8-22-2018

Analysis of Grade Variances in Multisection Calculus Classes

S. Paul Shaffer
Illinois State University, shaffersp@yahoo.com

ANALYSIS OF GRADE VARIANCES IN MULTISECTION CALCULUS CLASSES

S. Paul Shaffer

114 Pages

The purpose of this dissertation was to explore grade variances associated with multisection calculus grades. The method for the project was quantitative examining grade distribution patterns that existed between calculus sections at a public university void of an overarching grading policy during the case study. Results showed that different grade distribution patterns by different instructors were primarily due to student-centric characteristics. The significant findings suggest that student grades are firmly in control of students. Only when course grade averages were at the extreme did evidence suggest that instructor-centric characteristics played meaningful role in student grade determination.

KEYWORDS: Calculus, Grading, University

ANALYSIS OF GRADE VARIANCES IN MULTISECTION CALCULUS CLASSES

S. PAUL SHAFFER

A Dissertation Submitted in Partial
Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Department of Educational Administration and Foundations

ILLINOIS STATE UNIVERSITY

2018

ANALYSIS OF GRADE VARIANCES IN MULTISECTION CALCULUS CLASSES

S. PAUL SHAFFER

COMMITTEE MEMBERS:

Elizabeth Lugg, Chair

Olcay Akman

Diane Dean

John Rugut

ACKNOWLEDGMENTS

The purpose of this acknowledgement is to thank those individuals who have provided assistance, guidance, and encouragement through the number of years it takes to pursue a PhD. Commitment and effort were fundamental elements for the completion of my doctoral dissertation, but even more was the support of my family. Without their support, this journey could not have been successful nor would I have had the fortitude to endure it. To my wife and the two greatest projects of my life: my daughters Samantha and Annabelle, I dedicate this professional achievement because without their presence and support, I would have not achieved my goal.

I would like to thank Dr. Elizabeth Lugg who persevered with me as both my advisor and dissertation chair. Her willingness to believe in me is a testament to how special a person she truly is. Quite simply, this dissertation would not have come to fruition without her patience, guidance and support. A special thanks to Dr. Diane Dean, Dr. John Rugutt and Dr. Olcay Akman is also warranted as they provided valuable insight serving on my dissertation committee.

I would also like thank my grandmother, Doris Patterson for believing in me and encouraging me so many years ago that life is worth living. Lastly, I cannot say enough about my father who battled courageously for over two years with cancer. His guidance over these many years was invaluable as he taught all of those around him to live life to the fullest. Words cannot begin to express the admiration I had for him and how fortunate I was to be his son. I love you and miss you deeply!

S. P. S.

CONTENTS

FIGURES

TABLES

CHAPTER I: THE PROBLEM AND ITS BACKGROUND

Importance of the Topic

Assessing college students used to be a collegial affair between students about to graduate and professors done to communicate the progress that students had made (Smallwood, 1935). Today, grades attained in college have evolved to become the fundamental currency of our educational system signaling both "academic achievement and noncognitive skills to parents, employers, postsecondary gatekeepers, and students themselves" (Pattison, Grodsky, & Muller, 2013, p. 259). Today's students recognize this and view grades as a key determinant in their future job prospects and admittance to graduate programs (Becker, Geer, & Hughes, 1995).

Given the importance of grades, it is not surprising that grading practices in higher education have attracted significant attention inside and outside academia (Bar, Kadiyali, & Zussman, 2009; Hu, 2005). While there is extensive literature examining the impact of letter grades on student course selection and major choice (Rask, 2010; Rask & Tiefenthaler, 2008; Sabot & Wakeman-Linn, 1991), studies reported in higher education literature provide little empirical data about how the interaction of individual professor and student characteristics affect student course grades. The goal of this study is to uncover pertinent student and instructor characteristics that contribute to our understanding of the variances in the grades of students in the multisection calculus classes observed.

From a university administrative perspective, low success rates in calculus courses can have wide-ranging implications for the university itself. If disadvantaged students are unable to secure gainful employment in their chosen field, it may become increasingly difficult for the university to increase future student enrollment and alumni donations. These unsatisfied students could channel their displeasure through multiple media outlets, further amplifying problems

associated with uneven grading. If future enrollment numbers decrease, a university's ranking, prestige, and incoming tuition dollars may suffer as a result.

With this dissertation, I examine correlates and predictors of multisection Calculus I course grades at a public university using hierarchical linear modeling (HLM) for the purpose of increasing student success rates in multisection Calculus I courses. In Chapter II, I provide a historical overview of grades in higher education, along with a review of the contemporary literature on grades through a role theory lens. In Chapter III, I describe the quantitative approach and steps taken using IBM SPSS Statistics 20 (SPSS) and two-level hierarchical linear modeling (HLM) for this study. I use Chapter IV to examine the results of regression and multilevel analysis of student and professor characteristics in an effort to create models that explain the variances for grade disparities uncovered among students in different Calculus I sections. Finally, in Chapter V, I provide an examination of the significant findings of the study and describe potential solutions to increase student success rates in multisection calculus courses.

Background

There are typically between seven and ten sections of Calculus I offered each semester at the university studied. Students are allowed to register for any Calculus I section based on factors such as seat openings, convenience of class lectures, and fit with the rest of their individual schedules. At the time of course registration, instructor assignments to each section were not publicly posted. The public assignment of instructors was typically not made available to students until a few weeks before the first class session. As such, the identity of the course instructor played no role for the overwhelmingly majority of students when deciding what Calculus I section to register for during open course registration.

At the end of each semester, the Provost's office at the university studied produces a report entitled "Student Grade Summary Report." The report contains detailed student grade distributions for each class taught in the previous semester. The fall 2012 report showed a wide discrepancy in grades received by students varying by Calculus I sections. The report uncovered some sections where As or Bs were awarded to over half the students while other sections had a student failure rate approaching fifty percent. This activity may violate what Younts and Mueller (2001) described as distributive justice, or "the fairness of the outcomes or rewards that an individual or group receives" (p. 125).

Grade discrepancies found in the "Student Grade Summary Report" may have resulted from gender or racial bias, professor grading standards, student ability, or something else entirely. With this in mind, further exploration of the current grading and characteristics of students and professors found in Calculus I classes was warranted to help explain the grading discrepancies uncovered with the goal of looking for ways to increase student pass rates.

University Demographics

The university studied is an established public university located in a small, Midwestern city. The university has a total enrollment of approximately 20,000 students with the sizable number of the students being undergraduates who live on campus. The percentage of students admitted hovers around 80% with admitted students having an average American College Testing (ACT) score of 24 and high school GPA of 3.36 on a 4.0 scale (Princeton Review, n.d.). Over 55% of the students are female with close to 20% of the student body consisting of minority students (Princeton Review, n.d.). First-year student retention is approximately 80% with about half of the students graduating in four years (Princeton Review, n.d.). The calculus

classes studied were taught by both tenure track and nontenure track faculty members. Each professor teaching these courses had a PhD in mathematics or similar field.

Students are typically required to pass a course in precalculus or have scored 27 or better on the ACT math section before being allowed into Calculus I. A passing grade of C or better is required in order to move on to the next calculus class. Each calculus class examines a number of calculus-based concepts that become more advanced in the following class. Typically, each Calculus I section is taught in a traditional format and has approximately 28 students.

Purpose of Grades

Historically, grades were developed to provide information to students and help in administrative functions (Smallwood, 1935). As universities grew in complexity, grades continued to evolve over time to serve multiple purposes for multiple stakeholders. In addition, the issue of grade meanings varying over time with grades not conveying the same message or importance for the grade giver, the grade receiver, or the grade user is only one of the many issues surrounding grading. (Milton, Pollio, & Eison, 1986).

According to studies by Marzano (2000) and Walvoord and Anderson (2010), both educators and noneducators perceive providing information about student achievement as the primary purpose of grading. Grades provided a vehicle by which academic performance information is conveyed to students, parents, future employees, graduate schools, and others (Hunt, 2008). This information hopefully conveys some degree of how much actual learning the student has achieved over a given period of time (Milton et al., 1986).

Marzano (2000) also noted that professors use grades for instructional planning. How a student performs on a test or given set of tasks can provide feedback to professors and teachers alike to aid in grouping students or deciding how much time to allocate to various subjects

(Austin & McCann, 1992). Grades help students identify areas of special ability or inability, in deciding on the advisability of enrolling in certain courses and staying away from others, and in determining course load (National Education Association (NEA), 1974). Most "students want to know, and have a legitimate interest in knowing, whether their performance conforms to standards…at the current stage" in their educational career (Hunt, 2008, p. 74).

Rightly or wrongly, grades also serve as both motivational and disciplinary tools for students (NEA, 1974). For a certain percentage of students, grades are used as a pedagogical tool that motivates them to perform well in order to be eligible for scholarships, improved employment prospects, and graduate school opportunities (Brighouse, 2008). It is not surprising to find that a high GPA is required in order to be considered for these scarce positions. As a result, grades serve as a determinant to access enviable jobs and prestigious graduate programs.

A final purpose of grades is to help in administrative functions, as grades provide an efficient way for administrators to "distinguish between the quality of performance of the students in a given class" or program (Hunt, 2008, p. xvii). Grades help colleges and universities screen prospective applicants, they help professors and administrators differentiate students, and they indicate whether a student has passed or failed a course, whether a student should repeat a course, and whether a student has met graduation requirements (NEA, 1974). In this way, various parties use grades to compare, rank, and sort students in any number of ways.

Purpose of Study

> In 1913, I. E. Finkelstein opined in *The Marking System in Theory and Practice*: When we consider the practically universal use in all educational institutions of a system of marks, whether numbers or letters, to indicate scholastic attainment of the pupils or students in these institutions, and when we remember how very great stress is laid by teachers and pupils alike upon these marks as real measures or indicators of attainment, we can but be astonished at the blind faith that has been felt in the reliability of the marking system. School administrators have been using with confidence an absolutely uncalibrated instrument….What faults appear in the marking system that we are now using, and how can these be avoided or minimized? (p. 1)

The purpose of this empirical, quantitative, and exploratory study is threefold. One purpose is to identify the different grading patterns existing in multisection calculus classes. Another purpose of this study is to identify what student and instructor characteristics, using regression analysis and HLM procedures, are pertinent in explaining variances in student course grades. It could be that any interactions between student and instructor characteristics may have an additional influence on course grades. Lastly, the results of this study will help professors and administrators gain a deeper understanding of the role student and professor characteristics play in our understanding of determining a student's grade with the ultimate goal of improving student success rates in calculus. While grading inconsistencies exist in virtually every class, department, and university, the magnitude and origins of these inconsistencies are not well understood.

Sung (2003) argued that one of the core missions of public universities should be the maintenance of standardized grading policies. Without a uniform grading philosophy, professors are almost entirely in charge of what constitutes earning an A, B, or C grade. Grade disparities across multisection courses based upon individual faculty grading philosophies rather than student merit is unfair. Faculty grading philosophies are numerous and may arise from such things as employment status or personal connection to the university (Schutz, Drake, & Lessner, 2013). As such, these inconsistent standards have led, in part, to a "declining legitimacy of grades among stakeholders, especially employers" (Volkwein & Yin, 2010, p. 142).

Milton, Pollio, and Eison(1986) lament that efforts to improve grading systems have tended to focus on rank ordering rather than the teaching and learning possibilities inherent in course assignment, tests, and grades. Given the salience of grading, it is surprising that universities generally do not consider tailoring the student class enrollment process to consider the individual and professor characteristics that make up each class in order to improve student success rates. It would seem that universities and students are in need of better qualitative and quantitative tools in regard to the student class enrollment process in order to improve the consistency of grading in multisection courses and to reduce the grading variance.

This paper seeks to fill an important gap in the research literature using hierarchical linear modeling to better understand how student grades are affected by the complex interaction of individual student and professor characteristics. The results of the SPSS and HLM analyses completed in this study may be used to better match students to multisection Calculus I courses with the goal of increasing their chances of success based on student and professor characteristics.

Research Questions

Multisectioned courses that have class averages variegate by an entire grade point does not prove anything concerning different standards being applied by different professors. Students in a particular class may be far superior academically in one section. Class size or the time that a class was scheduled may have influenced the final grades that students received in any given section. Students in one particular class may have been far more motivated as well. With this in mind, four broad research questions are addressed in this dissertation:

1. Are there differences in student Calculus I grades among the college students between course sections?
2. To what extent do student characteristics (gender, ACT, high school GPA, etc.) explain differences in course grades?
3. To what extent do instructor characteristics (age, tenure, experience, gender, etc.) explain the differences in mean course grades?
4. Do instructor characteristics influence the magnitude of the student success measures on student course grades?

Methodology

For this study, I collected and examined grade data from over 2,000 students on the campus of a four-year regional public university. Phase one of the project involved gathering data from every multisection Calculus I class taken in the calendar years 2009–2013. I retrieved the raw data from the university's mainframe system. Phase two involved inputting the data in a useable format into Excel and IBM SPSS Statistics 20 to analyze. I developed several Excel spreadsheets for each of the classes analyzed. Phase three involved using HLM to develop a two-level model (with a student level and a professor level) to understand interactions between

student and professor characteristics and individual grades achieved in Calculus I. Phase four involved comparing and contrasting the findings to help understand reasons for different grading patterns with the goal of increasing student success rates.

The calculus classes chosen for study were classes typically taken primarily by either freshmen or sophomores. By selecting these classes, the goal was to get a more homogeneous mix of students. In addition, it was felt that by selecting classes geared for freshmen and sophomores, their ACT scores and high school GPAs would be more relevant. All of the seven to ten sections of Calculus I typically offered each semester used the same textbook but all homework, exams, grading scales, and teaching methods were unique to each section.

Limitations of the Study

As a quantitative study conducted on a single public campus, a number of limitations apply to the study. The study was convenient in being conducted on a campus at a four-year single Midwestern public university but may not be representative of other universities and colleges across different regions. While other universities would have made such grade data available upon request, it would have come in several different formats and at considerable expense. It is recognized that sampling the grade distribution patterns of multisection calculus classes on other campuses would have established greater generalizability to the findings in this dissertation.

It should be noted that all college and high school GPAs presented in this paper are based on a four-point scale. In addition, college GPAs used for this study were for coursework achieved only at the university studied. It is the policy of the university studied to only consider courses completed at the university to be factored into a student's overall GPA. University policy dictates that if a student retakes a course, the student's overall GPA will factor in only the results of their last attempt.

A limitation of the research design was that other unaccounted factors might have influenced the diverse grading patterns uncovered. There were hosts of socioeconomic factors unavailable to incorporate into the predictive models for this study. The vast majority of students attending the university under study took the ACT test prior to admittance to the institution. This study utilized high school GPAs, the ACT composite score, and ACT subscores as predictors of students' successes in college and as a proxy for making certain that class makeups were relatively equal in terms of students' abilities. It is recognized that making direct comparisons by this means is an imperfect measure for capturing differences in the abilities of students. Interviewing students to measure such things as effort put forth towards their studies, ethnic background, and hours worked while attending school may have improved the validity of the results uncovered in this study.

Interviewing instructors on their grading philosophy may also have improved the validity of the results uncovered in this study. This study utilized average section grades taught by each instructor as a proxy for each instructor's grading philosophy. It is recognized that making direct comparisons by this means is an imperfect measure for capturing differences in grading philosophies among instructors.

Data for this study was collected in January 2015. A few of the students for which data was collected were still enrolled at the university at the time of data collection. A student's overall GPA and Course Hours Completed may have changed slightly since the time the data was collected. It is believed that this issue is immaterial given the large number of students in the overall study.

I made every effort to cleanse the data, but it is impossible to ensure that all data extracted from the institutional database is without error. The possibility of miscoded data exists along with other unknown factors that may have introduced inaccuracies into the data set.

Summary

While research on grading in multisection classes in higher education is relatively limited, there has been no shortage of debate about grades (Allen, 2005; Hobbs, 1974; Hu, 2005; Kohn, 2012). What one takes away from these debates is that grades are not fixed across institutions, subjects, or time. The meaning of grades depends on each individual's frame of reference (Milton et al., 1986). A student, parent, professor, school official, college, or job recruiter each has a different perspective on grades. Within each of these categories, a cornucopia of grade meanings exists. A professor may view letter grades "to stand for real differences in ability between pupils [while others] merely temporary successes or failures in a series of test and quizzes; often he is not certain after all his pains that they represent the facts accurately in either case" (Dearborn, 1910, p. 5). In any university, there are a myriad of ways professors teach and grade. It is within this environment that multisection courses exist. As a result, students taking these multisection classes at institutions without consistent grading standards stand to be negatively affected academically by different grade patterns.

As such, this study aims to show, using SPSS and HLM, that the potential exists for changing the course registration process to better match students with professors in hopes of improved student outcomes. With improved student outcomes in multisection Calculus I, it is possible that students can benefit in a variety of ways. Research by Butcher, McEwan, and Weerapana (2014) provided evidence that students' choices about courses and majors are sensitive to grades. Other research has further shown that grades play a large role in a student's decision to take a subsequent course in a discipline (Sabot & Wakeman-Linn, 1991). The

importance of grades achieved early in a student's college career plays an outsized role in the student's success. A study of undergraduate students showed students' first quarter GPA to be a strong predictor of retention (Murtaugh, Burns, & Schuster, 1999). Crocker, Karpinski, Quinn, and Chase (2003) noted that some students measure their "value or worth as a human being because they stake their self-worth…on their academic performance" (p. 507).

With this in mind, it is not hard to believe that success in early semesters at college ultimately influences students' post-college experiences, such as career choice, personal income, and level of success, as well as the degree and nature of participation in community life. As such, the experience that a student has in the introductory college classes can have a significant influence on the course of that student's adult life. By creating a best-fit statistical model between students and professors based on the studied characteristics of both professors and students, department chairs and university administrators may be more open to changing the course registration process to incorporate some of the findings from this dissertation.

CHAPTER II: LITERATUR REVIEW AND CONTEMPORARY THEMES

Literature Review

Grades have not always been a part of student assessment in the United States (Marzano, 2000; Schneider & Hutt, 2014). Just as it took time for colleges and universities to appear in America, it took time before grades were introduced onto college campuses. Drawing on researchers, I will present a concise history of grading practices in American education.

Early History of Assessing Students

Established in 1636, Harvard University is the oldest university in the United States with the first students on campus typically being local young men ages 14–19 who studied a mix of religion, philosophy, Greek and Hebrew, history, and natural science (Smallwood, 1935; Thelin, 2004). Student assessments during this time consisted of faculty members providing narrative feedback to students based primarily on their ability (Marzano, 2000). In addition to students' scholastic merit, the judging of students during this early period included such things as parents' social status, attendance, obedience to faculty members, and religious worship (Smallwood, 1935). Failure to follow university rules could result in fines and whippings or, in severe cases, expulsion, also known as "rustication" (Thelin, 2004).

Yale University opened its doors in 1701 to become the third oldest university in the United States and, like Harvard, was established in order to train local clergy and political leaders (Smallwood, 1935). Until 1775, degrees at Yale and other universities "were obtained by the students passing a public examination after remaining in residence at a college for a varying number of years" (Smallwood, 1935, p. 9). These public, oral examinations were of two types: disputations (a type of frank discussion or debate) and declamations (a type of poetry or prose recitation) (Schneider & Hutt, 2014). Yale University was also the first known university to rank

students. According to the diary of then-president Ezra Stiles, Yale awarded 58 students one of four ranks (twenty Optimi, sixteen second Optimi, twelve Inferiores, ten Pejores) on their examinations in 1785 (Stiles, 1901). Former Yale student Charles Bristed (1963) noted that while these rankings were determined by the perceived learnedness of a student's response and by their ability to demonstrate knowledge publicly, they were imperfect rankings since there was no objective means of determining who the best in each class was.

From my research, it appeared that during most of the 19th century, universities experimented with various marking systems to assess students with no grading system becoming a de facto standard. By 1813, Yale University had adopted a four-point numeric scale and recorded the average marks of students into an actual book (Smallwood, 1935). Harvard University switched from a 20-point scale to a 100-point scale for assessing students and grouped students into one of five categories (Milton et al., 1986). By 1860, the University of Michigan was marking students on a scale of 1–100, with 50 being the passing mark (Durm, 1993). By 1877, Harvard University had switched marking systems again and was now classifying students into one of six divisions (Smallwood, 1935). It should be noted that several colleges still made use of descriptive adjectives when ranking students, foregoing any type of numerical scale (Milton et al., 1986).

