the final results.

Student ethnicity and its effect on scholastic achievement has been a topic of interest debated by educators and school administrators for many years (Trueba, E.H.,

1997). Studies indicate there are differences associated with student success as it depends on ethnicity (Strage, A., 1999) however there is no definitive method for classifying ethnicity. Classification can be limited by the ethnicities a student is "allowed" to choose from when submitting demographic information to the University. As a result, the Investigator retained this factor, as well, to account for differences between students' demographic background as it pertains to ethnicity.

FUTURE STUDIES

Results from this study are consistent with the notion that quizzing has a positive impact on student success in a college level core statistics course. Because this is an observational study, the Investigator was limited in terms of "matching, randomization, random sampling, and other methods of controlling extraneous variation." (Rubin, D.B., 1974) Limitations associated for control over random effects were accounted for through use of a hierarchical linear model. Non-significant variables with potentially confounding effects on the response variables evaluated were retained to further control for variation between students. To control for variation between students and instructors the Investigator proposes that a randomized, designed experiment be performed. The purpose of a designed experiment is to be able to estimate the counter factual through random assignment. Also, to conclude that quizzing causes students to perform better, a well designed experiment is the preferred method over "the use of carefully controlled nonrandomized data to estimate causal effects." (Rubin, D.B., 1974) Rubin goes on to state that use of controlled nonrandomized data is often "a reasonable

and necessary procedure in many cases," however experimental data should be used, when possible, in lieu of nonrandomized data "especially in the social sciences where much of the variability is often unassigned to particular causes." (Rubin, D.B., 1974)

One possible experiment is to create four equal size sections for STAT 145, with two instructors teaching two sections each. Students would be randomly assigned to each section in an attempt to account for demographic differences between students (gender, ethnicity, etc.). To account for time of day, the first two sections would be taught simultaneously, with one section/instructor administering quizzes while the other does not. For example, Instructor A would teach Section 1 at 10:00 AM and administer quizzes, while Instructor B, teaching Section 2 at 10:00 AM would not administer quizzes. At 11:00 AM Instructor A would teach Section 3, but not administer quizzes while Instructor B would teach Section 4 and administer quizzes. The format of all four sections, in terms of course structure, grading format, number of quizzes, the type of quizzes, etc. would be the same. This design would help control for the potential confounding effects on the response variable of interest and allow us to conclude if quizzing causes improved student performance. Of course there still is variation between different students and different instructors; however it is hoped that this design would reduce a substantial amount of this variation, allowing for a more direct analysis of the association between quizzing and student success.

CHAPTER 6: CONCLUSION

Education administrators and instructors are constantly seeking pedagogical methods to improve student learning of core subject content in mathematics and statistics. In addition, there is a need for administrators and instructors to assess how well students understand the material being taught. This can be challenging for classes with a large number of students, as is often the case with entry level or "core" courses offered by colleges and universities. One technique of "low-stakes" assessment and instruction is the implementation of periodic quizzing.

Quizzing, when effectively administered, allows students to solve problems in a "testing environment" without the grading impact of an exam, which typically accounts for a higher percentage of students' final course grade. Another benefit for students is that, through the process of studying for and taking a quiz, they become better prepared for exams which ultimately can assist them passing the course. Quizzing also allows educators to periodically assess student understanding of specific subject matter, as quizzes usually focus on only one or two concepts. This methodology allows the instructor to augment instruction during the course of a semester so that they can focus attention on subject matter that students may be struggling with as assessed through quizzing, which further benefits the students.

This study is consistent with the hypothesis that quizzing positively impacts student achievement in a college level statistics course as measured by final letter grades and final exam scores. Although further investigations are needed to

demonstrate a causal link between quizzing and performance as well as to determine the best way to employ quizzes it, should be encouraging to students, instructors and education administrators that this pedagogical technique shows promise in assisting students to become proficient in the subjects they choose to study.

