2023

Making the grade: Do teacher-created grading systems result in disparate outcomes for high school students?

Laura L. Mayberry
University of North Florida, laura.mayberry.island@gmail.com

Making the Grade: Do Teacher-Created Grading Systems Result in Disparate Outcomes for High School Students?

by

Laura L. Mayberry

A Dissertation submitted to the Department of Leadership,

School Counseling & Sport Management

in partial fulfillment of the requirements for the degree of

Doctor of Education

UNIVERSITY OF NORTH FLORIDA

COLLEGE OF EDUCATION AND HUMAN SERVICES

May 2023

This proposal or dissertation titled Making the Grade: Do Teacher Created Grading Systems Result in Disparate Outcomes for High School Students?

_____ _____

Dr. Daniel Dinsmore, Committee Chair

_____ _____

Dr. Diane Yendol-Hoppey, Committee Member 1

_____ _____

Dr. Rudy Jamison, Committee Member 2

_____ _____

Dr. Amanda Kulp, Committee Member 3

Dedication

For Pat Gallaher, who should have been first.

Acknowledgments

I would like to thank my professors and committee members for all of their help throughout this process. Their guidance and feedback made this an invaluable learning experience for me. I would also like to thank my family, friends, and coworkers for their support. Their check-ins and words of encouragement reminded me of the importance of this work. Finally, I would like to thank the members of my cohort, the Divine 29. It has been a privileged to work with such an amazing group of educators.

Table of Contents

Abstract

The lack of formal grading policies within individual schools and school districts results in differential grading across classrooms. In many secondary schools, an individual course may be taught by multiple teachers. This study aimed to determine if the existence of differential grading results in disparate outcomes for students based on their assigned teacher. Data about students (demographics, teacher-issued grades, and standardized test scores) was collected from four core content classes at one high school. Regression analysis was used to isolate the impact of students' assigned teacher on their course grade after controlling for demographics and academic performance as measured by the standardized test associated with each subject. The findings showed that the *teacher* factor was significant for all four courses. Given the importance of grades to students' academic trajectories, these findings should be of concern to all stakeholders, from students and parents to classroom teachers and school administrators.

Chapter 1: Introduction

Differential grading "occurs when students in courses with the same content and curriculum receive inconsistent grades across teachers, schools, or districts," (Rauschenberg, 2014, p. 3). Given the important role that grades play in students' educational trajectories, the existence of differential grading may be cause for concern. These are not isolated incidents, rather there appears to be a systematic trend regarding differential grading procedures. According to Guskey and Link (2018), multiple studies have shown that the criteria used to grade students can vary widely across teachers, making them unreliable measures of student performance. Similarly, Rauschenberg (2014) argued that high school teachers are given significant leeway in determining their grading practices, even with the existence of school-level grading policies. For example, a school or department may have a policy that tests must count for a certain percentage of students' grades, but individual teachers may have different policies about test corrections, retakes, or even their determination of what qualifies as a test. Finally, Carifio and Carey (2009) discovered that teachers were using far more factors than expected when determining grades. The greater the number of factors being considered, the greater the likelihood that there will be inconsistencies across classrooms.

One potential reason for the existence of differential grading may be the variety of beliefs about what grades should mean and convey. Carifio and Carey (2009) referred to two different groups of teachers when it comes to beliefs about grading practices. Some teachers believe that grades exist to help guide students as they learn, while others believe that grades serve a sorting

function. The sorting function refers to when students are ranked to determine access to future courses and other school-related activities. These competing beliefs - aiding versus sorting - result in a hodgepodge grade that may represent numerous factors, including, but not limited to, attitude, effort, and achievement (Brookhart, 1991). In other words, a grade may represent multiple constructs rather than one unitary construct such as academic achievement.

These differing beliefs mean that students of similar ability may have entirely different grade outcomes depending on which teacher they have for a particular course. Carifio and Carey (2009) noted that a uniform grading policy is lacking in many schools. According to the authors, this may result in significant differences from teacher to teacher and possibly from student to student within the same teacher's classroom. These variations can result in disparate educational outcomes for students even after that particular course is over.

Problem Statement

The dual-purpose and potential conflict of using grades to measure student achievement while also functioning as a gatekeeper for subsequent opportunities lead to issues with what Carifio and Carey (2009) refer to as a social contract. While teachers may view grades as a means of internal communication with their students, the world outside of their classroom often lays claim on that information to make decisions that impact students beyond the classroom. Scholarships, college applications, and job applications may all ask for a student's GPA to compare them to other candidates for the same reward or position. Given this conflict and the potential harm to students due to inequitable policies, it is important to examine the underlying

causes of differential grading and to gather data to examine whether and how students are impacted.

Purpose Statement

The purpose of this study is to analyze the effects of differential grading on students' course grades. For this analysis, a quantitative design will be used. I will analyze high school student course grades and test score data from the 2018-2019 school year to determine if differences exist in average quarterly grades across multiple teachers of the same course after controlling for demographic characteristics and academic achievement (as measured by standardized test scores in those subjects).