The College of William and Mary in the early 19th century sent the following report to parents about their sons:

> The Society took into consideration the situation of the different Classes, and the demeanors and improvement of each of the Students during the Course which has just terminated, Whereupon Resolved that a Communication to the following

effect be transmitted to the Parent or Guardian of every young man at the Institution.

No. 1. (Names listed) The first in their respective classes, orderly and attentive and have made the most flattering improvement.

No. 2. Orderly, correct and attentive and their Improvement has been respectable.

No. 3. They have made very little improvement and as we apprehend from want of Diligence.

No. 4. They have learnt little or nothing and we believe on account of escapade and Idleness. (Smallwood, 1935, p. 44)

Given both the number of marking systems and constant changes to them by universities in the 19th century, it is safe to conclude that both professors and universities were not wholly satisfied with the current assessment processes. Milton et al. (1986) noted that some professors at Harvard held the belief that minute distinctions in student achievement could be made and assessed while others admitted that it was impossible to get within five or ten percent of absolute exactness. Letter grades would have to wait until the latter part of the century before being introduced on college campuses as a potential solution.

Competition for gaining admittance to college started to become a major factor after the Civil War due to a number of confluences (NEA, 1974). Previously, at most elite universities, the ability of a male student to afford college and a reasonable degree of intelligence was all that was required to gain admission. Educational enrollments in grades K–12 greatly increased the number of college-eligible students due to compulsory schooling enforcement and newly enacted child labor laws that came into existence after the war (Schneider & Hutt, 2014). Between 1870 and 1910, the number of public high schools exploded from 500 to 10,000 to meet the increased

demand for education (Gutek, 1986). Some colleges had now loosened their Greek and Latin entrance requirements, which aided public high school students, who traditionally were not offered these classes, in gaining admittance to college (Giles, 1926). In addition, women were now gaining admittance to some previously male-only universities, further swelling college enrollments (Thelin, 2004).

In 1883, Harvard introduced letter grades, a system that spread quickly across other college campuses (Hammons & Barnsley, 1992). Student grades from high school were now starting to play a factor in college admissions as a method of sorting the increasing number of students (NEA, 1974). As a result of these and other events playing out across the country, grades provided the rapidly expanding American educational system "a unified and scalable mechanism for measurement and communication" between high schools and colleges as students were becoming increasingly mobile (Schneider & Hutt, 2014, p. 208).

Due to increasing enrollments at colleges after the Civil War, it would seem inevitable that assessing students based on oral examinations was going to decline as various numerical ranking systems began to solidify their status on campuses. Horace Mann had been arguing for some time that these "new examinations" were better because of their impartiality and greater completeness (Ruch, 1929). Students at some universities were now allowed to choose certain courses and avoid others, and this further undermined the ability of professors to rank students as any one professor would not have seen every student in a class (Schneider & Hutt, 2014). While professors and administrators were coming to a general consensus about utilizing grades to evaluate students, the actual assigning and distribution of grading was far from settled territory as problems soon arose "with the range of variability and reliability of letter grades as institutional guidelines for failure differed greatly" (Hammons & Barnsley, 1992, p. 52).

Smallwood (1935) noted that in 1870, Harvard University passed a resolution that grades should only be assigned based on academic achievement. From the Harvard University Faculty Records, Smallwood uncovered the official results of the vote that said:

> Voted, that the deductions from the scale of scholarship for absence, tardiness and misconduct, as well as those which accompany the various admonitions shall be discontinued; but the numbers representing such deductions and other similar numerical marks of censure, which the Faculty may establish shall be regularly entered in the Dean's office, and shall form a separate scale of demerit. (1935, p. 74)

Interestingly, one of the first known controversies surrounding grades also arose at Harvard University when a 1894 report from *The Committee on Raising the Standard* contained:

> Grades A and B are sometimes given too readily—Grade A for work of no very high merit, and Grade B for work not far above mediocrity.…One of the chief obstacles to raising the standard of the degree is the readiness with which insincere students gain passable grades by sham work. (p. na)

The century would end with numerous universities using some form of a 0–100 grading scale in combination with letter grades (Smallwood, 1935). The issue of assignment and distribution of grades would soon attract serious scholarly research in the upcoming years that would result in professors challenging the purpose of grades, grade distributions, calculations, and their benefits to students.

Studies Between 1900 and 1930

Numerous studies conducted between 1900 and 1930 shed light on the subjective nature of grading with a number of researchers calling for grades to become clearer and more objective.

Rugg (1918) noted that more than sixty articles appeared in various journals discussing the problems over the standardization of school marks during the previous decade. Influential studies by Missouri professor Max Meyer in 1908, Cornell professor Ira Finkelstein in 1913, and Wisconsin professors Daniel Starch and Edward Elliott in 1912 and 1913 dramatically highlighted the subjective nature of grades that reflected the value systems of educators in the early part of the 20th century.

M. Meyer (1908) examined grading practices at the University of Missouri after a fellow professor had failed his entire class. M. Meyer's investigation, which examined grades in forty classes with most of them in the College of Arts and Science, illuminated a wide number of grading distributions that varied by both professor and department. M. Meyer's work also uncovered evidence of students selecting both courses and professors based on class grade distributions in order to improve their grades. As a solution to these two issues, M. Meyer proposed that grades for each class be based on a binomial distribution curve in which the top 3% of students would be called excellent; the next 22%, superior; the middle 50%, medium; the next lowest 22%, inferior; and the bottom 3%, failure, with the assignment of letter grades A, B, C, D, and F, respectively. M. I. Meyer (1911) would later write: "the grade has in more than one sense a cash value, and if there is no uniformity of grading in an institution, this means directly that values are stolen from some and undeservedly presented to others" (p. 661).

In 1913, I. E. Finkelstein opined in *The Marking System in Theory and Practice*:

> Variability in the marks given for the same subject and to the same pupils by different instructors is so great as frequently to work real injustice to the students…Nor may anyone seek refuge in the assertion that the marks of the students are of little real importance. The evidence is clear that marks constitute a

> very real and a very strong inducement to work, that they are accepted as real and fairly exact measurements of ability or of performance. Moreover, they not infrequently are determiners of the student's career. (p. 6)

Finkelstein (1913) examined student grades at Cornell University from the years 1902, 1903, and 1911. Like Meyer (1911), Finkelstein uncovered a wide number of grading distribution patterns varying by both professor and department. He found evidence that courses in pure and applied science tended to have the stingiest grading distributions in the college. To rectify this situation, Finkelstein proposed a distribution curve in which the top 3% of students would be called excellent; the next 21%, superior; the middle 45%, medium; the next lowest 19%, inferior; and the bottom 12%, very poor, with the assignment of letter grades A, B, C, D, and F, respectively. In addition, Finkelstein proposed that each class's grade distribution be made public.

During this same time period, University of Wisconsin professors Daniel Starch and Edward Elliott (1912) believed inequalities in grading standards existed throughout both high schools and colleges. They believed these inequalities were largely due to different personal values and expectations held by teachers when assessing students. To prove this, Starch and Elliott mailed two high school English papers, written by two high school students, to English teachers in 200 different high schools throughout the Midwest to be graded. One hundred fifty-two high school teachers responded by carefully examining and grading the English papers. The returned high school grade results showed a wide grade disparity among the returned papers. On a scale of 0–100, Starch and Elliott's findings showed a range of marks to be as large as 35–40 points. Approximately 12% of the respondents gave scores in excess of 90 while over 12% of the graders gave a failing score to at least one of the English papers. Starch and Elliott concluded

that grades were far less precise than teachers and students believed. In addition, both professors noted the absurdity of trying to come up with a precise grade such as 79.9%, given the noted wide distribution in grading.

Critics of the research conducted by Starch and Elliott pointed to the inherent subjective nature of grading English papers. To appease and counter their critics, Starch and Elliott (1913) sent two high school students' geometry finals and answer key to 180 high schools to be graded by math teachers. One hundred forty high school math teachers responded with results showing the variability in grading even greater than grades assigned to the English papers with one geometry final exhibiting a 67 grade point range (28 to 95). Reasons put forth by Starch and Elliott for the grade variability was that some teachers assigned partial credit and deducted points for spelling and/or poor form while other teachers did not. The amounts deducted for these infractions varied as well.

Given the discontent with current grading practices, it is no surprise that standardized testing (objective testing), multiple choice, and true/false tests began making their first appearances as ways to lessen the perceived ills of grading. Standardized tests were, in part, developed as a way for the military to screen applicants during World War I (NEA, 1974). These aptitude tests provided a method for the military to use a common set of benchmarks for incoming recruits. One of the creators of these aptitude tests, First Lieutenant Carl Brigham, would later go on to become a professor at Princeton University and create the Scholastic Aptitude Test (SAT) by modifying the Alpha Army test (Gray, 2015; Hannon, 2016).

After World War I, men and women began attending college in record numbers. Northern universities expanded their student bodies from primarily white males to also include women and African Americans, traditionalists and conformists, radicals and intellectuals, and those with no

religious affiliation to various degrees (Thelin, 2004). Latin and Greek were still language entrance requirements at top academic institutions but more and more institutions offered some latitude concerning foreign language (Giles, 1926). Gouveia (2004) noted that college students typically still took classes centered on English, mathematics, history, and science during this time. The second decade of the new century would end with most universities issuing grades and professors pursuing their own unique approaches to grading standards, meanings, and distributions. Some instructors felt that at least grades motivated students as Arthur Campbell (1921) wrote in *The School Review*:

> That our marking systems of today are fraught with innumerable weaknesses and inconsistencies, their most adherents cannot deny; on the other hand, that they do serve as a spur to the laggard, even their most outspoken opponents must admit. (p. 510)

Research studies conducted after World War I and into the 1920s continued focusing on the applications of intelligence tests and normal distribution curves as a way to address both student admittance into college and college grading. Numerous researchers (J. Abell, 1928; Ellis, 1926; Gaw, 1926) concluded that grades should not vary significantly from a normal distribution curve. In addition, many educators felt that student grades should be a reflection of intelligence tests because studies showed a perfect positive correlation between ability and achievement under ideal situations (Ohlson, 1927).

Another closely related theme during this time was that professors should be ranking students against other students. University of Oregon professor Giles Ruch (1929) felt that the grading of students was arbitrary with grades having no consensus meaning. Ruch's solution involved the use of standard tests of general intelligence to rank students with respect to others.

Ruch acknowledged these tests may not be perfect but at least the tests would ensure the validity of grades when properly designed. Ruch summed up his views on grading with the following statements:

1. Students must be placed in correct, relative positions or ranks with respect to each other;
2. The adopted marking scheme must be valid, reliable and well-defined. (p. 376)

J. Abell (1928) suggested grades on exams be reduced to a 100 percent scale and then be distributed based on the median score achieved by the students. J. Abell felt his grading system would appropriately rank each student when measured by the performance of the middle or average group of students. J. Abell's proposed grading system is listed below:

A: For students scoring +12 above the median score

B: For students scoring between +4 and +11 above the median score

C: For students scoring between +3 and -4 around the median score

P: For students scoring between -5 and -12 below the median score

F: For students scoring -13 below the median score

The focus on standardization of grades during this time coincided with a period of immense standardization of product types across the nation that ranged from agricultural products like wheat to objective, psychological measurements (Freyd, 1923; Schneider & Hutt, 2014). As such, it is not without merit that professors' beliefs in standardized testing and grading stemmed from the general belief that whatever exists at all, exists in some amount, and anything that exists in some finite amount can be measured (Wahlquist, 1940). While the general assumption of this argument was that measurements in student learning are no different from measurements in the physical sciences, this was not a universally shared belief.

While research continued on the use of intelligence tests and normal distribution curves to fix the perceived ills of grading, the majority of educators rejected the idea of issuing grades according to the parameters of a fixed distribution scale (Schneider & Hutt, 2014) with several influential leaders having serious misgivings about grades in general. Odell (1928) questioned the subjectivity of grades by listing fifteen attributes that included student attitude and degree of interest in a subject coming into play when professors determined grades. Based on the perceived inaccuracy and psychological harm grades imposed on students, Korey (1926) rejected the use of grades outright. Northwestern University professor Henry Crew (1930) believed that grades were a hindrance to learning as they diverted both students' and professors' attention. University of Chicago president Max Mason delivered a speech in 1928 to members of the Institute of Higher Learning and said, "The present system of college marking is a hindrance to genuine learning. It is scarcely more than a form of academic booking" (Hillbrand, 1931, p. 726).

In an effort to investigate the distribution and reliability of marks issued by twelve different departments, Crawford (1930) investigated six years of freshmen grade data at Yale University. Crawford observed that, depending on the course selection, a freshman could have a one in three chance of failing two courses and consequently be expelled from the university. Crawford noted that a freshman selecting a different set of courses could reduce his chances of failing two courses to less than one in ten. Due to these grading differences, Crawford recognized that individual students could be unjustly harmed by the current grading process. Crawford proposed a simple system of transmutation to equate marks from different departments in an effort to reduce the grade variability among departments.

Despite the economic depression that enveloped the country during the end of the 1920s and continued into the 1930s, undergraduate enrollment remained steady at approximately 1.1

million during this period (Hostetler, 1989). As noted earlier, critics of the current grading practices during the early part of the 20th century had a tendency to favor various forms of a grading curve while the vast majority of educators rejected this notion. Research examining the relationship between college grades and later success in business generally concluded there was a positive relationship. Unfortunately, the decade would end without a national consensus on what grades should represent, their purpose, and what a proper grade distribution for a college class should be.

Studies Between 1930 and 1960

The next decade began with a deepening economic decline throughout the country. While the Depression took center stage on most college campuses, colleges continued expanding their enrollments from 1930 to the start of World War II (Thelin, 2004). With colleges and the nation primarily focused on the Depression, the awarding of letter grades to students continued without many changes across college landscapes. While most colleges did not change their grading policies during this time, one interesting change did occur on the campus of the University of Chicago. Sometime during the Depression, University of Chicago president Robert Hutchins instituted the separation of teaching and evaluating of students (Milton et al., 1986). This change would not become permanent however, as it was abandoned two decades later (Mayer, 1993).

Research on grading processes in the early 1930s still continued to focus on IQ tests and normal distribution of grades as ways to properly assess students (Eells, 1930; Middleton, 1933; Reeder, 1934; Segel, 1934). However, not all professors and administrators were influenced by the mental testing movement or convinced about the superiority of normal distributions when it came to grade distributions. A body of research developed during the 1930s and 1940s indicated that “teachers, on the basis of their judgment and such tests as they would of themselves

construct, cannot reliably mark pupils on the basis of percentages" (Segel, 1936, p. 34). A study conducted by Travers and Gronlund (1950) revealed the subjective nature of the grading process by professors. Based on the results of a questionnaire sent to 50 faculty members at the University of Michigan, Travers and Gronlund concluded that professors included numerous features like attendance, attitude, and student interest into grade determinations. The researchers opined that grades have a limited value because of their unpredictable determination. Travers and Gronlund proposed that grading systems needed to become more precise with grade meanings needing to be carefully crafted and understood in order to remedy the situation.

By the 1940s, the A–F grading system had emerged as the dominant grading system as college admissions and employers viewed grades as useful determinants in their admittance and hiring decisions (Schneider & Hutt, 2013). While the A–F grading system now dominated higher education, there was still no consensus to determine the makeup and assignment of grades with the grading of college students still varying tremendously by professor, department, and university.

In 1944, President Franklin D. Roosevelt signed into law the Servicemen's Readjustment Act, commonly known as the GI Bill, which continued the transformation of college from a privilege of the rich to the middle class (Thelin, 2004). With the United States government providing veterans returning from World War II with cash payments to offset tuition and living expenses in order to further their education, over two million servicemen would take advantage of the program over the decade that followed World War II (Greenberg, 2004). Most colleges and universities eagerly responded to the influx of students by adding more buildings, courses, professors, and degrees (Thelin, 2004).

Research after World War II continued focusing on both the subjective nature of grading and the potential harm to students caused by grading. Good (1945) forcefully opined that most educators believed marks should be abolished because of the harm they imposed on students. Good worried that despite educators doing their best to conscientiously grade students, grades were not reliable and they hindered students' ability to learn because of their discouraging signals. In addition Good stated: "everybody knows that it's well-nigh impossible to tell the difference between a high C and a low B…some teachers give nearly all A's and others hardly think anyone deserves an A" (p. 12).

Edwin Broome (1945), in writing to the influential *School and Society*, spoke to the negatives associated with grading:

> They [grades] have stimulated unwholesome rivalry, have encouraged dishonesty, have created bitter feeling among pupils, have transferred the interest of the pupils from the subject to the symbol, and, most unfortunate of all, marks have resulted in undue elation in many cases, and in devastating depression in other cases, even resulting in suicide. (p. 76)

Even movies touched on the subject of grading during this time period. In the 1945 classic movie, *The Bells of St. Mary*, Bing Crosby played a priest named Father O'Malley who was charged with evaluating a parochial school (McCarey, 1954). During a scene in the movie, Father O'Malley gets into an argument with Sister Mary Benedict (played by Ingrid Bergman). The argument centered over grading standards and what grade to assign a young student named Patsy. Sister Mary asks Father O'Malley during the argument, "Do you believe in just passing everybody, Father?" Father O'Malley replies, "Maybe. Be better than breaking their hearts."

Marcuse (1950), upon the advent of the "IBM Scoring Machine," investigated whether students preferred taking an objective test graded by this new invention or a subjective test graded by a professor. After evaluating student responses to a questionnaire, Marcuse concluded that subjective tests were the "most personally satisfying type of work" (p. 137). It should be noted that both Pease and Beardsley (1950) and Albright **(**1950) vehemently disagreed with Marcuse's conclusion that subjective grading was superior to objective grading.

Hadley **(**1954) wondered whether the grades were "fact or fancy" given the disconnect between student achievement as measured by standardized tests and as reflected in teacher grades. Hadley found evidence that girls were consistently awarded higher grades than warranted based on their achievements on standardized tests as compared to boys. Hadley believed this was due to girls being awarded higher grades because of a "halo" effect since girls tended to conform better in a classroom setting.

Studies in the early 1930s focused on tying student grades to various types of normal distribution curves as a solution to the subjective grading. Over the next two decades, research would also focus on the capriciousness of grading while the A–F grading scheme would become firmly cemented throughout higher education. The 1950s would see World War II veterans flooding higher education institutions as a result of the GI Bill with universities increasing their size and scope to handle the influx. The 1950s would end without a consensus in higher education about how to grade and what to grade.

Studies Between 1960 to 2017

Studies centering on college grading in the 1960s began to branch out and touch upon subjects like grade inflation, grade policies, assessment practices, and how students and professors view and react to grades.

Grade Inflation. Rojstaczer and Healy (2012) noted that in 1960 the average grade point average for all public and private institutions was about 2.4. Yet, by 2006, the authors noted the average grade point average had risen to approximately 3.0. Some scholars have suggested that the well-studied subject of grade inflation had its origins in the 1960s and was caused by implementation of a compulsory draft for the Vietnam War (Hunt, 2008; Rojstaczer & Healy, 2012; Rosovsky & Hartley, 2002). Professors across a number of campuses graded leniently during the Vietnam War with the aim of keeping students enrolled to avoid exposing them to the military draft (Rosovsky & Hartley, 2002). Miller (2014) noted that while grade inflation seemed to level off at the end of the Vietnam War, the trend began anew in the 1980s.

One of the first researchers to investigate the increase in college students' grade point averages was Dr. Arvo Juola. Juola (1976) used data compiled from a national survey of 134 colleges to examine student grade point averages and found evidence that over the time period from 1965 to 1973, student grade point averages increased .404 points. Juola's evidence supported the notion that a rise in student grade point averages across a wide variety of colleges existed. As a result of Juola's work, the term "grade inflation" came into existence.

Grade Policies. Beginning in the 1960s and continuing into the 1970s, schools began to experiment once again with altering their grading systems. Some universities wanted to recognize the growth of the total student and attempted to devise and apply new means of evaluating and reporting student progress (NEA, 1974). Milton et al. (1986) described noted the moving from oral assessment to written exams and the introduction of pluses and minuses to the five-letter grading system. The well-regarded California Institute of Technology (Caltech) implemented a pass/fail grading scheme for all freshmen during the 1960s (B. Abell, 1967). A few universities followed Caltech's lead while others converted to a three-point system (honors,

pass, fail); some schools implemented these changes across their entire campus while others allowed students a grading choice for certain courses (Kirschenbaum, Napier, & Simon, 1971). For a few years, Yale University abandoned its numerical scale in favor of a four-point scale (honors, high pass, pass and fail) where no cumulative GPA was calculated (Rodman, 2013). In the 1970s, Brown University adopted a grading policy of offering a pass/fail grading system where students could elect to be assessed in lieu of traditional grading for virtually every class (Pippin, 2014). University of California at Berkley adopted student-friendly policies concerning when a student could elect to take a class pass/fail and when a student could drop a class (Milton et al., 1986). Grade forgiveness policies came into existence during this time on numerous campuses, allowing students a second chance at succeeding in college **(**Rosovsky & Hartley, 2002**)**. Percentage grades gained in popularity among educators in the 1990s as grading software and online grade books became available (Guskey, 2013).