REFERENCES

1. ACT Inc. *The Condition of College & Career Readiness 2010*. 2010 <http://www.ncsl.org/issues-research/educ/improving-college-completion-reforming-remedial.aspx>

2. Agarwal, P. K., Roediger, H. L. I. I. I., McDaniel, M. A., McDermott, K. B., & Society for Research on Educational Effectiveness (SREE). "*Improving Student Learning through the Use of Classroom Quizzes: Three Years of Evidence from the Columbia Middle School Project*." Society for Research on Educational Effectiveness. 2010.

3. Bunce, D.M., VandenPlas, J.R., and Havanki, K.L. "Comparing the Effectiveness on Student Achievement of a Student Response System versus Online WebCT Quizzes." *Journal of Chemical Education.* 83 (3) 2006: 488.

4. Christensen, R. *Log-Linear Models*. 480p. New York, NY: Springer-Verlag, 1991.

5. Downer, R.G. and Richardson, P.J. "*Illustrative Logistic Regression Examples using PROC LOGISTIC: New Features in SAS/STAT® 9.2."* Paper SP03-2009. 2009.

6. Felson, R.B. and Trudeau, L. "Gender Differences in Mathematics Performance." *Social Psychology Quarterly*, Vol. 54, No. 2 Jun., 1991: pp. 113-126. Published by: American Sociological Association. Article Stable URL: <http://www.jstor.org/stable/2786930>.

7. Gauderman WJ, et al. "Association between air pollution and lung function growth in southern California children." *American Journal of Respiratory and Critical Care Medicine,* Vol. 162, 2000: 1383–90.

8. Goldstein, J. and High, R. V. "*Identifying cognitive and affective variables as they relate to the successful completion of business statistics."* Paper presented at the Adelphi University Colloquium. Garden City, New York. May, 1992. [ERIC Document Reproduction Service No. ED 350 921].

9. Greenless, J.S., Reece, W.S., and Zieschang, K.D. "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed." *Journal of the American Statistical Association*. Vol. 77, No. 378 Jun., 1982: pp. 251-261. Published by: American Statistical Association. Article Stable URL: <http://www.jstor.org/stable/2287228>.

10. Heath, Suzanne. "No Child Left Behind Act: What Teachers, Principals & School Administrators Need to Know". Wrightslaw. 2002. <http://www.wrightslaw.com/info/nclb.teachers.admins.htm>.

11. Jones, C.M. and Athanasiou, T. (2005). "Summary Receiver Operating Characteristic Curve Analysis Techniques in the Evaluation of Diagnostic Tests". *The Annals of Thoracic Surgery*. Volume 79, Issue 1, January 2005: Pages 16-20. ISSN 0003-4975, 10.1016/j.athoracsur.2004.09.040. <http://www.sciencedirect.com/science/article/pii/S0003497504019861>.

12. Klein, Joseph. "Attention, Scholastic Achievement and Timing of Lessons". *Scandinavian Journal of Educational Research*. 45:3, 2001: 301-309.

13. Kohavi, R. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In C.S. Mellish, ed., *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, Inc. 1995. pp. 1137 – 1143. http://robotics.stanford.edu/~ronnyk/ronnyk-bib.html.

14. Lee, V. E. and Bryk, A. S. "A multilevel model of the social distribution of high school achievement." *Sociology of Education*. Vol. 62, 1989: 172-192.

15. Lee, V.E., Dedrick, R.F. and Smith, J.B. "The Effect of the Social Organization of Schools on Teachers' Efficacy and Satisfaction." *Sociology of Education,* Vol. 64, No. 3, July, 1991: pp 190-208.

16. Little, R.J.A. and Rubin, D.B. *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons, 1987.

17. Lloyd, Michele E. *The Effects of Weekly Quizzing in a Ninth Grade Mathematics Classroom on Academic Achievement, Classroom Satisfaction, and Study Habits*. Oswego, NY: State University of New York, College of Arts and Science, School of Education, 1995. Print.