With the number of grading policy changes occurring on college campuses, researchers weighed in on how colleges and universities treat grades and GPAs in widely different manners. Chansky (1964), after reviewing literature with regard to GPAs, concluded that GPAs have several shortcomings with no inherent, stable meaning among universities. Hills (1972) opined that college grading practices are chaotic at times and potentially inhumane towards students. Hu (2005) noted that Princeton University experimented with a strict policy on the number of As and Bs awarded in its undergraduate classes while Harvard University, by contrast, had no such policy and the reported median grade given to undergraduates was an A-. Universities also varied in their policies regarding overall GPAs as they relate to undergraduate coursework repeated by students and coursework completed at other institutions (Marx & Meeler, 2013). A primary conclusion of Shouping Hu's book, *Beyond Grade Inflation*, is that grading disparity "is

a serious threat to the integrity of college grading. It not only affects students' choice of courses but also provides incentives for faculty to lower their grading standards" (Hu, 2005, pp. 42–43).

Student Behavior in Response to Grades. A growing body of research developed over the last few decades examined student behavior regarding course selection and major choice because of grades. In an effort to categorize students based on their views of grading, Oberlin College professor John Hobbs (1974) anecdotally categorized students into one of five categories.

1. The Grubber is willing to sacrifice his social life and mental stability in an effort to secure high grades.
2. The Pragmatist has a clear vocational goal, like medicine or law, in mind. While he is willing to work hard to attain the necessary grades to achieve his goals, the Pragmatist doesn't consider grades to be the life-and-death proposition that the Grubber does.
3. The Loser feels intellectually inferior and frequently achieves low grades. The Loser feels alienated from academic life and finds ways to protect his/her self-esteem from the grading system.
4. The Idealist pretends to everyone, including himself, that grades do not matter compared to learning. The Idealist is likely a reformed Grubber.
5. The Conscientious Objector protests against grading standards but continues working hard to achieve academic success. (p. 239)

It should come as no surprise that students often perceive their grades as the major cause of success or failure in a college setting (Covington, 2000). Most students will work to maximize the grade they receive in a class, believing that the more effort they put forth in a class, the better their grade will be (Milton et al., 1986). Covington (2000) and Kohn (2004) noted that for some

students, the motivation to protect self-worth becomes even more important than the need to perform well and that they may engage in self-worth protective strategies that may cause the very failures they are trying to avoid.

Research by Chizmar (2000) and Rask and Tiefenthaler (2008) noted that students in general are less likely to take additional courses in economics if they received a low grade in their introductory economics course. Sabot and Wakeman-Linn (1991) found that the effect of grades received in the first course in a subject area influenced the probability of a student enrolling in the second course. Bar, Kadiyali, and Zussman (2009) reviewed the effects of a grade policy change on student behavior at Cornell University and found that the new grade policy changed student behavior in course selection as students increasingly selected leniently graded courses in the hopes of achieving higher grades at the expense of truly challenging themselves. Some studies have suggested that because of students' obsession with grades, students are apt to think less deeply and actively seek out easier graded classes (Kohn, 2004).

Milton et al. (1986) conducted an empirical study in an attempt to better assess student attitudes and beliefs on grading. Based on the replies to a questionnaire sent to students, the authors divided college students into two groups depending on whether students were more grade-oriented or learning-oriented. Results of the study showed that grade-oriented students had generally positive views on grades, felt that studying was a "necessary evil" completed only to achieve a good grade, and would withdraw from an interesting class if they thought they would receive a poor grade. Students with a more learning-centered approach enjoyed learning new material, and had good study habits, positive educational views, and the lowest levels of test anxiety. The questionnaire also revealed that students with a high grade-oriented predisposition

to academics tended to have limited beliefs about their ability to control their own success or failure in the classroom.

Recent studies have shown today's students have a sense of academic entitlement whereby students hold the expectation of receiving high grades without taking personal responsibility to earn those grades (Ciani, Summers, & Easter, 2008; Vallade, Martin, & Weber, 2014). Today, going to college is viewed as an inalienable right by a majority of students (Kopp, Zinn, Finney, & Jurich, 2011). In addition, because either the student or their parents pay tuition, some of these students have the expectation that knowledge and high grades will be delivered to them while requiring a minimum level of effort or discomfort on behalf of the student, that any problems in learning are due to inadequacies of the professor, that students should have control over classroom policies, and that students deserve high grades (Chowning & Campbell, 2009). This academic entitlement behavior in students has been found to be associated with poor study habits, less effort, and low levels of mastery over goal orientation (Kopp et al., 2011).

These studies only reinforce that idea that inconsistent grading in multisection courses can have a profound impact upon student behavior as some students are apt to change their major or future course selections because of inconsistent grading found in multisection courses. Students taking these classes are typically adjusting from high school to college, which only adds to their anxieties (Chowning & Campbell, 2009). In some cases, these students recovered academically while others did not. In multisection classes where grading standards and objectives are not standardized, the signals students receive about their academic abilities are distorted and may not accurately represent their abilities in a given subject.

Professor Views on Grading. Examining faculties' roles and their connections to grade disparities between sections of the same class can provide an understanding of core concepts and

meaning about assigning grades in higher education. Do instructors view grades in a way that signifies such traits as accomplishment, performance, or ability? Do professors view themselves primarily as gatekeepers or facilitators concerning the grading of students? Is the professor tenured, nontenured, or recently hired?

According to Barker (2003), role theory is a "group of concepts, based on sociocultural and anthropological investigations, that pertain to the way people are influenced in their behaviors by the variety of social positions they hold and the expectations that accompany those positions" (p. 375). Role theory in this context accounts for the interactions between professors and students in higher education by focusing on the various roles professors play (Katz & Kahn, 1978). Using the lens of role theory, professors determine their stance on grading based on their status within the university. It seems logical to assume a new professor is bound to have different expectations and roles than a tenured professor. The same can be said for differential roles between tenured and nontenured professors. Grading differences may exist between professors who have a sense of belonging to the university versus those professors who feel less connected (Schutz et al., 2013).

Research supports the idea that adjunct faculty differ from full-time faculty in their perceptions of rigor in assigning grades (Schutz et al., 2013; Sonner, 2000). Van Ness, Van Ness, and Kamery (1999) concluded that part-time instructors typically assign higher grades than full-time instructors. Grading by university graduate assistants, who tend to have little formal training or supervision, only adds to the diversity of views concerning grades (Milton et al., 1986).

How professors views themselves in higher education can have a large impact on how they assess students (Schutz et al., 2013). Milton et al. (1986) advanced "adaptive-level theory" to explain grading differences among faculty. The theory posits that different grading standards

among different departments reflect both the students and professors in each department (Milton et al., 1986). As a result, the theory suggests both students and professors tend to gravitate to academic disciplines that align with their grading beliefs. Traditionally, the hard sciences have not been generous in awarding As and Bs (Hu, 2005). As such, it is not unrealistic to think that instructors in these disciplines view themselves as gatekeepers with the goal of limiting the number of students receiving passing marks. Seymour (2001) disagreed with this style claiming, "using teaching and assessment methods as a means to discover 'the few' by weeding out the rest is not only dysfunctional to this end, it is irrelevant" (p. 84).

Barnes, Bull, Campbell, and Perry (2001), in an effort to offer an alternative to "adaptive-level theory," hold that individual professors tend to favor either a criterion-referenced grading system or a norm-referenced grading system. While not as widely accepted as in the past, grading on a curve (norm-referenced grading) continues to be touted by some influential professors. Professors in this camp hold the belief that the purpose of grading is a mechanism to sort, rank, or act as a gatekeeper (Hunt, 2008). Professor Richard Kamber praised grading on the curve as a reliable and valid measure of student achievement along with Professor Harvey Mansfield, who stated, "*students should be set against one another in a race for artificially scarce rewards*" (Hunt, 2008, p. 6, italics in original). Kohn (2004) offered an opposing view that favors a criterion-referenced grading system. Kohn believes the number of peers that a student has bested tells us little about how much a student knows and is able to perform. Kohn criticized the notion that the bell curve accurately describes the range of knowledge in any given class as this may or may not be the case.

A study by Ekstrom and Villegas (1994) provided validation to both theories. Ekstrom and Villegas surveyed 500 faculty members at 14 colleges and universities about their attitudes

towards grading. Their results provided evidence that faculty orientation towards grading was primarily either criterion-referenced or norm-referenced. In addition, the authors noted significant differences across academic departments with professors in the STEM departments typically favoring norm-referenced grading practices while non-STEM departments favored criterion-referenced grading practices.

In addition to the above-mentioned theories surrounding professor dispositions towards criterion-referenced and norm-referenced grading, Geisinger, Wilson, and Naumann (1980) added yet another with the idea that some professors grade students based on the students' individual abilities and motivation. Professors in this camp would vary their grading based student anxieties, student ability, academic integrity, and other personal philosophies about the course (Schutz et al., 2013).

In addition to positive and negative views on grading, professors have differing views on how and what to grade. While some professors use meritocratic rules to distribute grades, other professors have adopted particularistic rules to assess students based on a student's individual circumstances (Milton et al., 1986). Despite the view shared by most faculty that grades should not be based on nonacademic issues, Guskey and Bailey (2001) report that "nearly every instructor wants to add individual exceptions" (p. 141) to standard grading procedures. Professors often include such behaviors as effort, attendance, and participation in their course grading schemes as a classroom control device (Marzano, 2000).

It would seem most professors today use grades as an evaluation and communication tool to assess the progress a student has made in learning the material over the course of the semester. A potential issue in multisection classes centers on the differences in academic standards that professors are applying, the employment status of professors, and differing views on how and

what to grade. These differing standards are bound to affect students in an uneven fashion. It could be quite possible that some students are forced to abandon their pursuit of a medical career and reevaluate their career options after taking a chemistry or physics class taught by a strict grader, whereas they may have stayed the course with a different instructor.

Assessment Practices

Another well-researched topic related to grading is the proper way for professors to assess students. Wiggins, in *Educative Assessment*, believes "assessment is primarily to educate and improve student performance, not merely audit it" (1998, p. 7). Walvoord and Anderson (2010) documented the types of communication grades conveyed to students and the effects of those grades on students. The authors believe student assessment can be effective when using a well-designed rubric. A positive aspect of grading noted by Walvoord and Anderson was that grades serve to organize and prepare students, and, by using grades as a proxy, students can decide how to best allocate their time.

While both Wiggins (1998) and Walvoord and Anderson (2010) entertain positive views on student assessment, they point out several issues with how grading is currently conducted in higher education. Wiggins wrote that too often, current grading practices do not offer students any "performer-friendly feedback" (1998, p. 21), are simplified to aid in scoring, are disconnected and disjointed among faculty, allow for no individualism by students, and simply provide only a score to students. Walvoord and Anderson acknowledged that "you can't use grades for assessment" (2010, p. xvi). It would seem no matter the assessment style, inconsistency issues remain as assessment models vary from class to class, are based primarily on a professor's grading views, and reduce everything a student has learned to one of five letters of the alphabet.

Urda and Ramocki (2015) reported that among the four types of assessment—memorization, analysis, creativity, and practical application—students preferred tests based on memorization even though researchers noted that students typically did not perform statistically better on tests based on memorization. In addition, Su (2007) noted that students' preference of team-based learning, grading, and assessment methods were, in part, dependent on how the students viewed their own academic ability.

Summary of Grading Research

A historical review of the assessment practices reveals universities have employed several different types of assessment since the founding of Harvard University in 1936. Oral examinations given to college students in the beginning gave way to other ranking systems, which came into vogue as universities expanded and become more complex. It was not until around the turn of the 20th century that the five-letter system came into existence and became the dominant grading system found on college campuses today. Even with the five-letter system firmly established, universities continued tinkering with pass/fail, grade forgiveness, and other changes in hopes of finding a better assessment option.

Research highlighted in this paper has shown the many inconsistencies with which grades are derived. Where assessing students was once relatively simple, college grading has now become exceptionally complex due to forces inside and out of academia (Hu, 2005). Early researchers like M. Meyer in 1908 and Finkelstein in 1913 uncovered wide divergences in grades given to students on their campuses. The mental testing movement of the 1920s and 1930s sought to come up with purported objective ways in which to measure students, even though they had little success in winning over the majority of their colleagues. Research by Good (1945) and

Travers and Gronlund (1950) during the 1940s and 1950s continued to point out the inconsistency in which grades were distributed.

Research in the 1960s that continues to today brought the issue of "grade inflation" into our vocabulary. Other lines of recent research involving grading inconsistencies examined student behavior regarding course selection and major choice because of grades and how professors assess students. As such, while letter grades have become the predominant assessment system of students, their meanings, purpose, distribution, impact on students, and how they are derived remains unsettled territory. Perhaps Dressel (1976, p. 2) summed it up best when he defined current grade reporting practices as "an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has achieved an undefined level of mastery of an unknown proportion of an indefinite material."

At the end of the day, it would appear educators are stuck in a bind until a better way to assess students is found and accepted by the majority of universities. Even though many educators and researchers have doubts over the legitimacy, representation, purpose, and usefulness of grades, the vast majority of professors and universities will continue issuing them for the foreseeable future.

Why Grades Are Still With Us

Chansky (1964) reported over fifty years ago that undergraduate grades continued to represent different things to different people in higher education. Considering the criticisms that have been raised regarding grading, it is easy to wonder why grades have become the dominant means of reporting student achievement. There are a number of factors that continue to allow grades to flourish on college campuses.

Students, professors, and administrators generally appreciate something familiar to them even if they are somewhat dissatisfied with it. As Kohn (2004) put it, "students expect to receive grades and even seem addicted to them" (p. 81). In addition, people tend to prefer simple and grades represent simple concepts (NEA, 1974). Supporters of grades like to point to the fact that people are familiar with grades, even if they mean different things to different groups. When professors or universities have proposed changes to grading, they are often met with negative reactions. Professors David Beito and Charles Nuckolls found that university administration and other professors had little interest in grade reform when they led a failed campaign at the University of Alabama to halt grade inflation at their institution (Hunt, 2008).

Grades also offer an easy way for professors to assess students and for administrators to have the ability to store the assessments over time. Other alternatives would require a much greater time investment on the part of professors and administrators and are thus viewed negatively (NEA, 1974). As a result, there are very few universities and colleges like Alverno College in Wisconsin where students receive a narrative transcript about their abilities instead of traditional grades.

Despite the criticisms that some students are being negatively affected by grades, there are students who are positively motivated by grades. Studies have shown that when grades were

removed for some students and retained for others, student output for the ungraded students decreased in comparison to the graded group (Milton et al., 1986). In addition, grades are used to screen students for college admission, as well as graduate and professional programs (NEA, 1974). Without the use of grades, the screening process would invariably become more complex and costly. Companies interviewing students may not want to deal with this additional complexity, as noted earlier, that people prefer simple to complex. Any individual school would be very hesitant to move away from grades if it felt that would harm their students' employment or graduate school opportunities.

Finally, the best argument for the continuation of grades is that no other alternative has been shown to be superior (NEA, 1974). Society is very grade-oriented and as a result, so are the majority of students. The majority of students believe that grades are important and accept university grading systems so completely that they cannot conceive of alternatives (Milton et al., 1986). Society attaches an outsized premium to high grades. As such, students recognize that it pays to have excellent grades. Scholarships, top jobs, and entry into prestigious graduate programs are awarded on the basis of grades. Until society's value system changes about grades, it is likely that grades will continue to be with us.

It appears that the standardization of grading is still not imminent in higher education. While some institutions have implemented policies for assessing students based on a bell curve, other institutions allow the students the use of pass/fail grading options, grade forgiveness, and late class withdrawals. Still other institutions use letter grades but assess students on a wide range of factors that are not consistent from one campus and classroom to the next. The next chapter will cover the design of this study used to explore student and instructor characteristics

that contribute to our understanding of the variances in student grades in multisection calculus classes.

CHAPTER III: RESEARCH DESIGN

Overview

In the previous chapter, I reviewed literature on the history of grades, current literature regarding stakeholders' perceptions of grading, and why grades still dominate the college landscape. This chapter describes the methods used in conducting the research study to develop two-level statistical models (student-level and professor-level) to understand the interrelationships between student-level and professor characteristics and Calculus I grades. In addition, I use this chapter to identify the student population and unique student and instructor attributes and provide a discussion on the research design. I conclude this chapter presenting information about the steps involved in data analysis, model designs, and interpretation.

Purpose of Research

I created this dissertation to investigate correlates and predictors of multisection Calculus I course grades at a mid-sized, four-year public Midwestern university using both SPSS and HLM for the purpose of creating statistical models to better understand how the relationships between student and professor characteristics influence course grades. Findings may contribute to a closer examination of the course registration process to better match students with professors to improve student success rates in multisection Calculus I courses. Accordingly, the results of this study are of practical significance to professors and administrators interested in increasing success rates across multisection courses. Fairness to students concerning grading should be a goal of both professors and administrators to ensure that student grades are primarily a product of their achievement versus which professor is teaching a given section. Administrative officials should be concerned that if disadvantaged students are unable to secure gainful employment in their chosen field, it may become increasingly difficult for a university to increase future student

enrollment and solicit alumni donations as these students voice their concerns to current and prospective students.

Research Questions

Based on the literature review, four broad research questions were addressed in this quantitative study:

1. Are there differences in student Calculus I course grades among the college students between the course sections?
2. To what degree do student characteristics (gender, ACT, high school GPA, etc.) explain differences in course grades?
3. To what degree do instructor characteristics (age, tenure, experience, gender, etc.) explain the differences in mean course grades?
4. Do instructor characteristics influence the magnitude of the student success measures on student course grades?

Research Design

To address the listed research questions, I selected a secondary data analysis design using archived student grade data of students from a traditional, four-year public university in the Midwest. To investigate correlates and predictors between students taking Calculus I, archived student grade data and instructor characteristics for Calculus I courses were retrieved from the mainframe database of the university. Student and instructor data was organized into a univariate format with dichotomous variables being reformatted to either 0 or 1.

SPSS and Excel were used to organize the student and instructor data into two files. Using regression analysis, predictor variables associated with students, including gender, grades, ACT scores, earned college hours, ethnicity, and calculus course sections, were compared to

determine the significance and degree to which these variables explained the variances of student grades in Calculus I. Predictor variables associated with instructors, including gender, age, and experience, were also compared using regression analysis to determine the significance and degree to which these variables explained the variances of student grades in Calculus I.

The student-level (or level-one) file was sorted by Instructor IDs. The instructor-level (or level-two) file was initially sorted by Instructor ID. HLM software requires the data in the instructor-level file be aggregated to the higher level (instructor-level data). The break variable in the aggregation process is the Instructor ID variable. By having a two-level hierarchical linear model, the variance associated with each level can be separated and measured for both the level-one (student data) intercept and slope. Aggregated instructor variables include:

- Instructor tenure track status
- Instructor gender
- Instructor age
- Instructor years teaching
- Instructor course grades by section

Using HLM, two-level hierarchical linear models with student-level variables at the first level and instructor-level variables at the second level were constructed. The need for simultaneous evaluation of student-level and instructor-level variables and the hierarchical nature of the data guided the decision to use HLM (Raudenbush & Bryk, 2002). Hierarchical linear modeling techniques enable us to conceptualize in terms of multiple levels and then identify, measure, and model any between-group differences uncovered in ways that linear regression analysis fails to account for (Raudenbush & Bryk, 2002).

The Population

The target population for this study was student grade data from a four-year public Midwestern university. The data set included archived student information from the 2009 through 2013 calendar years for students who enrolled in a section of Calculus I. Students' grades, ACT scores, ethnicity, hours earned and gender were among the attributes selected to analyze. At the university studied, there were 73 sections of Calculus I during that time frame.

There was an average of 29 students in each of the 73 Calculus I sections that were observed in this study, meaning 2,086 students received a grade in a Calculus I class from 2009–2013. A majority of the students that enrolled in a Calculus I course were identified as either freshman or sophomores. Slightly over 61% of the students were identified as male, and 85% of the students identified themselves as white. Two of the sections included in the results of this study were honors Calculus I, but it should be noted that students in these sections had statistically higher high school GPAs and ACT scores.