18. McDaniel, Mark A., et al. "Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement." *Journal of Educational Psychology*, Vol 103(2), May 2011: 399-414. doi: 10.1037/a0021782 .

19. Morling, B., et al. "Efficacy of personal response systems ("clickers") in large, introductory psychology classes." *Teaching of Psychology, 35,* 1, January 01, 2008: 45-50.

20. Morton, L.L. and Kershner, J.R. "Time of day effects upon children's memory and analogical reasoning." *The Alberta Journal of Educational Research, 31*, 1985: 26–34.

21. Morton, L.L. and Kershner, J.R. “Time of day and attentional-order influences on dichotic processing of digits in learning-disabled and normal-achieving children.” *Journal of Neuroscience, 71*, 1993: 51–61.

22. Park, S. H., Goo, J. M. and Jo, C.H. “Receiver operating characteristic (ROC) curve: practical review for radiologists.” *Korean J. Radiology*, 5, 2004: 11-18.

23. Peduzzi P, et al. “A simulation study of the number of events per variable in logistic regression analysis.” *Journal of Clinical Epidemiology*, 49, 1996: 1373-9.

24. Prentice, R. L. “Use of the logistic model in retrospective studies.” *Biometrics* 32, 1976: 599-606.

25. Raudenbush, Stephen W. “Educational Applications of Hierarchical Linear Models: A Review.” *Journal of Educational and Behavioral Statistics*, vol. 13 no. 2, June 20, 1988: 85-116.

26. Raudenbush, S. and Bryck, A.S. “A Hierarchical Model for Studying School Effects.” *Sociology of Education*, Vol. 59, No. 1, Jan., 1986: pp. 1-17. Published by: American Sociological Association. Article Stable URL: http://www.jstor.org/stable/2112482

27. Refaeilzadeh, P., Tang, L. and Liu, H. “On comparison of feature selection algorithms.” *In Association for the Advancement of Artificial Intelligence (AAAI), Workshop on Evaluation Methods for Machine Learning II*. Vancouver: AAAI Press, 2007: 34-39.

28. Roediger, H. and Karpicke, J. “The Power of Testing Memory: Basic Research and Implications for Educational Practice.” *Perspectives on Psychological Science,* 1, 3, January 01, 2006: 181-210.

29. Rubin, D.B. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”. *Journal of Educational Psychology*, Vol. 66, No. 5, 1974: 688 - 701

30. SAS Institute (2012). The output, code and data analysis for this paper was generated using SAS/STAT software, Version 9.3 of the SAS System for Windows. Copyright © 2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

31. Strage, A. A. “Social and academic integration and college success: Similarities and differences as a function of ethnicity and family educational background.” *College Student Journal*, 33, 1999: 198-205.

32. The ACT. ACT Profile Report – National, Graduating Class 2010, National. 2010.

33. The College Board. 2010 College-Bound Seniors Total Group Profile Report. 2010.

34. Trueba, E.H. “Ethnicity and education forum: What difference does difference make.” *Harvard Educational Review*, Vol. 67-2, 1997: 169. ISSN: 0017-8055.

35. Urtel, M. G., et al. “On-Line Quizzing and Its Effect on Student Engagement and Academic Performance.” *Journal of Scholarship of Teaching and Learning, 6,* 2, October 01, 2006: 84-92.

36. Van Dongen, S., Lens, L. and Molenberghs, G. (1999) *“*Recent Developments and Shortcomings in the Analysis of Individual Asymmetry: A Review and Introduction of a Bayesian Statistical Approach”*. In M.Polak* (ed.). *Developmental Instability: Causes and Consequences*. New York: Oxford University Press, 2003: pp. *325*.

37. Volante L. “Teaching to the Test: What Every Educator and Policy-Maker Should Know.” *Canadian Journal Of Educational Administration And Policy.* [serial on the Internet]. (2004, Sep 25), [cited June 12, 2012]; (35): Available from: ERIC.

38. Zusman M., Knox D., and Lieberman M. “Gender Differences In Reactions To College Course Requirements or “Why Females Are Better Students”.” *College Student Journal [serial online],* 39 (4), December 2005: 621-626. Available from: SPORTDiscus with Full Text, Ipswich, MA. Accessed June 7, 2012.

APPENDIX A

POST – SEMESTER QUESTIONNAIRE:

Research Study: Post- Semester Questionnaire (Fall 2011)

"The Relationship of Quizzing and Student Success in A College Level Core Statistics Course"

You have agreed to participate in a research study being conducted by David M. Glavin, Graduate TA in Statistics on the relationship between quizzing and student success in college level core statistics courses.

Please answer the following questions regarding the section of STAT 145 you taught during the Fall 2011 Semester at the University of New Mexico:

1. *List the section number, days of week and associated time you taught STAT 145 during the Fall Semester, 2011:*

Section No.	Day(s) of Week	Time
_____	_____	_____

2. *Did you administer quizzes as part of instruction for your section?*

YES NO

If you answered **YES** respond to questions (3) – (7).

If you answered **NO** skip to question (8).

3. *How many quizzes (total number) did you administer during the semester?*

4. *How often were quizzes administered (approximate # per week or per exam, please specify)?*

5. *On average, approximately how long was each quiz (number of questions)?*

6. *How much was **each** quiz worth in terms of the students' overall grade (percent of final grade)? Please indicate if quizzes administered did not have any value with respect to overall grade.*

7. *How was the quiz administered (i.e.: written, web-based, clicker/PowerPoint, take-home)? If multiple methods were used please indicate the methods and number of quizzes administered for each method.*

8. ***Submission of Final Exam Scores and Final Letter Grade***

    a. *Please use an MS Excel Spreadsheet to submit this information.*
    b. ***Strip all student identification information (name, banner id, etc.) from the Excel Spreadsheet. The spreadsheet should include only three columns**:*
        i. *The Research Identification Numbers (RIDN) provided by the UNM Dept. of Mathematics and Statistics Department's IT Support Manager*
        ii. *Final Exam Scores associated with the student's RIDN*
        iii. *Final Letter Grades associated with the student's RIDN*
    c. *If a student has withdrawn prior to the Final Exam, please leave this cell blank and enter only the appropriate Final Letter Grade (WP, WF, I, F, etc.)*
    d. *Please email the Excel Spreadsheet. to: dglavin@unm.edu. If this is not feasible, please submit a printed copy of the spreadsheet in a sealed envelope to the Dept. of Mathematics and Statistics Office located on the 2nd Floor of the Science and Math Learning Center (SMLC), addressed to: **ATTN: David M. Glavin, MSC01 1115.***

Please sign and date this form. A signed hard-copy of this form will be returned to you by David M. Glavin by no later than Jan. 31, 2012.

Name of Instructor (printed)	Instructor's Signature / Date
Name of Researcher (printed)	Researcher's Signature / Date

Supplementary Figures and Tables

Table A.1
Separation Table – Including Factor: TIME

Obs	QUIZ	Gender	race	TIME	COUNT	PERCENT
1	N	F	H	AM	37	5.3161
2	N	F	H	PM	82	11.7816
3	N	F	O	AM	32	4.5977
4	N	F	O	PM	41	5.8908
5	N	F	W	AM	44	6.3218
6	N	F	W	PM	71	10.2011
7	N	M	H	AM	30	4.3103
8	N	M	H	PM	34	4.8851
9	N	M	O	AM	10	1.4368
10	N	M	O	PM	24	3.4483
11	N	M	W	AM	20	2.8736
12	N	M	W	PM	52	7.4713
13	Q	F	H	AM	35	5.0287
14	Q	F	H	PM	6	0.8621
15	Q	F	O	AM	21	3.0172
16	Q	F	O	PM	8	1.1494
17	Q	F	W	AM	44	6.3218
18	Q	F	W	PM	7	1.0057
19	Q	M	H	AM	29	4.1667
20	Q	M	H	PM	9	1.2931
21	Q	M	O	AM	17	2.4425
22	Q	M	O	PM	6	0.8621
23	Q	M	W	AM	32	4.5977
24	Q	M	W	PM	5	0.7184