In order to help explain course grade differences among the Calculus I sections, data on instructors that taught a Calculus I section from 2009–2013 was collected as well. Seventeen different instructors were identified in this group, and all of them earned a PhD in a STEM-related field. The majority of the instructors were men, while fewer women taught more sections of the Calculus I classes (40 of the 73).

Variables in the Study

The student characteristic variables used in this study were selected by analyzing prior research (Credé, Roch, & Kieszczynka , 2010; Sabot & Wakeman-Linn, 1991) that found gender, grades, college hours earned, and ACT scores were predictors of student GPAs in college. Astin (1991) believed that, among the current admissions data available, students' high

school GPAs and standardized test scores were the two strongest predictors of their college successes. According to Mertler and Vannatta (2005):

> In nonexperimental research…the researcher has no control over the levels of the [independent variables]. The research can define the [independent variables], but cannot assign participants to its various levels. The participants enter the study already "belonging" to one of the levels. (p. na)

For this study, the students placed themselves into the various classes. Table 1 contains a list of independent student variables, descriptions, and the scale of measurement initially used in conducting this research in order to answer the four research questions.

Table 1
Independent Student Variables Explored in Study

Name	Description	Values
ACT Composite Score	Composite ACT score	Range from 1–36
ACT Math Subscore	Math ACT score	Range from 1–36
Age of student	Age of student at time of study	
College GPA	Composite college GPA	Range from 0–4.00
College Major GPA	Composite college major GPA	Range from 0–4.00
College Hours Comp	Completed college hours	Range from 0–150+
Ethnicity of student	White, Latin American, African American, Asian	
Gender	Male or female	1 = Male, 0 = Female
Grade in Calculus I	Grade earned in Calculus I	A = 4, B = 3, C = 2, D = 1, F = 0
High School GPA	Composite high school GPA	Range from 0–4.00

Predictor variables initially examined in the study but not used in determining the results included "ACT Composite Score," "Age of student," "Ethnicity of student," and "College Major GPA." It was determined that each students' ACT Math Subscore would be a better a predictor in determining their Calculus I grades as compared to their overall ACT score. As such, the ACT Composite Score variable was not analyzed further in this study. The students in the study were very homogeneous in terms of both age and ethnicity, so these student attributes were not investigated further. It was determined that a student's major grade point average achieved in

college would be sufficiently reflected in their overall grade point average. As such, this student attribute was also not investigated further.

Figure 1 presents a pictorial view of student characteristics analyzed in this study.

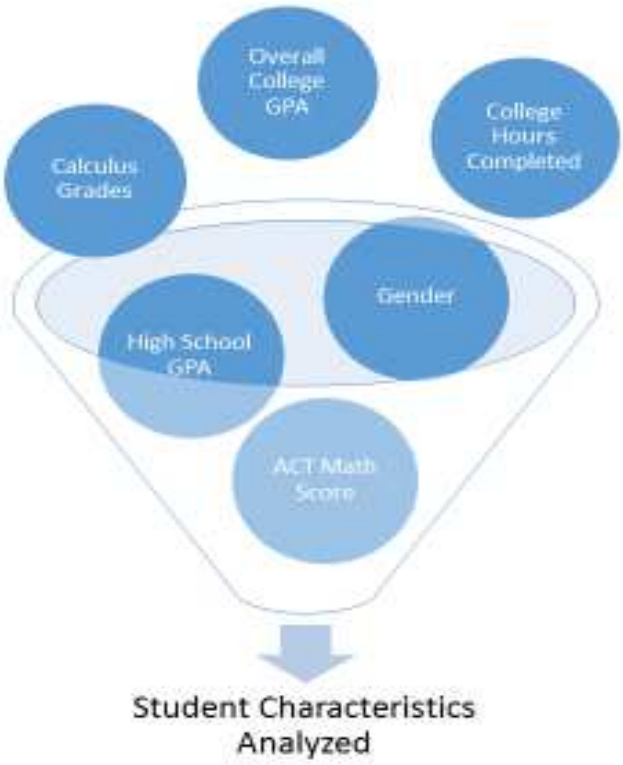

Figure 1. Pictorial of college student characteristics analyzed.

With the goal of building the best predictive models for student grade outcomes in Calculus I courses, archived data on instructors that taught a Calculus I course from 2009–2013 was collected as well. Table 2 contains an initial list of independent instructor variables, descriptions, and the scale of measurement initially explored in conducting this research in order to answer the three previously highlighted research questions.

Table 2
Independent Instructor Variables Explored in Study

Name	Description	Values
Tenured track or nontenured track	Tenured track professor or nontenured track	1 = Tenure, 0 = Nontenure Track
Earned PhD	Earned PhD or not	1 = Earned PhD, 0 = No PhD
Years of experience	Years of experience	
Age of instructor	Age of professor	
Ethnicity	White, Latin American, African American, Asian	
Gender of instructor	Male or female	1 = Male, 0 = Female
Grade in Calculus I	Average course grade assigned	Range from 0–4.00

Predictor variables related to instructors that were initially examined in the study but not used in determining the results were "Earned PhD" and "Ethnicity." As noted earlier, all instructors that taught Calculus I had earned a PhD in a STEM-related field. As such, this instructor attribute were not investigated further. In order to conceal the identity of an instructor who taught any of the Calculus I courses, ethnicity was not considered.

Figure 2 presents a pictorial view of instructor characteristics analyzed in this study.

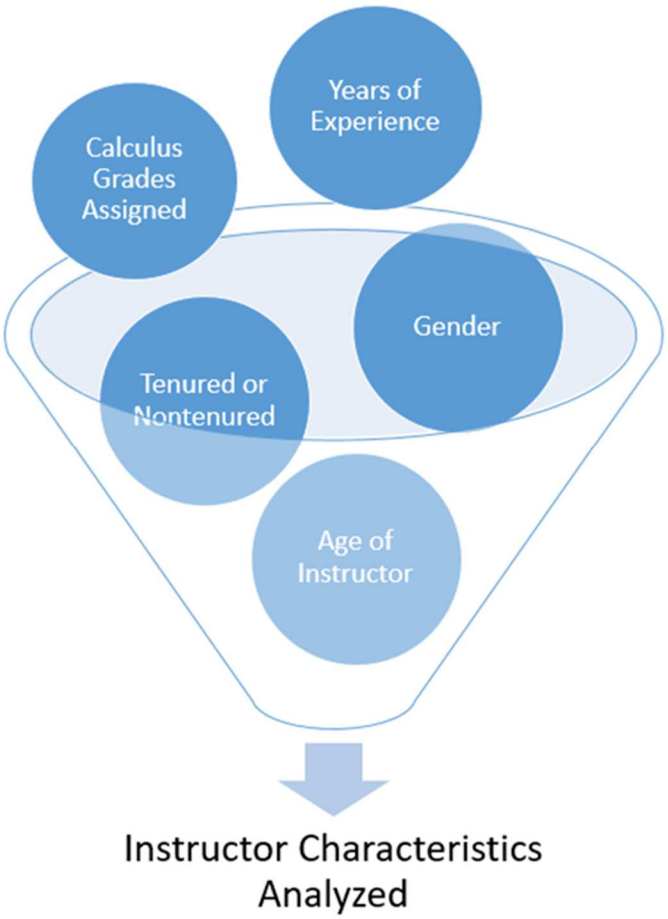

Figure 2. Pictorial of instructor characteristics analyzed.

Data Analysis

For purposes of this study, correlation analysis, regression analysis, and HLM were used to create statistical models that addressed the above research questions. MacDonell and Wylie (2014) noted that predictive modeling (regression analysis) is a useful process used to create statistical models that can then be used to answer critical questions. Regression analysis was used to estimate the predictive value of the relationship among the various student characteristics. Multiple regression is used to create a predictive model for the dependent variable of student grades in Calculus I. The variables of student gender, instructor, ACT Math

score, and high school GPA are used as independent variables. Both Excel and SPSS were the primary software packages used in this study to calculate means and compute regression analyses with F-ratio, p-value, Beta, R square, and Adjusted R square values reported.

The data was also analyzed using HLM software with the aim of building two-level models that explain grade variances between students in Calculus I using both student and instructor characteristics. While students within a particular classroom tended to have homogeneous characteristics in terms of ACT scores and high school GPAs, the environment in which they took Calculus I differed because of the instructor. Given that both students and instructors have individual characteristics, HLM can be a useful tool to deal with the problem of how to deal with this cross-level data problem.

Missing Values

The archived data collected came from the mainframe database of a four-year public Midwest university. Any cases that contained missing values were either discarded or averages for the specific value were substituted.

Summary

By creating best-fit statistical models between students and professors based on the studied characteristics of both professors and students across multisection Calculus I courses at the university studied, this study hopes to offer predictive models showing how students can best succeed in multisection Calculus I classes. The findings may lead professors and students, department chairs, and university administrators to be more open to changing the course registration process to incorporate some of the findings from this dissertation. The next chapter will review the findings that resulted from this study.

CHAPTER IV: ANALYSIS OF RESULTS

Overview

In this chapter, I seek to discover if different grading variances exist between the calculus course grades of students in multisection Calculus I classes and to what degree instructor and student characteristics explain the calculus course grade variances. Seventy-three Calculus I sections were observed in this study. There was an average of 29 students in each class with all students typically required to have passed a precalculus class unless they achieved a score of 27 or better on the math section of the ACT exam before admittance to the class. The 73 Calculus I sections were initially ordered by course grade averages and then graphed to visually show the distribution of course grade averages in Figure 3.

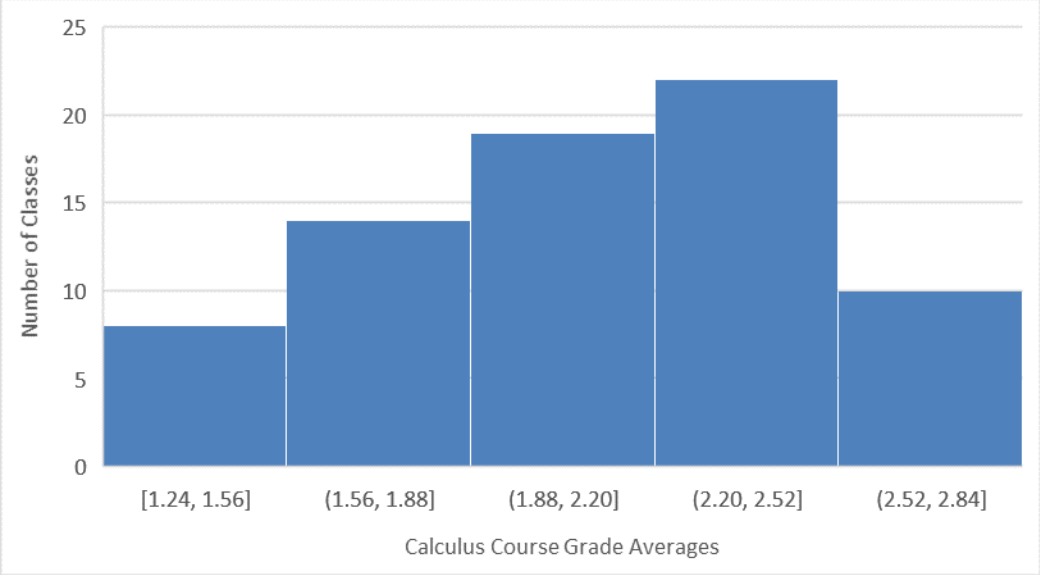

Figure 3. Graph of average course grades assigned in Calculus I classes between 2009–2013.

By looking at Figure 3, it is readily apparent that the 73 Calculus I course grade averages varied a great deal between the years 2009 and 2013. The lowest graded section average was 1.24 on a 0.00 to 4.00 scale. Eight calculus sections had an average course grade of 1.56 or less while ten calculus sections had an average course grade of 2.52 or greater. The highest graded

section average was 2.80 with 2.10 being the average section grade of the observed 73 Calculus I sections.

The 73 Calculus I sections were next ordered by course grade averages and then artificially set into groups of six. The artificial grouping was done to better understand and discover the reasons for the divergent course grade averages among the various sections. Table 3 presents the average grading distribution results for the Calculus I classes set into groups of six, ordered by course averages in each section along with some key metrics related to students enrolled in the course.

Table 3
Summary of Grade Distributions, Number of Students, and Success Rates in Calculus I Courses Ordered by GPA in Groups of Six Sections

Class section	Section Grade Average	Average # of Students[a]	Student Success Rate[b]	Student Retake Rate[c]	Composite College GPA[d]	Male / Female Ratio[e]	High School GPA[f]	ACT Math Subscore[g]
Sections 01–06	2.71	30.00	82.77%	8.90%	2.81	0.59	3.44	26.76
Sections 07–12	2.55	28.00	76.45%	8.30%	3.08	0.60	3.59	27.12
Sections 13–18	2.43	29.50	72.16%	13.64%	2.90	0.56	3.46	25.85
Sections 19–24	2.36	26.83	70.37%	11.11%	2.93	0.62	3.52	27.01
Sections 25–30	2.23	29.83	74.86%	12.85%	2.78	0.71	3.35	26.72
Sections 31–36	2.20	27.50	66.06%	16.97%	2.83	0.62	3.47	26.57
Sections 38–43	2.13	29.17	69.34%	11.93%	2.81	0.62	3.46	26.23
Sections 44–49	2.02	29.50	61.76%	15.61%	2.88	0.56	3.42	26.42
Sections 50–55	1.88	29.83	59.78%	13.79%	2.86	0.65	3.41	26.18
Sections 56–61	1.78	29.67	54.49%	19.66%	2.77	0.62	3.42	26.51
Sections 62–67	1.59	28.67	45.35%	25.82%	2.76	0.64	3.40	26.16
Sections 68–73	1.34	24.33	42.47%	30.82%	2.65	0.65	3.48	27.08
Average Section[h]	2.10	28.58	65.00%	16.22%	2.84	0.62	3.44	26.58

Note: With number 73 being a prime number, the middle course (course 37) was not included.

[a]Average number of students in each Calculus I section. [b]Student Success Rate reports the percentage of students earning a letter grade of either A, B, or C. [c]Student Retake Rate reports the percentage of students who chose to repeat Calculus I. [d]Average GPA of students who were in Calculus I. The GPA average includes only coursework completed at the university studied and includes grades from coursework before and after this course. [e]Ratio of male to female students in each section. [f]Average High School GPA of students who were in each section. [g]ACT Math Subscore average of students who were in each section. [h]Averages of all 73 sections of Calculus I.

Examination of Table 3 shows 42.47% of the students in the six sections with the lowest average section grades (Sections 68–73) passed Calculus I. In the six sections with the highest average section grades (Sections 01–06), 82.77% of the students passed Calculus I. The results also show a wide distribution in section grade averages between the top six and bottom six sections of Calculus I (2.71 and 1.34 respectively) while appearing to have students with similar High School GPAs (3.44 and 3.48 respectively) and ACT math scores (26.76 and 27.08 respectively). Table 3 shows Student Success Rates decreasing and Student Retake Rates increasing in a fairly uniform fashion as the Section Grade Averages decrease.

Data from all 73 sections was next reordered by instructors who taught the courses in order to examine the extent an individual instructor may have played in assigning calculus course grades. Table 4 presents the average grading distribution results for the 73 sections of the Calculus I classes grouped by instructor.

Table 4
Summary of Grade Distributions, Number of Students, and Success Rates in Calculus I Courses Grouped by Instructor

Instructor	Section Grade Average	Average # of Students[a]	Student Success Rate[b]	Student Retake Rate[c]	Composite College GPA[d]	Male / Female Ratio[e]	High School GPA[f]	ACT Math Subscore[g]
Instructor 01	1.34	24.33	42.44%	31.62%	2.96	0.65	3.48	27.08
Instructor 02	1.87	29.00	56.90%	24.14%	3.11	0.78	3.40	26.92
Instructor 03	1.91	30.00	58.32%	16.51%	3.02	0.59	3.51	26.75
Instructor 04	1.99	29.33	53.12%	20.63%	2.98	0.58	3.47	26.09
Instructor 05	2.00	5.00	40.00%	40.00%	3.13	0.60	3.26	25.00
Instructor 06	2.10	28.20	67.66%	13.01%	2.97	0.64	3.41	26.22
Instructor 07	2.15	30.00	68.89%	11.11%	3.12	0.64	3.36	26.46
Instructor 08	2.20	28.25	76.06%	11.49%	3.07	0.66	3.43	26.83
Instructor 09	2.21	27.00	66.67%	22.22%	2.90	0.56	3.36	26.41
Instructor 10	2.21	30.00	68.07%	16.15%	2.98	0.64	3.38	26.42
Instructor 11	2.28	29.00	72.93%	19.93%	3.01	0.58	3.49	27.73
Instructor 12	2.39	29.75	69.64%	10.87%	3.07	0.64	3.54	26.72
Instructor 13	2.42	30.33	80.25%	5.45%	3.04	0.56	3.36	25.95
Instructor 14	2.42	30.50	68.70%	7.98%	3.19	0.53	2.97	25.90
Instructor 15	2.46	20.00	71.33%	26.00%	3.52	0.49	3.97	30.47
Instructor 16	2.61	31.00	77.60%	14.48%	2.99	0.60	3.40	25.85
Instructor 17	2.67	32.00	81.26%	9.38%	2.85	0.69	3.38	26.79

[a]Average number of students in each Calculus I section. [b]Student Success Rate reports the percentage of students earning a letter grade of either A, B, or C. [c]Student Retake Rate reports the percentage of students who chose to repeat Calculus I. [d]Average GPA of students who were in Calculus I. The GPA average includes only coursework completed at the university studied and includes grades from coursework before and after this course. [e]Ratio of male to female students in each section. [f]Average High School GPA of students who were in each section. [g]ACT Math Subscore average of students who were in each section.

By examining Table 4, only 42.44% of the students who had Instructor 01 passed Calculus I compared to 81.26% of the students who had Instructor 17. In this instance, it is interesting to note that students having Instructor 01, on average, had both a higher High School GPA and ACT Math Subscore than students who had Instructor 17 for Calculus I. Table 4 provides circumstantial evidence that instructors who taught Calculus I might have played a meaningful role in determining calculus course grades.

With the results of Table 4 in mind, the calculus section grade averages associated with each instructor may be thought of as a proxy of the grading philosophy of each instructor. This continuous variable along with four other instructor-centric variables will be used to later examine the role an instructor plays in explaining the variance associated with calculus course grades.

I next analyzed the data using Microsoft Excel, Hierarchical Linear Modeling (HLM), and IBM SPSS Statistics 20 analysis to identify potential reasons for the variances in calculus course grades and the extent to which these grading differences exist at the student and instructor levels in order to answer the research questions. Three kinds of statistical analyses were completed in this study: correlation analysis, hierarchical linear modeling, and multiple regression analysis.

Based on the purpose of the study and the literature review, four broad research questions were addressed in the study.

Research Question One: Are There Differences in Student Calculus I Course Grades Among the College Students Between the Course Sections?

The possibility exists for standard statistical analysis techniques, like linear regression, to not account for cases where students taught by the same instructor have more similar academic and nonacademic characteristics among each other than to students taught by different instructors. If this situation arises, the results obtained by linear regression may be biased (Anderson, 2012; McCoach, n.d.; Stevens, 2007). HLM considers the statistical relations of dependent variables and independent variables as well as relations between the students and instructors that can be represented as different data levels (Raudenbush & Bryk, 2002).

As such, data collected on students and instructors was arranged into two-level models using HLM software to show how much the student-centric variables and instructor-centric variables separately explained the variances in course grade averages among the 73 Calculus I classes. The first SPSS file (level-one data) contained student-centric variables on 2,090 students. For this part of the analysis, any missing data pertaining to an individual student resulted in that student's information not being considered for further examination using HLM software. The remaining student data on 1,726 students that had complete information was then sorted by the unique ID assigned to the instructor that taught the calculus class and then by student ID. The second file (level-two data) contained instructor-centric variables and was sorted by a unique ID assigned to each of the 17 instructors who taught Calculus I and then by student ID.

Model 1: Unconditional (Null) Model

Using HLM, I initially created a baseline (null or unconditional) model for comparison to more complex models. For the baseline model, I selected neither student-centric variables nor instructor-centric variables were selected to examine their effects on calculus course grades. This baseline model is the equivalent to a one-way analysis of variance (ANOVA) and is used within a HLM framework primarily to establish a baseline model from which subsequent models can be compared and to capture the degree to which variance at Level-1 depends upon group membership at Level-2 (Anderson, 2012).