Table A.2

Ethnicity Least Squares Means

Ethnicity	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
H	5476.71	252.82	680	21.66	<.0001	0.05	4980.32	5973.11
O	5484.13	274.79	680	19.96	<.0001	0.05	4944.60	6023.67
W	5195.49	249.46	680	20.83	<.0001	0.05	4705.69	5685.30

Table A.3

Tukey-Kramer Adjusted t-test for Ethnicity

Differences of Ethnicity Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
Ethnicity	Ethnicity	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
H	O	-7.4198	233.11	680	-0.03	0.9746	0.9994	0.05	-465.13	450.29	-554.97	540.13
H	W	281.22	202.53	680	1.39	0.1654	0.3475	0.05	-116.45	678.89	-194.50	756.94
O	W	288.64	231.23	680	1.25	0.2123	0.4252	0.05	-165.36	742.64	-254.48	831.76

Figure A.1

LS Means Comparison of Final Exam Score-Squared by Ethnicity

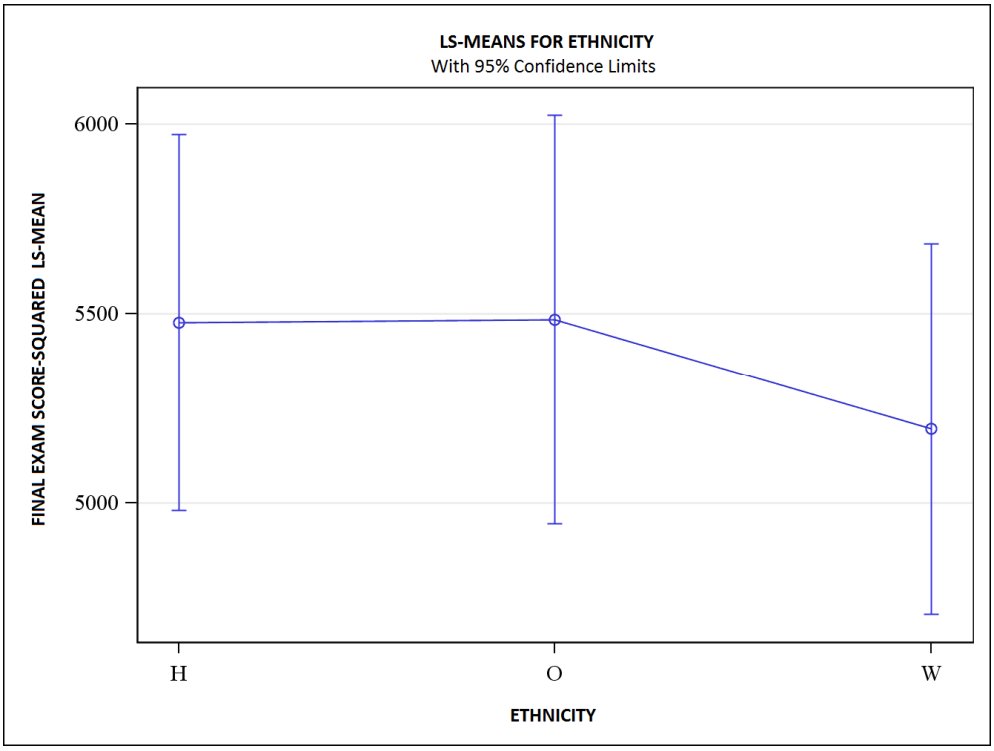


Table A.4

Time Least Squares Means

TIME	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
PM	5443.04	262.61	680	20.73	<.0001	0.05	4927.42	5958.67
AM	5327.85	257.57	680	20.69	<.0001	0.05	4822.13	5833.57