This baseline model contains just one fixed term (the calculus course grades of students) and variance terms at both the student and instructor level. In this way, the model provides a measure of the grade variation between students and grade variation between students with a given instructor. This will allow for the calculation of an intraclass correlation coefficient (ICC),

which will give the proportion of the total calculus grade variation of students at both the student and instructor levels. The ICC reports on the amount of variation unexplained by any predictors in the model that can be attributed to the grouping variable, as compared to the overall unexplained variance (within and between variance) (Department of Statistics and Data Sciences, 2015). Based on the covariance estimates, the ICC represents the proportion of the variance in calculus course grades attributable to instructors and ranges from 0 to 1.0 (Anderson, 2012).

With later models, predictor variables are added to each level to examine what happens to the unexplained variance found in the baseline model. With that in mind, the following models were initially created:

Level-1 Baseline Model

$COURSEGR_{ij} = \beta_{0j} + r_{ij}$

Level-2 Baseline Model

$\beta_{0j} = \Upsilon_{00} + u_{0j}$

The Level-1 Baseline Model and Level-2 Baseline Model can also be displayed in its mixed model form by simply substituting the Level-2 equation into the Level-1 equation. The resulting model obtained is:

Mixed Baseline Model

$COURSEGR_{ij} = \Upsilon_{00} + u_{0j} + r_{ij}$

For the Level-1 Baseline Model, $COURSEGR_{ij}$ (Calculus Course Grade) represents the outcome variable *COURSEGR* for Level-1 unit, i, nested in Level-2 unit, j, and is equal to a Level-1 intercept, β_{0j}, and residual or unexplained variance, r_{ij}. For the Level-2 Baseline Model, the Level-1 intercept, β_{0j}, is set as the outcome in a new regression equation with two components: the Level-2 intercept, Υ_{00}, and a random parameter, u_{0j}, which is the Level-2 residual variance. The Level-2 random parameter, u_{0j}, is what allows the model to vary by the higher-level unit.

After running the model, the parameter estimates obtained are:

Intercept = $\beta_{0j} = 2.19$

Variance at student-level = $\sigma^2 = Var(r_{ij}) = 1.562$

Variance at instructor-level = $\tau_{00} = Var(u_{0j}) = 0.109$

Total Variance = $Var(\Upsilon_{ij}) = \sigma^2 + \tau 0_0 = 1.562 + 0.109 + = 1.671$

With the unconditional model results, the total variation in calculus course grades can now be divided between the variation over students and variance on an instructor level.

$$\text{The ICC} = \frac{Var(u_{0j})}{Var(u_{0j} + r_{ij})} = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)} = \frac{0.109}{(0.109 + 1.562)} = 0.0652$$

$\tau_{00} = u_{0j}$ = variance at Level-2 (instructor-level)

$\sigma^2 = r_{ij}$ = variance at Level-1 (student-level)

Analysis shows 6.52% of the grade variance is the portion of the total variance that occurs between instructors. The remaining 93.48% of the calculus course grade variance occurs at the student level.

Research Question Two: To What Degree Do Student Characteristics (Gender, ACT, High School GPA, etc.) Explain Differences in Course Grades?

Using HLM to create a baseline model, it was discovered that over 93% of variance in calculus course grades resulted from student characteristics. In order to answer Research Question Two, I next examined student-centric factors using SPSS and HLM to determine the level of grade variance these characteristics were able to explain in course grades achieved in Calculus I.

Student-Centric Variables for Grade Predictors of Calculus I Grade Averages

Correlation analysis was also used to assess the possibility of multicollinearity between the student-centric grade predictors. Correlation coefficient is a statistical measurement that measures both the strength and direction of a linear relationship between two variables (Meyers, Gamst, & Guarino, 2017).

The issue of multicollinearity centers on if two independent variables are highly correlated (redundant) to each other; a regression model using both variables to explain the behavior of a dependent variable may not give valid results in regard to any individual predictor and about which predictors are redundant with respect to others (Vogt, Vogt, Gardner, & Haeffele, 2014). Meyers, Gamst, and Guarino (2017) note that a correlation value of less than .6 should pose no problems to using the independent variables together in a regression model.

Table 5 shows the results of correlation analysis in which the five student-centric characteristics (Male/Female Ratio in each class, High School GPA, Composite College GPA, College Hours Earned, and ACT Math scores investigated).

Table 5
Correlation of Student-Centric Predictors to Student Calculus Course Grades

Correlations[c]

Student Predictors		Course Grade	Student Gender	College GPA	HS GPA	ACT Math	College Hours
Course Grade	Pearson Correlation	1					
Student Gender	Pearson Correlation	-.146**	1				
College GPA	Pearson Correlation	.612**	-.266**	1			
HS GPA	Pearson Correlation	.359**	-.334**	.455**	1		
ACT Math	Pearson Correlation	.155**	.177**	.041*	.177**	1	
College Hours	Pearson Correlation	.199**	-.115**	.480**	.152**	-.059*	1

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

c. Listwise N=1726

Examination of Table 5 illustrates that four of the student-centric predictors (High School GPA, ACT Math, College Hours Earned, and College GPA) were positively correlated with calculus course grades. Of the positively correlated predictors, College GPA and High School GPA ($r = .612$ and $r = .359$ respectively) had the strongest association with a calculus course grades. Only a student's gender ($r = -.146$) was negatively correlated with a calculus course grades. All five student-centric predictors were significantly correlated at the 0.01 level with calculus course grades. All student-centric variables had correlation values of less than .5 when compared with each other so multicollinearity was not an issue.

Regression Model Using Student-Centric Variables

Regression analysis using SPSS was primarily used to determine what student-centric variables were statistically significant in explaining the variance in calculus course grades. Regression analysis is widely used for predicting and forecasting the relationships among independent and dependent variables (Meyers et al., 2017). With this in mind, regression analysis was used to create a model showing the relationship between the dependent variable (student course grade) and a linear combination of all five, independent student-centric variables (College GPA, College Hours Earned, Student Gender, High School GPA, and ACT Math Score). Variables College GPA, College Hours Earned, High School GPA, and ACT Math Score were mean centered.

The regression prediction model is presented below.

$$COURSEGR_i = \beta_0 + \beta_1*(COLLEGEGPA_i) + \beta_2*(HOURS_i) + \beta_3*(SGEN_i) + \beta_4*(HSGPA_i) + \beta_5*(ACTMATH_i) + r_i$$

A summary of the analysis is presented in Tables 6 and 7.

Table 6
Regression Table Summary Using Student-Centric Independent Variables

Regression Statistics	
Multiple R[a]	0.627
R Square[b]	0.393
Adjusted R Square	0.392
Standard Error	1.000
Observations[c]	1726

[a]Multiple R, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two or more variables. [b]R Square is the coefficient of determination and tells the proportion of the variance of one variable(s) that is predictable from the other variable(s). [c]Number of student observations.

Table 7
Coefficient Table Where the Dependent Variable is Calculus Course Grade Using Student-Centric Independent Variables

	Unstandardized Coefficients		Standardized Coefficients		
	B[a]	Std. Error[b]	Beta[c]	t[d]	Significance[e]
(Constant)	2.033	0.042		48.743	$P < 0.001$
HS GPA	0.190	0.061	0.070	3.105	$p = 0.002$
College GPA	1.072	0.042	0.627	25.436	$p < 0.001$
ACT Math Score	0.045	0.008	0.112	5.647	$p < 0.001$
College Hours	-0.004	0.001	-0.119	-5.403	$p < 0.001$
Student Gender	0.053	0.055	0.020	0.966	$p = 0.334$

[a]Beta coefficient represents the amount of change on the dependent variable if the independent variable is multiplied by the coefficient. [b]Standard Error of the estimate is a measure of the accuracy of predictions (Meyers et al., 2017). [c]Coefficients obtained if all the variables in regression model were standardized. [d]Statistical measure of significance of whether regression coefficient is significant (Meyers et al., 2017). [e]Measure of how meaningful changes in independent variable affects the dependent variable.

Results from SPSS show the predictive power of the student-centric variables on calculus course grades is moderate with approximately 39% of the total variation in calculus course grades being accounted for by the linear relationship between calculus course grades and the five student-centric variables (High School GPA, College GPA, College Hours Earned, Student Gender, and ACT Math Score). The results show that College GPA explains the vast majority of

the variance. The other 60% of the total variation in calculus course grades remains unexplained. Of the five student-centric variables examined, Student Gender was determined not to have any statistically significant predictive power when used in combination with the other four student-centric variables in examining the variation in calculus course grades. As such, the variable Student Gender was not included in HLM analysis.

Model 2: Model with Level-1 Predictors and No Level-2 Predictors

The results of Model 1 showed over 93% of the course grade variance concentrated at the student level. Model 2 will expand on the null model by adding the four significant student-centric variables.

The creation of Model 2 is motivated by the question of whether the addition of the four significant student-centric variables play a meaningful role in explaining the variance seen in calculus course grades. Regression analysis completed earlier determined Student Gender was not statistically significant when used in combination with the other four student-centric variables in determining calculus course grades and was not included in this analysis.

In terms of regression equations, the following equations are generated:

Level-1 Baseline Model

$$COURSEGR_{ij} = \beta_{0j} + \beta_{1j}*(COLLEGEGPA_{ij}) + \beta_{2j}*(HOURS_{ij}) + \beta_{3j}*(HSGPA_{ij}) + \beta_{4j}*(ACTMATH_{ij}) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \Upsilon_{00} + \Upsilon_{01} + u_{0j}$$
$$\beta_{1j} = \Upsilon_{10}$$
$$\beta_{2j} = \Upsilon_{20}$$
$$\beta_{3j} = \Upsilon_{30}$$
$$\beta_{4j} = \Upsilon_{40}$$

Mixed Baseline Model

$COURSEGR_{ij} = \Upsilon_{00} + \Upsilon_{10j}*(COLLEGEGPA_{ij}) + \Upsilon_{20j}*(HOURS_{ij}) + \Upsilon_{30j}*(HSGPA_{ij}) + \Upsilon_{40j}*(ACTMATH_{ij}) + r_{ij} + u_{0j}$

After running the model, the parameter estimates derived are:

Intercept = $\beta_{0j} = 2.20$ $\Upsilon_{00} = 2.20$

Coefficient for College GPA $= \Upsilon_{10} = 1.05$

Coefficient for College Hours $= \Upsilon_{20} = -0.004$

Coefficient for High School GPA $= \Upsilon_{30} = 0.24$

Coefficient for ACT Math Score $= \Upsilon_{40} = 0.05$

Variance at student level $= \sigma^2 = Var(r_{ij}) = 0.931$

Variance at instructor level $= \tau_0 = Var(u_{0j}) = 0.116$

Total Variance = $Var(\Upsilon_{ij}) = \sigma^2 + \tau_0 = 0.931 + 0.116 = 1.047$

The proportion of variance that can be explained in β_{0j} can be calculated as:

$$R^2 = \frac{\sigma^2_{00}\text{ (Null Model)} - \sigma^2_{00}\text{ (Model 2)}}{\sigma^2_{00}\text{ (Null Model)}} = \frac{1.562 - 0.931}{1.562} = 40.40$$

Model 2 shows that 40.40% of the course grade variance at the student-level can be accounted for by the four independent variables. Results show a Chi-square value of 195.12 and a df (degrees of freedom) of 16 producing a p-value of <0.001. By convention, the "cutoff" point for accepting these results is a p-value of 0.05 so these results provide evidence that the four student-centric variables are responsible for explaining a significant part of the variance in calculus course grades. These results are predictably similar to those obtained earlier using regression analysis.

Student-Centric Regression Model for Sections 01–06 and 68–73

Before moving on from Research Question Two, regression analysis was next used to create a model showing the relationship among the dependent variable (calculus course grades) and all five independent student-centric variables (College GPA, College Hours Earned, Student Gender, High School GPA, and ACT Math Score) for those students in Sections 01–06 and 68–73 from Table 3. Variables College GPA, College Hours Earned, High School GPA, and ACT Math Score were mean centered. Students in Sections 01–06 had the highest average calculus course grades while students in sections 68–73 had the lowest. The reason for conducting this analysis was to examine the role student-centric variables played in explaining the variances in calculus course grades for students at the extremes of the course grading spectrum and note any statistically significant differences from the analysis conducted on students in all of the course sections.

The regression formula is presented below.

$$COURSEGR_i = \beta_0 + \beta_1*(COLLEGEGPA_i) + \beta_2*(HOURS_i) + \beta_3*(SGEN_i) + \beta_4*(HSGPA_i) + \beta_5*(ACTMATH_i) + r_i$$

A summary of the analysis is presented in Tables 8 and 9.

Table 8
Regression Table Summary for Sections 01–06 and 68–73 Using Student-Centric Independent Variables

Regression Statistics	
Multiple R	0.637
R Square	0.406
Adjusted R Square	0.395
Standard Error	0.995
Observations	269

Table 9
Coefficient Table for Sections 01–06 and 68–73 Where the Dependent Variable is Student Calculus Course Grade Using Student-Centric Independent Variables

Student	Unstandardized Coefficients		Standardized Coefficients		
Predictors	β	Std. Error	Beta	t	Significance
(Constant)	2.065	0.101		20.454	$p < 0.001$
HS GPA	-0.181	0.151	-0.066	-1.193	$p = 0.234$
College GPA	1.065	0.093	0.723	11.400	$p < 0.001$
ACT Math Score	0.027	0.021	0.063	5.647	$p = 0.195$
College Hours	-0.004	0.002	-0.135	-2.361	$p = 0.019$
Student Gender	-0.035	0.134	-0.013	-0.258	$p = 0.796$

Results show little change from the regression model for students in Sections 01–06 and 68–73 as compared to the regression model for students in all 73 sections. The predictive power of the student-centric variables remains moderate with a little over 40% of the total variation in calculus course grades being accounted for by the linear relationship between calculus course grades and the five student-centric variables (High School GPA, College GPA, College Hours Earned, Student Gender, and ACT Math Score) with College GPA explaining the vast majority of the variance. The other 59% of the total variation in calculus course grades remains unexplained. Student Gender, ACT Math Score and High School GPA were determined not to have any statistically significant predictive power. The results of this analysis show that even for students in calculus sections where the section course averages are at the extreme, student-centric variables still explain approximately 40% of the total variation in calculus course grades.

Additional HLM analysis involving data from only Sections 01–06 and 68–73 was not an option due to the relatively small number of instructors for these sections as compared to the number of instructors for all 73 sections.

Research Question Three: To What Degree Do Instructor Characteristics (Age, Tenure, Experience, Gender, etc.) Explain the Differences in Mean Calculus Course Grades?

Research Question Three seeks to understand if calculus course grades differ as a function of who the instructor was. Using HLM to create a baseline model, it was discovered that approximately 6% of variance in calculus course grades resulted from instructor characteristics. In order to answer Research Question Three, I examine instructor-centric factors using SPSS and HLM to determine the level of calculus course grade variance these characteristics are able to explain. The goal is to determine how much calculus course grades vary as a function of the instructor-centric variables examined in this study.

Correlation analysis using SPSS was used to examine the relationships between course grades achieved in Calculus I and five instructor-centric characteristics. Table 10 shows the results of correlation analysis in which the five instructor-centric characteristics (Instructor Gender, Tenure Track Status, Years of teaching experience, Instructor Age, Section GPA average associated with each instructor) investigated.

Table 10
Correlation of Instructor-Centric Predictors to Calculus Course Grades

Correlations[c]

Instructor Variables		Course Grade	Instructor Gender_Mn	Tenure Track_Mn	Years of Teaching_Mn	Instructor Age_Mn	Section GPA_Mn
Course Grade	Pearson Correlation	1					
Instructor Gender_Mn	Pearson Correlation	.039	1				
Tenure Track_Mn	Pearson Correlation	-.005	.645**	1			
Years of Teaching_Mn	Pearson Correlation	-.075**	.267**	.415**	1		
Instructor Age_Mn	Pearson Correlation	-.095**	-.209**	-.165**	.725**	1	
Section GPA_Mn	Pearson Correlation	.220**	.164**	-.019**	-.352**	-.441**	1

**Correlation is significant at the 0.01 level (2-tailed).

c. Listwise N=1726

Examination of Table 10 illustrates that instructor-centric predictors Instructor Gender and Section GPA (.039, and .220 respectively) were positively correlated with calculus course grades. Of the positively correlated predictors, Section GPA (.220) had the strongest association with a calculus course grades. An instructor's age, years of teaching, and tenure track status variables were negatively correlated with a calculus course grades. Instructor Gender and Tenure Track Status were not significantly correlated at the .01 level.

With instructor-centric variables, there were two issues concerning multicollinearity. Correlation analysis revealed a highly correlated relationship between variables Instructor Gender and Tenure Track Status. A highly correlated relationship also exists between an instructor's age and the number of years they had taught. Further analysis in the next section using regression analysis would render these issues moot.

Regression Model Using Instructor-Centric Variables

In order to determine what instructor-centric variables were statistically significant in explaining the variance in calculus course grades, I used SPSS to conduct regression analysis. A model showing the relationship among the dependent variable (calculus course grade) and five instructor-centric variables (Instructor Section Grade Average, Years Teaching, Age, Gender, and Tenure Track Status) was created. Variables Instructor Section Grade Average, Years Teaching, and Age were mean centered. The regression formula is:

$$COURSEGR_i = \beta_0 + \beta_1*(SECTGRADE_i) + \beta_2*(TENURE_i) + \beta_3*(IGEN_i) + \beta_4*(IAGE_i) + \beta_5*(IEXPERIENCE_i) + r_i$$

A summary of the regression analysis is presented in Tables 11 and 12.

Table 11
Regression Table Summary Using Instructor-Centric Independent Variables

Regression Statistics	
Multiple R	0.291
R Square	0.085
Adjusted R Square	0.082
Standard Error	1.239
Observations	1976

Table 12
Coefficient Table Where the Dependent Variable is Student Calculus Grade Using Instructor-Centric Independent Variables

Instructor Variables	Unstandardized Coefficients β	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Significance
(Constant)	2.114	0.054		39.136	$p < 0.001$
Instr. Section GPA	1.003	0.080	0.290	12.567	$p < 0.001$
Tenure Status	-0.015	0.078	-0.005	-0.162	$p = 0.877$
Instr. Gender	0.010	0.079	0.004	0.132	$p = 0.895$
Instr. Age	0.000	0.006	-0.003	-0.064	$p = 0.949$
Years Teaching	0.000	0.006	0.004	0.072	$p = 0.942$

The results of regression analysis showed the predictive power of the instructor-centric variables on section grades being relatively weak in describing the variances of student calculus grades. Less than 9% of the total variation in calculus course grades can be explained by the linear relationship between course grades and the five instructor-centric variables (Instructor Section GPA, Gender, Age, Experience, and Tenure Track status). The other 91% of the total variation in calculus course grades remains unexplained. Of the instructor-centric variables examined, only average grades assigned by the instructor were found to have any significant predictive power in determining calculus course grades. With this finding, the issues of multicollinearity discovered during correlation analysis are rendered moot as the variables Gender, Age, Experience, and Tenure Track will no longer be included in HLM analysis.

Model 3: Model with No Level-1 Predictors and a Single Level-2 Predictor Instructor Section Grade

The results of Model 1 showed over 6% of the calculus course grade variance is concentrated at the instructor level. Model 3 will now expand on the null model by adding the one significant instructor-centric variable.

The creation of Model 3 is motivated by the question of whether the addition of the instructor section grade averages into the baseline model plays a meaningful role in explaining the variances seen in calculus course grades. In other words, we want to understand how the section grade averages of instructors help increase our understanding of differences in calculus course grades. Raudenbush and Bryk (2002) refer to this model as regression with Instructor-Section-Grade-as-outcomes. In terms of regression equations, HLM produces the following:

Level-1 Baseline Model

$COURSEGR_{ij} = \beta_{0j} + r_{ij}$

Level-2 Instructor Rank Model

$\beta_{0j} = \Upsilon_{00} + \Upsilon_{01}*(SECGPA\_MN_j) + u_{0j}$

Mixed Baseline Model

$COURSEGR_{ij} = \Upsilon_{00} + \Upsilon_{01}*(SECGPA\_MN_j) + u_{0j} + r_{ij}$

After running the model, the parameter estimates are:

Intercept = $\beta_{0j} = 2.18$ $\Upsilon_{00} = 2.18$

Coefficient for Section GPA = $\Upsilon_{01} = 1.062$

Variance at student level = $\sigma^2 = Var(r_{ij}) = 1.55$

Variance at instructor level = $\tau_0 = Var(u_{0j}) = 0.0004$

Given that the unconditional model and Model 3 share the same Level-1 model, the unconditional model is the appropriate choice as a baseline for comparisons. With the results of

the two models, I calculated a "pseudo R^2" value. Anderson (2012) wrote the following about deriving a "pseudo R^2" value:

> Unfortunately, there is no direct measure of the variance accounted for by HLM models—hence the name "pseudo." Pseudo R^2 statistics provide an indication of the amount of variance accounted for by comparing the variance component in an unconditional model to the same variance component in a conditional model. Pseudo R^2 can be calculated for the overall residual in the model, r_{ij}, or for any random parameter in the model (e.g., intercept variance). (p. 14)

With this model, the proportion of variance that can be explained in β_{0j} can be calculated as:

$$R^2 = \frac{\tau_{00}\text{ (Null Model)} - \tau_{00}\text{ (Model 3)}}{\tau_{00}\text{ (Null Model)}} = \frac{0.109 - 0.0004}{0.109} = 0.9963$$

Initial results from Research Question One, using HLM analysis, showed 6.52% of the grade variance occurring between instructors. This model shows that over 99% of the explainable calculus course grade variance associated with instructors can be accounted for by the section grading mean of the instructor. Results show a Chi-square value of 3.15 and a df of 15, producing a p-value of > 0.500. By convention, the "cutoff" point for accepting these results is a p-value of 0.05 or less so the results cannot be accepted with any certainty that the section grading mean of the instructor is responsible for a reduction in variance to student grades.

Instructor-Centric Regression Model for Sections 01–06 and 68–73

Before leaving Research Question Three, I used regression analysis to examine the relationship between the dependent variable (student grade) and the section course averages of

instructors for those students in Sections 01–06 and 68–73 from Table 3. As noted earlier, students in Sections 01–06 had the highest average course grades while students in sections 68–73 had the lowest. The reason for conducting this analysis was to examine the role section course averages of instructors played in explaining the variances in course grades for students who were enrolled in course sections that had the highest and lowest course averages.

The regression formula is presented below.

$$COURSEGR_i = \beta_0 + \beta_1*(SECTNGRADE_i) + r_i$$

A summary of the regression analysis is presented in Tables 13 and 14.

Table 13
Regression Table Summary for Sections 01–06 and 68–73 Using Instructor-Centric Independent Variables

Regression Statistics	
Multiple R	0.527
R Square	0.278
Adjusted R Square	0.275
Standard Error	1.110
Observations	306

Table 14
Coefficient Table for Sections 01–06 and 68–73 Where Dependent Variable is Student Calculus Grade Using Instructor-Centric Independent Variables

	Unstandardized Coefficients		Standardized Coefficients		
Instructor Variables	β	Std. Error	Beta	t	Significance
(Constant)	2.098	0.063		33.116	$p < 0.001$
Instr. Section GPA	1.008	0.093	0.527	10.831	$p < 0.001$

The predictive power of the Instructor Section GPA variable explaining the variance in calculus course grades is much more powerful for those students in Sections 01–06 and 68–73 as compared to the entire student dataset. In this case, almost 28% of the total variation in calculus course grades can be explained by the linear relationship between calculus course grades and section grade averages of an instructor. This compares to the Instructor Section GPA variable

explaining only 8% of the total variation in course grades for students in all of the sections. It appears that as you move to either end of the grading spectrum, the section grades of the instructor variable is able to explain more of the variance in calculus course grades. In these cases, the instructor's grading philosophy may be able to explain a greater part of the variance in calculus course grades.

As noted earlier, HLM analysis involving data from only Sections 01–06 and 68–73 is not an option due to the relatively small number of instructors for these sections as compared to the number of instructors for all 73 sections.

Research Question Four: Do Instructor Characteristics Influence the Magnitude of the Student Success Measures on Student Course Grades?

The last research question seeks to understand the relationship between calculus course grades and student-centric variables varying as a function of instructor calculus course averages. HLM was used to understand the degree of influence instructor course averages had on calculus course grades, while taking into account student-centric variables. Only statistically significant student and instructor-centric variables were used to create this two-level model.

Model 4: Model with Level-1 Predictors and One Level-2 Predictor

In terms of regression equations, HLM produces the following:

Level-1 Baseline Model

$$COURSEGR_{ij} = \beta_{0j} + \beta_{1j}*(COLLEGEGPA_{ij}) + \beta_{2j}*(HOURS_{ij}) + \beta_{3j}*(HSGPA_{ij}) + \beta_{4j}*(ACTMATH_{ij}) + r_{ij}$$

Level-2 Instructor Rank Model

$$\beta_{0j} = \Upsilon_{00} + \Upsilon_{01}*(SECGPA\_MN_j) + u_{0j}$$
$$\beta_{1j} = \Upsilon_{10}$$
$$\beta_{2j} = \Upsilon_{20}$$
$$\beta_{3j} = \Upsilon_{30}$$

$\beta_{4j} = \Upsilon_{40}$

Mixed Baseline Model

$COURSEGR_{ij} = \Upsilon_{00} + \Upsilon_{01}*(SECGPA\_MN_j) + \Upsilon_{10j}*(COLLEGEGPA_{ij}) + \Upsilon_{20j}*(HOURS_{ij}) + \Upsilon_{30j}*(HSGPA_{ij}) + \Upsilon_{40j}*(ACTMATH_{ij}) + r_{ij} + u_{0j}$

After running the model, the parameter estimates are:

Intercept = $\beta_{0j} = 2.19$ $\Upsilon_{00} = 2.18$

Coefficient for Section GPA $= \Upsilon_{01} = 1.06$

Coefficient for College GPA $= \Upsilon_{10} = 1.05$

Coefficient for College Hours $= \Upsilon_{20} = -0.004$

Coefficient for High School GPA $= \Upsilon_{30} = 0.24$

Coefficient for ACT Math Score $= \Upsilon_{40} = 0.05$

Variance at Instructor Level $= \tau_0 = Var(u_{0j}) = 0.0003$

Variance at Student Level $= \sigma^2 = Var(r_{ij}) = 0.92615$

Total Variance $= Var(\Upsilon_{ij}) = \tau_0 + \sigma^2 = 0.0003 + 0.92615 = 0.92618$

With this model, the proportion of variance that can be explained in β_{0j} can be calculated as:

$$R^2 = \frac{Var(\Upsilon_{ij})\ (Model\ 1) - Var(\Upsilon_{ij})\ (Model\ 4)}{Var(\Upsilon_{ij})\ (Model\ 1)} = \frac{1.671 - 0.926}{1.671} = 0.4458$$

This model shows that 44.58% of the calculus course grade variance can be accounted for by the four student-centric and one instructor-centric variable.

In addition, this model shows the magnitude to which instructor course averages influence student success measures on calculus course grades:

$$R^2 = \frac{\text{Var}(r_{ij})\ (\text{Model 2}) - \text{Var}(r_{ij})\ (\text{Model 4})}{\text{Var}(r_{ij})\ (\text{Model 2})} = \frac{0.931 - 0.926}{0.931} = 0.005$$

This model shows that the instructor course averages influence on student success measures concerning calculus course grades is minuscule. However, results show a Chi-square value of 5.27 and a df of 15, resulting in a p-value > 0.500. By convention, the "cutoff" point for not rejecting the null hypothesis is a p-value of 0.05, so the results indicate that the null hypothesis is not rejected with any certainty that the addition of the instructor course grade variable along with the four student-centric variables are responsible for explaining a significant reduction in variance to calculus course grades.

Summary

Using HLM, statistical models between students and professors based on the studied characteristics of both professors and students across multisection Calculus I courses at the university studied were created to answer four research questions.

Summary for Research Question One

Initial examination of Chart 3, along with Table 3, shows that the Calculus I course grade averages varied a great deal between the years 2009 and 2013 among the different sections and instructors. The bottom six sections (Sections 68–73 on Table 3) had a course average of 1.34 with over half of the students failing the course. The high achievement sections (Sections 01–06) had a course average of 2.71 with over 80% of the students successfully passing the course. A cursory appraisal of Table 3 shows that while section course grades varied a great deal, it does not appear that the makeup of students in regard to their academic characteristics varied much between instructors.

A similar appraisal of Table 4 shows that calculus section course grades varied a great deal by instructor while the makeup of students in regard to their academic characteristics did not appear to vary much between instructors. Students taught by Instructor 01 had a success rate of less than 50%, while students taught by Instructor 17 enjoyed a success rate in excess of 80%.

However, HLM analysis showed that 93.48% of the calculus course grade variance occurs at the student level with the remaining 6.52% of the grade variance occurring between instructors. While wide differences in calculus course averages exist by both sections and instructors (Tables 3 and 4), the differences in calculus course grades issued to students result primarily from differences in individual student characteristics. Instructor characteristics played only a minor role in explaining the calculus course grade variances when looking across the entire spectrum of the 73 classes.

Summary for Research Question Two

Using HLM to create a baseline model to answer Research Question One, it was discovered that over 93% of variance in student grades resulted from student characteristics. Research Question Two examined student-centric factors using SPSS and HLM to determine the level of course grade variance these characteristics were able to explain. Across the board, using both SPSS and HLM, the student-centric variables were able to explain approximately 40% of the variance associated with calculus course grades. The remaining 60% of the calculus course grade variance associated at the student level remains unexplained.

There was little change between the regression models produced in SPSS involving every student and with students only in Sections 01–06 and 68–73. Both models showed approximately 40% of the variance in calculus course grades being explained by student-centric variables.

Summary for Research Question Three

Using HLM to create a baseline model to answer Research Question One, it was discovered that approximately 6% of variance in calculus course grades resulted from instructor characteristics. Research Question Three examined instructor-centric factors using SPSS and HLM to determine the level of calculus course grade variance these characteristics were able to explain. Results from HLM showed that over 99% of the explainable calculus course grade variance associated with instructors can be accounted for by the section grading mean of the instructor. However, due to the small sample size of instructors, no conclusions with any certainty can be drawn that the section grading mean of the instructor is responsible for a reduction in variance to student grades.

Regression analysis showed that the Instructor Section GPA variable was able to explain a much greater portion of the variance in calculus course grades for those students in Sections 01–06 and 68–73 as compared to the entire student dataset. The results provide evidence that the instructor's grading philosophy would be able to explain a greater part of the variance in student grades as you move to either end of the grading spectrum.

Summary for Research Question Four

Results from HLM showed that over 44% of the explainable grade variance can be accounted for by the significant predictors used in this study. The remaining 56% of the calculus course grade variance remains unaccounted for.

The instructor course grade variable has virtually no effect on the student-centric variables in the model. However, due to the small sample size of instructors, no conclusions with any certainty can be drawn that the section grading means of the instructors are responsible for influencing the magnitude of student-centric variables on student course grades. The next chapter will further discuss the findings of this chapter.

CHAPTER V: SUMMARY, CONCLUSION, AND RECOMMENDATIONS

Statement of the Problem

Achieving exceptional grades in higher education has increasingly become an economic necessity for many students. Grades are required by deans, honor societies, graduate admission committees, and other governing powers beyond the classroom (Filene, 2005). Given their mission to improve student lives through education, it is incumbent upon colleges and universities to know each student's background and provide them with a learning environment that allows them to succeed in multisection calculus courses.

Low student success rates in multisection calculus courses have the potential to deliver a bevy of gloomy financial and academic problems to a large swath of students. This harm may include students dealing with increased course retakes and course droppings in order to graduate, costing them both time and money. These students may have to contend with switching majors, delaying graduating, and/or achieving a lower overall GPA. All of these issues may hinder their ability to graduate, procure internships, and hiring prospects upon graduation. As Ohland, Yuhasz, and Sill (2004) posit, "failing or withdrawing from Calculus I, or not being prepared for it, had a much greater effect on a student's progress than failing other required courses" (p. 253).

Issues surrounding students unsuccessful in passing calculus can also negatively affect a university. Impacted students may not have the means or the willingness to give back to the university in any capacity, hurting the university's ability to increase its endowment fund. Any decrease in a university's endowment fund hampers its ability to offer scholarships to students, expand academic and nonacademic programs, and building maintenance and expansion. Lower student satisfaction with a university is likely to lead to a decline in a university's ranking in

influential magazines and websites resulting in both a decline in prestige and future tuition dollars as some students opt to attend elsewhere.

Purpose of the Study

Habre and Abboud (2006) noted that calculus courses have high failure rates and therefore serve as a barrier for many students interested in well-paying STEM fields. These high fail rates present an opportunity for educators to better understand the challenges that students face in general calculus courses in order to reduce the failure rates. Student persistence in college and degree completion depend heavily on the instructors and instruction received during their first semester at college (Erickson, Peters, & Strommer, 2006). These challenges include grading policies that, as Hu (2005) posited, are not well understood. With that in mind, the purpose of this study was to shed a light on the grade variances in multisection calculus classes at a publicly supported university by using data mining techniques and predictive modeling in order to understand the reasons for the grade variances. Understanding the root causes of the grade variances will allow for more effective, tailored solutions that can be implemented across campuses to increase student success rates in calculus.

I devoted the first chapter of this dissertation to covering the issue of grading disparities in higher education and outlining the purpose of grades, along with the methodology used and limitations of the study. In Chapter Two, I reviewed the history of grading students in higher education and touched on current stakeholder views concerning grading. With Chapter Three, I laid out the research design of this project and used Chapter Four to review the results of this research. In this chapter, I present a summary and discussion of the findings and recommendations for increasing success rates in multisection calculus classes.

Research Questions

Walvoord and Anderson (2010) opined, "Most of us know that grades are too often inadequate, imprecise, and wildly idiosyncratic indicators of learning" (p. ix). Even so, there is a lack of scholarly research articles that critically look at grading differences in multisection classes and what causes those differences. This paper contributes to the literature by filling this evidential gap by seeking to answer the following research questions:

1. Are there differences in student Calculus I course grades among the college students between the course sections?
2. To what degree do student characteristics (gender, ACT, high school GPA, etc.) explain differences in course grades?
3. To what degree do instructor characteristics (age, tenure, experience, gender, etc.) explain the differences in mean course grades?
4. Do instructor characteristics influence the magnitude of the student success measures on student course grades?

Brief Literature Review

Student assessment has always been a divisive and constantly changing issue in higher education. The beginning period of higher education in this country witnessed student evaluations consisting of faculty members providing narrative feedback to students based primarily on their abilities (Marzano, 2000). The first grades of any kind were introduced at Yale University in 1785 (Stiles, 1901). During the 19th century, various marking systems came and went throughout higher education without any of them becoming a de facto standard.

After the Civil War, high school grades achieved by students increasingly became a factor in college admissions (NEA, 1974). By the beginning of the 20th century, most colleges

used some form of grades to evaluate students. Studies conducted by M. Meyer (1908), Finkelstein (1913), and other researchers showed grading by professors varied greatly at most colleges. Beginning around World War I, a belief arose among some academics that college grading should conform to either a normal distribution scale or a ranking of students against other students when assigning grades started. Studies by Gaw (1926), E. Abell (1928), and others helped shape these beliefs as they concluded that student grades and intelligence tests should be closely correlated. Other researchers, like Korey (1926) and Crew (1930), felt these and other forms of ranking grade systems did more harm than good. They held the belief that grades were not only an impediment to learning as they diverted students' attention away from actual learning but also had the power to inflict psychological harm on students. Opponents of grading, like Odell (1928), argued that due to the subjective nature of grades, they should not be used as a student-measuring device. A study by Travers and Gronlund (1950) showed that because some professors took into account features like attendance, attitude, and student interest when making grade determinations, grades, in their present form, offered limited value.

Despite the onset of the Great Depression, research on grading in higher education continued with the general belief that IQ tests and normal distribution of grades were the proper methods to assess students. Research by Eells (1930), Middleton (1933), Segel (1934), and others continued to adhere to this belief, positing that this would be one way to minimize the inherent subjectivity in grading. Still others, like Broome (1945) and Good (1945), continued a line of research pointing out not only the subjective nature of grades, but also the psychological harm grades inflicted upon students. In this vein of research, a study conducted by Hadley (1954) resulted in his belief that girls were awarded higher grades simply because they tended to be better behaved in the classroom in comparison to their male counterparts.

As a result of students receiving higher grades throughout the 1960s, scholarly research began focusing on the causes of grade inflation with researchers generally concluding that grade inflation resulted from the Vietnam War and implementation of a compulsory draft (Hunt, 2008; Rojstaczer & Healy, 2012; Rosovsky & Hartley, 2002). Changes to grading schemes continued in the 1970s with universities experimenting with new grading policies to evaluate students. Pass/Fail grading and grade forgiveness schemes were adopted with various degrees of success across college campuses (Milton et al., 1986). Influential researchers like Hu (2005) and Hills (1972) continued a line of research pointing out the inherent shortcomings of grading and general unfairness to students.

Research involving grading in higher education began to branch out during the 1980s to investigate student behaviors in response to grades. Chizmar (2000), Rask and Tiefenthaler (2008), Sabot and Wakeman-Linn (1991), and others concluded that student course enrollment into specific courses was affected by the assignment of grades. Other researchers noticed the rise of student entitlement concerning grades with students feeling they have earned a particular grade without putting in the corresponding effort (Ciani et al., 2008; Vallade et al., 2014).

Another grade-related avenue researchers have examined is how professor views and backgrounds influence the grading process. Role theory has been the primary lens researchers have used when examining various roles instructors play within the university setting concerning grades. Schutz, Drake, and Lessner (2013) and Sonner (2000) pointed out that grades meted out by full professors differed from those of part-time faculty and graduate students. Grading differences were also noted between professors who favored criterion-referenced grading systems versus professors employing norm-referenced grading systems (Geisinger, Wilson, & Naumann, 1980). Marzano (2000) noted that the grading differences also stemmed from what

material professors chose to grade and what student behaviors should be included when assessing students.

Since the founding of Harvard University, assessment practices in higher education have continued to evolve. While the actual assignment of letter grades on college campuses is practically universal, how they are derived remains in a constant state of flux varying from school to school, department to department, and instructor to instructor. Within this constant state of change exists these multisection courses where I investigated the variances in student grades in calculus classes in an effort to improve success rates in those classes.

Limitations of the Study

A number of limitations may have influenced the results of this study. I utilized only student-centric and instructor-centric variables provided by data residing on the university's mainframe system for this research. It is quite possible that socioeconomic factors unavailable to the researcher may have provided additional insight into the assignment of course grades. While a sincere effort was made to cleanse the data obtained from the university's mainframe system to minimize the possibility of miscoded data along with other unknown factors that may have adversely influenced the results of the study, there remains the possibility of corrupted data influencing the results of this study.

Data used for this study involved only participants from a single campus of a four-year Midwestern public university. It is possible that conclusions reached in this study may not be representative of other universities and colleges across different regions. While time and monetary constraints prevented the inclusion of data from other universities, the results of this study would have been enhanced if data from other universities were included. It is also recognized that some student-centric variables related to students' overall GPA and Course

Hours Completed may have changed slightly since the time the data was collected as a few of the students were still enrolled after the collection date.

It is recognized that while high school GPAs and ACT subscores utilized to serve as predictors of students' success in college and measurement of students' abilities, these measures are not foolproof and serve only as imprecise substitutes for understanding student effort and abilities. Conducting individual student interviews to uncover effort put forth towards their studies, engagement in the classroom, socioethnic backgrounds, and record of hours worked while attending school may have improved the generalizability of the outcomes unearthed in this study.

Instructor interviews to uncover their views on grading, teaching methods, and effort put forth towards instructing may have improved the generalizability of the outcomes unearthed in this study. Without interviewing individual instructors, the author used the average section GPA as a proxy for instructor grading views. Admittedly, student grades alone provide limited value about an instructor's value in the classroom. In addition, this study examined 73 calculus sections over a five-year period that were taught by only 17 different instructors. It is acknowledged that having additional instructor data would have made the results more generalizable.

Research Design

For this study, a secondary data analysis design was selected using grade data from a state-supported public institution of higher education in the Midwest. In nonexperimental design, according to Mertler and Vannatta (2005), test subjects enter the study already belonging to various groups with the researcher only defining the independent variables. The target population for this study was student grade data from a publically supported university. I examined data

from Calculus I classes taken by students between 2009 through 2013 to investigate predictors of the variance in student grades in a multisection format. During this time, over two thousand students received a grade in Calculus I at the university studied. Patterns of student course enrollment, grade point averages, persistence rates, and degree selection were examined.

Student and instructor data came from the university's mainframe system that housed the information. Excel and IBM SPSS Statistics 20 were used to put the data in a useable format to analyze. One-way ANOVA, regression and correlation analysis using Excel, Hierarchical Linear Modeling (HLM), and IBM SPSS Statistics 20 were used to analyze the data.