Table A.5

Tukey-Kramer Adjusted t-test for Time

Differences of TIME Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
TIME	TIME	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
PM	AM	115.19	260.54	680	0.44	0.6585	0.6585	0.05	-396.37	626.75	-396.37	626.75

Figure A.2

LS Means Comparison of Final Exam Score-Squared by Time

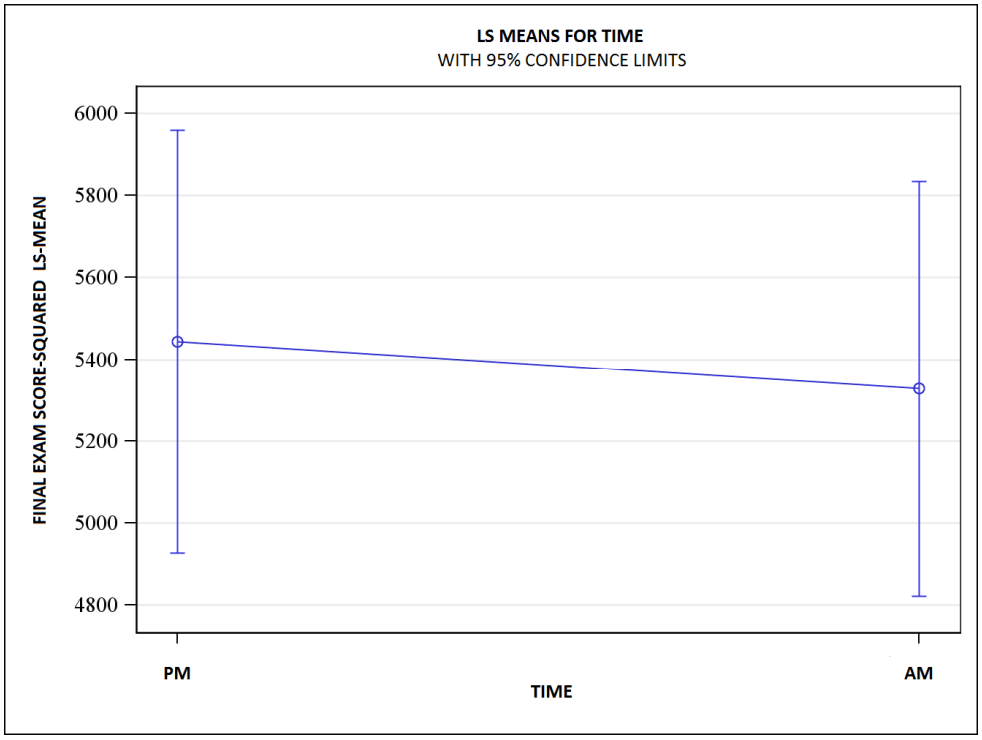


Table A.6

GPA (college) Least Squares Means

(Using indicator values for freshmen and upper classmen)

GPA	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Frosh	5343.90	281.23	680	19.00	<.0001	0.05	4791.71	5896.10
UC	5426.99	219.76	680	24.70	<.0001	0.05	4995.51	5858.47

Table A.7

Tukey-Kramer Adjusted t-test for GPA (college)

(Using indicator values for freshmen and upper classmen)

Differences of GPA (college) Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
GPA	GPA	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Frosh	UC	-83.0865	228.13	680	-0.36	0.7158	0.7158	0.05	-531.01	364.84	-531.01	364.84

Figure A.3

LS Means Comparison of Final Exam Score-Squared by GPA (college)

(Using indicator values for freshmen and upper classmen)