Summary of Findings

Research Question One: Are There Differences in Student Calculus I Course Grades Among the College Students Between the Course Sections?

Examination of Tables 3 and 4 in Chapter Four illuminated wildly different grading patterns existing between multisection Calculus I classes and between instructors who taught those classes. Further analysis showed that differences in course grades issued to students resulted primarily from differences in individual student characteristics. HLM analysis revealed that over 93% of the calculus course grade variance were explained by student-related characteristics with the remaining balance explained by instructor-related characteristics.

This proved to be a surprising finding considering the results in Tables 3 and 4. When starting this study, I made the initial assumption that the reasons behind the divergent course averages likely resulted from how individual professors assessed students in their abilities to master the course material. HLM analysis clearly showed that the divergent course averages resulted primarily from differences in the student-related characteristics of those students taking calculus.

My initial assumptions regarding the divergent grading patterns seen in Tables 3 and 4 were shaped by research concerning grading that has primarily focused on the subjective nature of grading, ways to improve grading, how to improve student assessment, and grading philosophies of instructors. Most of the criticisms regarding the subjective nature of grading were eloquently laid out decades ago by M. Meyer (1908), Finkelstein (1913), and Starch and Elliott (1912). Milton et al. (1986) describe higher education's attempts to change the grading process over time that included moving from oral assessment to written exams, the introduction of pluses and minuses to the five-letter grading system, and pass/fail grading schemes. Walvoord and Anderson (2010), among others, have spent a great deal of time developing tools to better assess college students. A number of research lines have been developed to help explain differences in grading philosophies among professors. Schutz et al. (2013) supported the idea that grading differences exist between professors who have a sense of belonging to the university versus those professors who feel less connected. Milton et al. (1986) advanced "adaptive-level theory" to explain varying grading standards among faculty as a reflection of what academic department they were working in. Hunt (2008) posited that some professors hold the belief that the purpose of grading is a mechanism to sort, rank, or act as a gatekeeper. Van Ness et al. (1999) advanced the idea that part-time instructors generally assign higher grades than full-time instructors do.

While all of these research studies have merit, they have not honed in on the idea that the majority of grade variation among students is explained by student-related characteristics. As such, previous research has focused on issues that may not greatly help students succeed in college. HLM results from this study reveal that educators need to focus on the idea of grades

being primarily determined by student-centric factors if the goal is to find way to improve student success rates.

In practical terms, college academic advisors can use these results to impress upon students to avoid misplacing blame on any poor grades students receive as the results of this study point to their grades being primarily decided by student attributes. Instructors can also take comfort in know that the grades they assign are primarily the result of student attributes. Students, armed with this information, can take solace in knowing that the grades they receive are not likely the result of overzealous grading standards imposed by the instructor.

Research Question Two: To What Degree Do Student Characteristics (Gender, ACT, High School GPA, etc.) Explain Differences in Course Grades?

The aim of this research question was to determine how much student grades vary as a function of the five student-centric variables examined in this study. Correlation analysis was initially done to probe the relationship strength between student course grades achieved in Calculus I and the student-centric characteristics (High School GPA, ACT Math Score, College Hours Completed, Student Gender, and College GPA). Correlation analysis also provided a check for multicollinearity issues. All five student-centric predictors were significantly correlated at the 0.01 level with student grades with no issues of multicollinearity.

Regression analysis was then administered to determine which of the five student-centric variables were statistically significant in explaining the variance in student course grades. Student Gender was the only student-centric variable determined not to be statistically significant in explaining the variances in course grades for students. College Hours Completed, while statistically significant, was determined to have only a negligible effect in explaining the variances in course grades for students. College GPA was found to be the most significant

student-centric variable in explaining course grades with all significant variables explaining approximately 40% of the total grade variance attributable to students. This finding was consistent when looking at the entire student population as well as only those students in sections 01–06 and 68–73 from Table 3.

HLM analysis was next conducted to help explain the variance in student grades with regard to changes in the four significant student-centric variables (High School GPA, ACT Math Score, College Hours, and College GPA). The HLM model, like regression analysis, showed approximately 40% of the grade variance associated at the student-level being accounted for by High School GPA, ACT Math Score, and College GPA. The remaining grade variance associated at the student-level is accounted for by student-level variables not evaluated in this study.

Numerous studies using college entrance exams and high school grade point averages have been used to predict first-year college GPAs (Cohn, Cohn, Balch, & Bradley, Jr, 2004; Hoffman & Lowitzki, 2005; Noble & Sawyer, 2002). Prior research from Sabot and Wakeman-Linn (1991) and Credé, Roch, and Kieszczynka (2010) determined that the student characteristics of gender, grades, college hours completed, and ACT scores were significant predictors of student GPAs in college. Among available admissions data, Astin (1991) believed that a student's high school GPA and standardized test score were the two strongest predictors of student success.

Results of this study confirm prior research findings that a student's math ACT score, high school grades, and college grades play a statistically significant role in explaining the variance in student course grades. This study found that college hours completed by students played only a small statistically significant role in explaining the variance associated with

student course grades. While both ACT scores and high school grade point averages partially explained the variance to student grades, it was not to the degree that was expected. The predictive ability of these two variables may have been muted to a degree in this study given most of the students in the calculus classes had high school GPAs and ACT math scores above the student body average at the university studied.

In 2005, Dr. Lawrence Summers was the featured speaker at a conference hosted by the National Bureau of Economic Research (Dillon, 2005). The then president of Harvard University declared that men outperform women in mathematics due to biological differences. A meta-analysis on gender differences in mathematical tasks conducted by Friedman (1989) provided concrete evidence against the idea that women are inferior in mathematical capabilities when compared to their male counterparts. Dr. Summers resigned in 2006 from his position at Harvard University after a vote of no confidence by the faculty (Wilson, Fain, Fogg, & Selingo, 2006). As noted earlier, results of this study determined a student's gender did not play a statistically significant role in in explaining the variance in student course grades.

What this dissertation and other studies lack is a more complete picture of student-characteristics that lead to high grades and how these characteristics interplay with each other. This study, using regression and HLM analysis, showed that approximately 40% of the grade variance associated at the student-level is accounted for by a student's high school GPA, ACT math score, college hours, and college GPA. That still leaves approximately 60% of the grade variance associated at the student level unaccounted for by student-level variables not evaluated in this study. Choy (2001) noted that a student's success in college is partially based on parental income and education levels. Kuh, Cruce, Shoup, Kinzie, and Gonyea (2008) found that student engagement in educational activities is positively correlated with better academic outcomes for

students. Knowing a greater percentage of student-characteristics and how they interact with one another at the individual level would be an important first step that can then be expanded upon to help educators create a better learning environment to increase student successes.

Research Question Three: To What Degree Do Instructor Characteristics (Age, Tenure, Experience, Gender, etc.) Explain the Differences in Mean Course Grades?

This question addressed the issue of wanting to know if student grades differed as a function of who the instructor was. Correlation analysis was first done to examine the relationship strength between student course grades achieved in Calculus I and five instructor-centric characteristics (Section GPA average associated with each Instructor, Tenure Track Status, Instructor Gender, Age of Instructor, Years of Teaching Experience). Correlation analysis also provided a check for issues concerning multicollinearity. Analysis showed that Instructor Gender and Tenure Track Status were not significantly correlated at the .01 level with student course grades. Instructor Section GPA was shown to have the strongest relationship with student course grades.

Regression analysis was used next to determine what instructor-centric variables were statistically significant in explaining the variance in student course grades. Only the Instructor Section GPA variable was determined to be statistically significant in explaining the variances in course grades for students. HLM analysis was then conducted to help increase our understanding of differences in student grades with regard to changes in Instructor Section GPA. The HLM model showed that over 99% of the explainable grade variance associated at the instructor level could be accounted for by the section-grading mean of the instructor. However, given that data for only 17 instructors was used in the study, the p-value from this analysis was not low enough to firmly conclude this finding with any certainty.

Using regression analysis, this study found evidence that the instructor's grading style played a more pronounced role when looking at only students in Sections 01–06 and 68–74 from Table 3. In these limited cases, the prior research noted above may have yielded limited insights as to why instructors choose to grade in a manner that differed from their peers assuming that the grading style of these professors could be determined. While it was beyond the scope of this study to determine the grading philosophies among professors teaching the calculus courses under investigation, the benefits of doing so may have been muted given that only a small percentage of the variation in student grades was determined to be at the instructor level.

Previous research has found some support for adjunct faculty differing from full-time faculty in their perceptions of rigor in assigning grades (Schutz et al., 2013; Sonner, 2000; Van Ness, Van Ness, and Kamery, 1999). The conclusions reached in these studies were that adjunct faculty tended to award higher grades than tenure track faculty. This study found no evidence to support the idea that grading differences existed between nontenure track and tenure track faculty concerning the calculus classes under examination.

DeBoer, Anderson, and Elfessi (2007) examined instructor attitudes towards grades concluding that "although there are significant differences between men and women…with respect to ascription of responsibility [of grades], there appears to be no consequence on grading behavior" (p. 62). Results from this dissertation showed that the age of the instructor, gender, and experience played little to no role in explaining the variance of student grades.

Research by Armstrong (2000) concluded that between 15 and 20 percent of the amount of variance in students' final grades in community college settings involving entry-level English and mathematics courses could be explained by instructor characteristics. Results from HLM analysis in this study concluded that approximately 7% of the amount of variance in students'

final grades could be explained by instructor characteristics. With all this in mind, research aimed at improving student success rates might best be served by focusing on identifying student characteristics that increase student pass rates and then making an effort to inform professors what these characteristics are and how they can use this information to increase student success rates.

Research Question Four: Do Instructor Characteristics Influence the Magnitude of the Student Success Measures on Student Course Grades?

HLM results showed that the only significant instructor-level variable in explaining the variance to student grades, Instructor Section GPA, had virtually no effect on the student-centric variables in the model. However, due to the small sample size of instructors, no conclusions with any certainty could be drawn using HLM that the section grading means of the instructors were responsible for influencing the magnitude of student-centric variables on student course grades.

These results were not surprising concerning an instructor's influence on student variables ACT Math Score and High School GPA, since those variables were determined before a student had set foot on campus. While it was beyond the scope of this research to interview instructors to ascertain their sense of belonging to the university and whether the instructors favored either a criterion-referenced grading system or a norm-referenced grading system, doing so may have yielded new insights. Schutz et al. (2013) found that grading differences exist between professors who have a sense of belonging to the university versus those professors who feel less connected. Barnes et al. (2001) found support that individual professors tend to favor either a criterion-referenced grading system or a norm-referenced grading system. The overall results of this research point to the surprising general lack of impact instructors have on student

grades. As such, regardless of how instructors view themselves in the environment in which they instruct, their impact on student grading appears muted.

Discussion

"Despite the fact that grading and reporting have been the subject of innumerable studies, current policies and practices tend to be based more on opinion than on thoughtful analysis" (Guskey & Bailey, 2001, p. 11). When I first started this study, it was my opinion that the reasons behind the divergent course grading averages seen in Tables 3 and 4 would rest primarily with the course instructor. This assumption should not come as a total shock given the great diversity of views faculty harbor concerning grading and assessment. A professor's views on grading are shaped by a myriad of events. In my travels, I have encountered numerous professors with no two having identical views on grading. It seems the questions surrounding grading are many without consensus that has vexed educators for centuries. What should be graded? What kind of test, if any, is needed? What grading scale should be imposed? How many points should this problem be worth? What about partial credit or points for attendance? Should norm- or criterion-referenced grading systems be used? What about the outlier student?

What the questions above bring out is that while what should be graded may seem straightforward, grading is layered in both complexity and ambiguity. Hu (2005) points out that "notions and standards of grading vary by educational context and among individuals with different philosophies" (p. xiii). I have encountered professors who stress some course material at the expense of other material despite what other professors teaching the same course do. Some professors have a greater understanding of certain topics over others and, as a result, cover that material in greater depth. Other professors have an affinity for obscure material that is

overrepresented in their exams. For a few of the exams I have taken over the course of my life, I still have no idea what topics the professor was trying to cover.

Those and other thoughts ran through the author's head as data collection began in earnest. Initial review of Charts 3 and 4 appeared to confirm initial suspicions that instructors could play a major role in explaining the grade variances of students in multisection calculus classes. Chart 3 unmistakably showed course averages varying a great deal between the sections while Chart 4 revealed course averages varying a great deal between the instructors.

However, section course averages proved to be much different from the individual student course grades that comprised the section course averages. HLM analysis used to answer Research Question 1 showed that over the 93% grade variances of students in multisection calculus classes being explained by student-centric variables. That left only a minor role for instructor-centric variables to explain the grade variances of students. Further analysis showed that of the student-centric variables examined, a student's college GPA was found to explain a significant amount of the grade variation among students taking calculus.

As noted earlier, a course grade alone says little about the benefits conferred to students from individual instructors. While debate will continue as to what to assign, what to grade, and how to assess, this study pointed to instructors playing only a minor role in determining a student's grade. Perhaps a more fruitful debate is determining ways for educators to understand the ways student centric characteristics effect their grades and then use those findings to improve student success rates in the classroom.

Policy Implications

In 1963, Adams published research concerning the issue of equity in the workplace environment. His seminal work focused on whether the distribution of resources was fair to employees and employers. Adams felt that if workers felt they were being treated fairly relative to one another and the company, they would be more motivated. This line of research has continued into the education arena with Chory-Assad (2002) asserting that student perceptions concerning grading fairness affect student motivation and learning. Nesbit and Burton (2006) concluded that poorly performing students that expressed concerns over grading practices appeared to be more negatively affected by those concerns as compared to other students who did not express the same level of concern over grading practices.

Students made aware of this study should take comfort in knowing that that over 93% of the grade variation among students seen in Tables 3 and 4 can be explained by student-related characteristics. As such, this study provides support that, for the vast majority of students, they are firmly in control when it comes to determining their grade regardless of the calculus section enrolled in. In addition, while the student-centric predictors used in this study explained some of the variance to calculus grades, this study could not explain a great deal of the variation in grades related to students, likely leaving plenty of room for a student's effort, determination, and class attendance to play a key role in grade determination. All these factors are within a student's control.

Given the high level of grade determination being explained by student-level characteristics, mathematics advisors need to ensure that high-quality resources are in place to aid students outside of the classroom. Unfortunately, most students in need of academic support services are less likely to take advantage the services offered at their university (Friedlander,

1980). Academic counselors could use the results of this study to impress upon students taking calculus, whose academic profile suggests they may struggle, the need to get extra help if needed early on in the semester to avoid failing the course.

The results of this study surmise that only when the section grade averages move to the extreme does it appear that the instructor's grading philosophy plays an outsized role in determining a student's grade. In many of these cases, instructors may not even be aware of the divergent course averages within the multisection courses. To help reduce some of the variation at the extremes, departmental chairs could share grading information that included student success rates with faculty allowing them the opportunity to gain some perspective on their grading philosophy compared to others within the department. This would allow professors teaching these courses to get a different frame of reference if they looked at the grading differences across the department and the impact these differences had on students.

This study lends a degree of support to the argument that resources be in place in the Provost office in an effort to monitor multisection courses when course grade averages move too far from the overall average. It should be salient to university policymakers that policies and procedures be put in place to rigorously examine class grade averages each semester. Classes with significant deviations from other multisection class averages should be fully investigated. The underlying causes may be outside the control of both the professor and department, in which case both the department and professor would have a legitimate reason for a deviation from the norm.

These recommendations are not meant to take away from any kind of academic freedom that professors enjoy. These recommendations need to be implemented so that academic freedom among professors continues in the search for knowledge and understanding. The individual

teaching styles that make for successful instructors need to remain in place. The variability seen in calculus grades among students, while largely due to student-centric characteristics, has the potential effect of harming students if the course grades stray too far from the norm. If this problem arises, professors run the risk of college administrators forcing changes upon them in an effort to solve the problem. One has to be cognizant of the fact that where there is noncompliance, the organization can enforce compliance by way of "overt" or "covert" sanctions (Katz & Kahn, 1978). Professors' rights to academic freedom are not without limits. In the 1989 case Parate v. Isibor, the court's opinion was a university has a right to "reasonably determine how classes are to be taught and grades are assigned" (Fossey, 2007, p. 161). It would be better for professors to get out in front of this issue before college administrators do.

Ironically, while I am a staunch believer in grading consistency and fairness, I am an advocate for having exceptions to rules. Overly burdensome, rigid rules have the potential to be harmful to students as well. A year or two ago, a student in my class missed taking his final without notice. The syllabus for the class clearly stated the penalty for such an action was a failing grade. Two days later, the student showed up in my office to inform me why he missed taking the final. We went for a walk around the quad where the student, with tears in his eyes, informed me that over the past weekend, his father had disowned him because of his lifestyle choices.

I saw a young man, already working fulltime with an incredibly bright future, broken by events much bigger than school. I listened and consoled the student as we made our way around the quad at least a half-dozen times. As far as his course grade was concerned, I did not have to do anything. The course syllabus coupled with academic freedom had granted me that privilege. That same academic freedom also allowed me to let the student take the exam a few days after

our walk. I do not remember how the student did on the exam, and I do not think any score on it would have prevented me from passing the young man anyway.

I ran into the student a few months later at a hardware store parking lot. He came up to me and thanked me profusely for listening to him in his time of need. He was in a much better place both physically and emotionally as he had graduated school and was moving to Phoenix because of a job promotion with his significant other. The conversation ended with a hug and each of us wishing the best for each other, knowing our paths would likely never cross again.

With compassion and sensibility, professors have the ability to positively affect student lives that extend far beyond sensible grading. If professors lose sight of compassion and doing right by students, perhaps it is not the students who should be evaluated. Even though it may be a time-consuming task, being aware of individual students' issues is one way instructors can help increase student successes and perhaps influence course grades by more than the 7% that the HLM analysis revealed.

Recommendations for Future Research

Based on the review of the literature and the review of the data collected and analyzed, the following recommendations are made for further research.

1. The results of the study suggest that there are a number of student-centric variables not used in this study that would help explain the variances in student grades. It was outside the scope of this study to interview students about their study habits, class attendance, and parental income and educational levels. Surveying students to ascertain their study habits, class attendance, and parental income and educational levels across academic disciplines combined with the student-centric variables used in this study would paint a clearer picture surrounding the

variances in student grades given that so much of the grade variance is attributable at the student level.

2. Additional analysis needs to be conducted on the grades of students in calculus sections where the average course grades were at the extreme. The study provided evidence that instructors' grading schemes took on a greater role in determining a student's grade in these sections. A study with a much larger set of classes and greater number of instructors would allow for the use of HLM analysis to help tease out and verify this information.

3. The student population used in this study may have influenced the results of the study. The student population at the university studied was relatively homogeneous. An expanded study of this nature conducted at other campuses with greater diversity among students may yield new insights in student outcomes resulting from grade variances in multisection courses.

4. It was beyond the scope of this research study to critically examine the individual characteristics of the professors who taught multisection calculus courses. A refinement of and/or additional instructor variables used in this study could be made in an effort to further examine course grading variances at the extremities in multisection courses. Further research in these areas could lead to useful insights as to why the extreme course grade variances found in this study exist.

Conclusion

Erickson, Peters, and Strommer (2006) surmise that a great deal of the anxiety that first-year students face stems from grading. The goal of this study from the onset was to discover the reasons behind the grade variances seen in multisection Calculus I courses in an effort to uncover ways to improve success rates among college students. This study observed that the primary cause of the observed variances in Calculus I grades appeared to be related to individual student-centric characteristics. The results of this study should promote consideration of policies and

procedures outlined in this chapter to increase the success rates of students in multisection Calculus I classes.

It is hoped that implementation of the policies outlined above will result in a reduction of the number of students who fail calculus, forcing a significant number of them to retake Calculus I, costing them both time and money. As result of doing poorly in calculus, a large number of students contend with a drop in their college GPA, likely affecting their future job prospects. As a plausible result of failing calculus, some students switch their major or opt not to continue taking additional calculus classes at the university where the study was conducted. There are no statistics on the issue, but it would not come as a surprise to learn that some of these students may have even left the university studied, in part, because of their experience taking calculus.

In an era where many institutions are faced with enrollment and budgetary challenges, it should behoove administrators to minimize the number of unsatisfied students attending their institution. Students unable to complete Calculus I may struggle to secure gainful employment in their chosen field resulting in lower alumni donations. If this happens, a university's ranking, prestige, and incoming tuition dollars would suffer as well. Given that so much of grade variance seen in Calculus I classes was explained by student-related characteristics, this study supports the idea that the key for universities seeking to minimize the number of students unable to complete Calculus I is to have a better understanding of students and then to develop tailored solutions to help them succeed.

REFERENCES

Abell, B. (Ed.). (1967, September). More pass-fail grading to ease the perils of curiosity. *Caltech News*, *1*(4), 1–6. Retrieved from http://caltechcampuspubs.library.caltech.edu/2290/1/1967_09_01_04.pdf

Abell, E. L. (1928). Grading from the median. *Journal of Educational Research*, *18*, 425–428.

Adams, J. S. (1963). Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, *67*(5), 422–436. doi: 10.1037/h0040968

Albright, P. B. (1950). Objective and subjective examinations: An answer to F. L. Marcuse. *School & Society*, *72*(2), 96.

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House*, *78*(5), 218–223. doi: 10.3200/TCHS.78.5.218-223

Anderson, D. (2012). *Hierarchical linear modeling (HLM): An introduction to key concepts within cross-sectional and growth modeling frameworks* (Technical Report #1308). Eugene, OR: Behavioral Research and Teaching, University of Oregon. Retrieved from https://files.eric.ed.gov/fulltext/ED545279.pdf

Armstrong, W. B. (2000). The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community College Journal of Research and Practice*, *24*(8), 681–695. doi: 10.1080/10668920050140837

Astin, A. W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York, NY: American Council on Education/Macmillan.

Austin, S., & McCann, R. (1992). “Here’s another arbitrary grade for your collection”: A statewide study of grading policies. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA. Retrieved from https://files.eric.ed.gov/fulltext/ED343944.pdf

Bar, T., Kadiyali, V., & Zussman, A. (2009, Summer). Grade information and grade inflation: The Cornell experiment. *The Journal of Economic Perspectives*, *23*(3), 93–108. Retrieved from http://www.jstor.org/stable/27740542

Barker, R. L. (2003). *The social work dictionary* (5th ed.). Washington, DC: NASW Press.

Barnes, L. L. B., Bull, K. S., Campbell, N. J., & Perry, K. M. (2001). Effects of academic discipline and teaching goals in predicting grading beliefs among undergraduate teaching faculty. *Research in Higher Education*, *42*(4), 455–467. doi: 10.1023/A:1011006909774

Becker, H. S., Geer, B., & Hughes, E. C. (1995). *Making the grade: The academic side of college life*. Piscataway, NJ: Transaction.

Brighouse, H. (2008, December 5). Grading medical students (and more on grade inflation). *Out of the Crooked Timber of Humanity, No Straight Thing Was Ever Made*. Retrieved from http://crookedtimber.org/2008/12/05/grading-medical-students-and-more-on-grade-inflation/#more-8718

Bristed, C. A. (1963). *Five years in an English university*. London, United Kingdom: Sampson, Low, Marston, Low & Searle.

Broome, E. C. (1945). Marks, marks, marks! *School & Society*, *62*, 76.

Butcher, K. F., McEwan, P. J., & Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. *The Journal of Economic Perspectives*, *28*(3), 189–204. doi: 10.1257/jep.28.3.189

Campbell, A. L. (1921, September). Keeping the score. *The School Review*, *29*(7), 510–519. doi: 10.1086/437427

Chansky, N. M. (1964). A note on the grade point average in research. *Educational and Psychological Measurement*, *24*(1), 95–99. doi: 10.1177/001316446402400109

Chizmar, J. F. (2000, Spring). A discrete-time hazard analysis of the role of gender in persistence in the economics major. *The Journal of Economic Education*, *31*(2), 107–118. doi: 10.2307/1183183

Chory-Assad, R. M. (2002). Classroom justice: Perceptions of fairness as a predictor of student motivation, learning, and aggression. *Communication Quarterly*, *50*(1), 58–77. doi: 10.1080/01463370209385646

Chowning, K., & Campbell, N. J. (2009, November). Development and validation of a measure of academic entitlement: Individual differences in students' externalized responsibility and entitled expectations. *Journal of Educational Psychology*, *101*(4), 982–997. doi: 10.1037/a0016351

Choy, S. P. (2002). Students whose parents did not go to college: Postsecondary access, persistence, and attainment. *PsycEXTRA Dataset*. doi: 10.1037/e492182006-021

Ciani, K. D., Summers, J. J., & Easter, M. A. (2008, December). Gender differences in academic entitlement among college students. *Journal of Genetic Psychology*, *169*(4), 332–344. doi: 10.3200/GNTP.169.4.332-344

Cohn, E., Cohn, S., Balch, D. C., & Bradley, Jr., J. (2004, December). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. *Economics of Education Review*, *23*(6), 577–586. doi: 10.1016/j.econedurev.2004.01.001

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*, 171–200. doi: 10.1146/annurev.psych.51.1.171

Crawford, A. B. (1930). Rubber micrometers. *School & Society*, *32*(2), 233–240.

Credé, M., Roch, S. G., Kieszczynka, U. M. (2010, June 1). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, *80*(2), 272–295. doi: 10.3102/0034654310362998

Crew, H. (1930, February). Annual meeting: Address of President Crew. *Bulletin of the American Association of University Professors (1915–1955)*, *16*(2), 103–111. Retrieved from http://www.jstor.org/stable/40218210

Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. K. (2003, September). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology*, *85*(3), 507–516. doi: 10.1037/0022-3514.85.3.507

Dearborn, W.F. (1910). School and university grades. *Bulletin of The University of Wisconsin*, No. 368: High School Series, No. 9.

DeBoer, B. V., Anderson, D. M., & Elfessi, A. M. (2007). Grading styles and instructor attitudes. *College Teaching*, *55*(2), 57–64. doi: 10.3200/CTCH.55.2.57-64

Department of Statistics and Data Sciences at The University of Texas at Austin. (2015, March). Multilevel modeling tutorial: Using SAS, Stata, HLM, R, SPSS, and Mplus. Retrieved from https://stat.utexas.edu/images/SSC/documents/SoftwareTutorials/MultilevelModeling.pdf

Dillon, S. (2005, January 18). Harvard chief defends his talk on women. *The New York Times*. Retrieved from https://www.nytimes.com/2005/01/18/us/harvard-chief-defends-his-talk-on-women.html

Durm, M. W. (1993). An A is not an A is not an A: A history of grading. *Educational Forum*, *57*(3), 294–297. doi: 10.1080/00131729309335429

Eells, W. (1930). An improvement in the theoretical basis of five point grading systems based on the normal probability curve. *Journal of Educational Psychology*, *21*, 128–135.

Ekstrom, R. B., & Villegas, A. M. (1994). College grades: An exploratory study of policies and practices (College Board Report No. 94-1). New York, NY: College Board Publications.

Ellis, R. S. (1926). The correction of constant errors in college marks. *School & Society*, *24*, 432–436.

Erickson, B. L., Peters, C. B., & Strommer, D. W.. (2006). *Teaching first-year college students* (Revised and expanded edition of *Teaching college freshman*). San Francisco, CA: Jossey-Bass.

Filene, P. (2005). *The joy of teaching: A practical guide for new college instructors.* Chapel Hill: University of North Carolina Press.

Finkelstein, I. E. (1913). *The marking system in theory and practice*. Baltimore, MD: Warwick & York.

Fossey, R. (2007, December). University oversight of professors' teaching activities: A professor's academic freedom does not mean freedom from institutional regulation. *Journal of Personnel Evaluation in Education*, *19*(3–4), 159–173. doi: 10.1007/s11092-007-9043-6

Freyd , M. (1923). The graphic rating scale. *Journal of Educational Psychology*, *14*(2), 83–102. doi: 10.1037/h0074329

Friedlander, J. (1980, January). Are college support programs and services reaching high-risk students? *Journal of College Student Personnel*, *21*(1), 23–28.

Friedman, L. (1989, Summer). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, *59*(2), 185–213. doi: 10.2307/1170414

Gaw, E. A. (1926). College grades. *School & Society*, *24*, 648–651.

Geisinger, K. F., Wilson, A. N., & Naumann, J. J. (1980). A construct validation of faculty orientations toward grading: Comparative data from three institutions. *Educational & Psychological Measurement*, *40*(2), 413–417. doi: 10.1177/001316448004000219

Giles, B. (1926). *Latin and Greek in college entrance and college graduation requirements*. Washington, DC: Catholic University of America.

Good, W. R. (1945, December). Should school marks be abolished? *The Education Digest 11*(4), 11.

Gouveia, M. (2004, September). Running wild: College students in the 1920s [Web log post]. Retrieved from http://www.sarahbaker.org/flapperjane/September%2004/running_wild.htm

Gray, E. (2015, October 1). Bubble trouble for standardized testing. *Time*, *186*(14), 44–49.

Greenberg, M. (2004, June 18). How the GI Bill changed higher education. *The Chronicle of Higher Education*, *50*(41), B9.

Guskey, T. R. (2013, September). The case against percentage grades. *Educational Leadership*, *71*(1), 68–72.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin Press.

Gutek, G. L. (1986). *Education in the United States: An historical perspective*. Englewood Cliffs, NJ: Prentice College Division.

Hadley, S. T. (1954). School mark: Fact or fancy? *Educational Administration and Supervision*, *40*, 305–312.

Hannon, B. (2016, November–December). General and non-general intelligence factors simultaneously influence SAT, SAT-V, and SAT-M performance. *Intelligence*, *59*, 51–63. doi: 10.1016/j.intell.2016.07.002

Hammons, J. O., & Barnsley, J. R. (1992). Everything you need to know about developing a grading plan for your course (well, almost). *Journal on Excellence in College Teaching, 3*, 51–68.

Habre, S., & Abboud, M. (2006, December). Students' conceptual understanding of a function and its derivative in an experimental calculus course. *The Journal of Mathematical Behavior*, *25*(1), 57–72. doi: 10.1016/j.jmathb.2005.11.004

Hillbrand, E. K. (1931). College marks, and remarks. *School & Society*, *33*, 725–729.

Hills, J. R. (1972). Consistent college grading standards through equating. *Educational and Psychological Measurement*, *32*, 137–146.

Hobbs, J. N. (1974). Grading versus teaching. *Improving College and University Teaching*, *4*, 239.

Hoffman, J. L., & Lowitzki, K. E. (2005, Summer). Predicting college success with high school grades and test scores: Limitations for minority students. *The Review of Higher Education*, *28*(4), 455–474. doi: 10.1353/rhe.2005.0042

Hostetler, J. M. (1989, December). Effects of the Great Depression on private higher education: Impact on private college and university planning (Doctoral dissertation, Western Michigan University). Retrieved from http://scholarworks.wmich.edu/dissertations/2113

Hu, S. (2005, May). *Beyond grade inflation: Grading problems in higher education*. San Francisco, CA: Jossey-Bass.

Hunt, L. H. (Ed.). (2008). *Grade inflation: Academic standards in higher education*. Albany: State University of New York Press.

"In the present practice": Report of the Committee on Raising the Standard, Harvard University, January 16, 1894, UAIII 5.15, Harvard University Archives.

Juola, A. E. (1976, April). *Grade inflation in higher education: What can or should we do?* Paper presented at the Annual Meeting of National Council on Measurement in Education, San Francisco, CA. Retrieved from https://files.eric.ed.gov/fulltext/ED129917.pdf

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). Hoboken, NJ: Wiley.

Kirschenbaum, H., Napier, R., & Simon, S. B. (1971). *Wad-ja-get? The grading game in American education*. New York, NY: Hart Publishing.

Kohn, A. (2004). *What does it mean to be well educated? And more essays on standards, grading, and other follies*. Boston, MA: Beacon Press.

Kohn, A. (2012). The case against grades. *Education Digest*, *77*(5), 8–16.

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the academic entitlement questionnaire. *Measurement and Evaluation in Counseling and Development*, *44*(2), 105–129. doi: 10.1177/0748175611400292

Korey, R. A. (1926). The question of marks. *School & Society*, *24*, 175–176.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008, September). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, *79*(5), 540–563. doi: 10.1353/jhe.0.0019

MacDonell, K., & Wylie, P. (2014). *Score! Data-driven success for your advancement team*. Washington, DC: Council for Advancement and Support of Education.

Marcuse, F. L. (1950, August 26). Objective and subjective examinations. *School & Society*, *72*, 136-137.

Marx, J., & Meeler, D. (2013). "Strike four! Do-over policies institutionalize GPA distortion." *Quality Assurance in Education: An International Perspective*, *21*(1), 39–53.

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Mayer, M. (1993). *Robert Maynard Hutchins: A memoir*. Berkeley: University of California Press.

McCarey, L. (Producer & Director). (1945). The Bells of St. Mary's [Motion picture]. United States: Rainbow Productions.

McCoach, D. B. (n.d.). HLM: A gentle introduction. Retrieved from http://assessment.education.uconn.edu/assessment/assets/File/HLM_Intro.pdf

Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Glendale, CA: Pyrczak Publishing.

Meyer, M. (1908). The grading of students. *Science*, *28*(712), 243.

Meyer, M. I. (1911). Experiences with the grading system of the University of Missouri. *Science*, *33*, 661–667.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2017). *Applied multivariate research: Design and interpretation* (3rd ed.). Los Angeles, CA: Sage.

Middleton, W. C. (1933). Some general trends in grading procedure. *Education,* 545–510.

Miller, G. (2014). Grade inflation, gatekeeping, and social work education: Ethics and perils. *Journal of Social Work Values and Ethics*, *11*(1), 12–22.

Milton, O., Pollio, H. R., & Eison, J. A. (1986). *Making sense of college grades: Why the grading system does not work and what can be done about it*. San Francisco, CA: Jossey-Bass.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999, June). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355–371.

National Education Association (NEA) of the United States. (1974). *What research says to the teacher: Evaluation and reporting of student achievement*. Washington, DC: Author.

Nesbit, P. L., & Burton, S. (2006). Student justice perceptions following assignment feedback. *Assessment & Evaluation in Higher Education*, *31*(6), 655–670. doi: 10.1080/02602930600760868

Noble, J., & Sawyer, R. (2002, August). *Predicting different levels of academic success in college using high school GPA and ACT composite score* (ACT Research Report Series). Retrieved from https://files.eric.ed.gov/fulltext/ED469746.pdf

Odell, C. W. (1928). *Traditional examinations and new-type tests*. New York, NY: The Century.

Ohland, M. W., Yuhasz, A. G., & Sill, B. L. (2004, July). Identifying and removing a calculus prerequisite as a bottleneck in Clemson's general engineering curriculum. *Journal of Engineering Education*, *93*(3), 253–257. doi: 10.1002/j.2168-9830.2004.tb00812.x

Ohlson, D. (1927, February). School marks vs. intelligence rating. *Educational Administration and Supervision*, *13*, 90–102.

Pattison, E., Grodsky, E., & Muller, C. (2013, June 1). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259–265. doi: 10.3102/0013189X13481382

Pease, K., & Beardsley, S. W. (1950). Objective and subjective examinations: An answer to F. L. Marcuse. *School & Society*, *72*, 294–296.

Pippin, T. (2014, June). Roundtable on pedagogy: Response: Renounce grading? *Journal of the American Academy of Religion*, *82*(2), 348–355. doi: 10.1093/jaarel/lfu002

Princeton Review, The. (n.d.). College Profile: Illinois State University. Retrieved from https://www.princetonreview.com/college/illinois-state-university-1023608

Rask, K. (2010, December). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, *29*(6), 892–900. doi: 10.1016/j.econedurev.2010.06.013

Rask, K., & Tiefenthaler, J. (2008, December). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, *27*(6), 676–687. doi: 10.1016/j.econedurev.2007.09.010

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Reeder, C. W. (1934, September 22). Should increases in grades follow increases in intelligence test scores? *School & Society*, *40*, 399–400.

Rodman, M. (2013, April 13). Making the grades. *The Yale Herald*. Retrieved from http://yaleherald.com/news-and-features/making-the-grades/

Rojstaczer, S., & Healy, C. (2012, July). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, *114*(7).

Rosovsky, H., & Hartley, M. (2002). *Evaluation and the academy: Are we doing the right thing?: Grade inflation and letters of recommendation*. Cambridge, MA: American Academy of Arts and Sciences.

Ruch, G. M. (1929). The objective or new-type examination: An introduction to educational measurement. Chicago, IL: Scott, Foresman and Company.

Rugg, H. O. (1918, May). Teachers' marks and the reconstruction of the marking systems. *The Elementary School Journal*, *18*(9), 701–719. Retrieved from https://www.jstor.org/stable/994119

Sabot, R., & Wakeman-Linn, J. (1991, Winter). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*(1), 159–170. doi: 10.1257/jep.5.1.159

Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies*, *46*(2), 201–224.

Schutz, K. R., Drake, B. M., & Lessner, J. (2013, November). Do community college full-time and adjunct faculties differ in their perceptions of rigor in assigning grades? *American Journal of Educational Studies*, *6*(2), 59–79.

Segel, D. (1934). Prediction of success in college: A handbook for administrators and investigators concerned with the problems of college admittance or guidance of college students. *Bulletin (Issue 15) of the U.S. Department of the Interior, Office of Education.*

Segel, D. (1936). To mark or not to mark, an unsolved problem. *School Life*, 2234.

Seymour, E. (2001, December 19). Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology. *Science Education*, *86*(1), 79–105.

Smallwood, M. L. (1935). *An historical study of examinations and grading systems in early American universities: A critical study of the original records of Harvard, William and Mary, Yale, Mount Holyoke, and Michigan from their founding to 1900*. Cambridge, MA: Harvard University Press.

Sonner, B. S. (2000). A is for "adjunct": Examining grade inflation in higher education. *Journal of Education for Business*, *76*, 5–8. doi: 10.1080/08832320009599042

Starch, D., & Elliott, E. C. (1912, September). Reliability of the grading of high-school work in English. *The School Review*, *20*(7), 442–457. Retrieved from https://www.jstor.org/stable/1076706

Starch, D., & Elliott, E. C. (1913, April). Reliability of grading work in mathematics. *The School Review*, *21*(4), 254–259. Retrieved from https://www.jstor.org/stable/1076246

Stevens, J. (2007). Hierarchical Linear Models. Retrieved from http://pages.uoregon.edu/stevensj/workshops/hlmslides.pdf

Stiles, E. (1901). *The Literary Diary of Ezra Stiles: 1782–1795* (*Vol. 3:January 1,1782–May 6,1795*). New York, NY: Charles Scribner's Sons.

Su, A. Y. (2007, Winter). The impact of individual ability, favorable team member scores, and student perception of course importance on student preference of team-based learning and grading methods. *Adolescence*, *42*(168), 805–826.

Sung, E. (2003, October). Mending the federal circuit split on the First Amendment right of public university professors to assign grades. *New York University Law Review*, *78*(4), 1550–1580.

Thelin, J. R. (2004). A history of American higher education. Baltimore, MD: The Johns Hopkins University Press.

Travers, R. W., & Gronlund, N. E. (1950). Meaning of marks. *Journal of Higher Education*, *21*, 369–374.

Urda, J., & Ramocki, S. P. (2015). Assessing students' performance by measured patterns of perceived strengths: Does preference make a difference? *Assessment and Evaluation in Higher Education*, *40*(1), 33–44.

Vallade, J. I., Martin, M. M., & Weber, K. (2014). Academic entitlement, grade orientation, and classroom justice as predictors of instructional beliefs and learning outcomes. *Communication Quarterly*, *62*(5), 497–517. doi: 10.1080/01463373.2014.949386

Van Ness, B. F., Van Ness, R. A., & Kamery, R. H. (1999). The effect of part-time instruction on grades in principles of finance. *Financial Practice & Education*, *9*(2), 105–110.

Vogt, W. P., Vogt, E. R., Gardner, D. C., & Haeffele, L. M. (2014). *Selecting the right analyses for your data: Quantitative, qualitative, and mixed methods*. New York, NY: The Guilford Press.

Volkwein, J. F., & Yin, A. C. (2010, Autumn). Measurement issues in assessment. *New Directions for Institutional Research*, *2010*(S1), 141–154.

Wahlquist, J. T. (1940). Is the IQ controversy philosophical? *School & Society*, *52*, 539–547.

Walvoord, B. E., & Anderson, V. J. (2010). *Effective grading: A tool for learning and assessment in college*. San Francisco, CA: Jossey-Bass.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Wilson, R., Fain, P., Fogg, P., & Selingo, J. (2006, March 3). The power of professors. *The Chronicle of Higher Education*, *52*(26), A10.

Younts, C. W., & Mueller, C. W. (2001, February). Justice processes: Specifying the mediating role of perceptions of distributive justice. *American Sociological Review*, *66*(1), 125–145. doi: 10.2307/2657396