Researcher Positionality

While some quantitative studies have shown that different subgroups of students are disproportionately affected by grading policies that do not align with standardized test results (Angelo & Balcão Reis, 2021; Beswick et al., 2005; Martinez et al., 2009; Thorsen & Cliffordson, 2012), differences across teachers of the same course have not been explored in any significant way. The goal of this study is to explore whether differential grading results in meaningful differences in outcomes for students between their teacher-assigned grades and standardized test scores. This will provide teachers, school-based administrators, and district-level administrators with insights into the potential negative impacts that differential grading may have on students. This is especially important if multiple teachers teach the same course and students are randomly assigned to those teachers.

I spent my last few years in the classroom revamping my grading practices in an attempt to have my students' grades more closely reflect their academic achievement (as opposed to other factors such as attendance, organization, etc.). I converted to a standards-based grading system and believe that it made a difference in how my students viewed their work. A standards-based grading system relies solely on summative assessments that are tied to course standards. Formative assessments, such as homework, classwork, and quizzes, are not counted as part of a student's grade because those represent learning *in process*, as opposed to the final product. After becoming an assistant principal, I began to look more closely at the grading practices of the teachers at my school. I saw differences in practices from teacher to teacher, including those teaching the same course. I also saw students whose grades did not always correlate with their test scores (both teacher-made tests and standardized tests). Given that students' grades affect their future academic trajectory (e.g., access to higher-level courses the next year, college admissions, scholarships), I think it is important to understand how teachers determine their grading practices and if these practices 1) accurately measure students' achievement, and 2) are equitable across all students and courses.

Assumptions

The main underlying assumption made in this study is that differential grading exists and that differential grading is present in the school site chosen. The district and school site do not have any official grading policies, but teachers within a department or school may work together on common grading policies. While some level of coordination may have occurred with the

teachers and courses in this data set, I believe that it is safe to assume that even coordinated, written policies may be enacted differently across classrooms.

Research Questions

The following questions will be addressed for 4 courses (Biology, U.S. History, English 2, and Geometry) from the 2018-2019 school year.

1) How do teacher-issued course grades relate to students' scores on standardized tests within the same course?
2) How do students' demographic characteristics (gender, race/ethnicity, ELL status, and ESE status) and what teacher they were assigned to affect the difference between students' course grades and their standardized test scores?

Overview of Theoretical Framework

If differential grading results in disparate outcomes for students, this should be a cause for concern among students, parents, teachers, and administrators. I plan to look at this problem of practice through the lens of street-level bureaucracy (Lipsky & Weatherley, 1977). This framework focuses on how ground-level decisions about policy implementation can involve a large amount of discretion because those responsible for completing complex and widely-variable tasks cannot possibly develop rules or policies for every possible situation that may arise (Sattin-Bajaj, Jennings, Corocran, Baker-Smith, & Hiley, 2018). Barberis and Buchowicz (2015) noted that discretion is not necessarily positive or negative and does not imply illegitimacy. They go on to explain that this perspective is a way to "look into *teachers' coping*

strategies, in classroom management, in curricular choices, and in teachers' decisions on curriculum, maneuvering between external pressures and their own views" (p. 63). Without any guidance from above or collaboration among peers, these *coping strategies* may result in a wide variety of grading policies and varied implementation of shared policies.

Teachers are expected to issue grades to their students with little to no guidance on how these grades should be determined. This leaves them in the position of developing their policies based on past experience, the policies of their peers, and the pressure to conform to the expectations of their students, their students' parents, and other stakeholders. Proitz (2013) notes that teachers may view recommended practices as irrelevant, especially when they are asked to grade students based on performance and not on non-academic factors that can be used as motivators for struggling students. He also points out that policy-makers often support contradictory policies and leave it to teachers to figure out how to implement them.

In addition to the lack of guidance from above, teachers must also deal with the issue of efficiency when it comes to grading. Secondary school teachers can have more than 100 students on their caseload. Given the limited time with which teachers have to develop and grade assignments, it can be easy to rely on certain stereotypes to drive grading practices. According to Harrits (2018), "there is a consensus that stereotypes function as a tool to reduce the uncertainty of the information and situations characterizing street-level work (p. 94). For example, a teacher may award a daily grade for participation as an efficient way to deal with classroom management issues. However, their particular definition of participation may favor

some students over others. A student who calls out in class or does not speak at all may have the same level of academic knowledge as their peer who sits quietly and raises their hand to answer a question. In the context of my research, I believe that teachers developing their grading systems within the loosely defined guidelines and expectations of most school systems may create policies that do not serve the best interests of their students. If teachers' policies are even partially driven by stereotypes about different subpopulations of students, then outcomes for those students will not be equitable. Barberis and Buchowicz (2015) point out that "educational professionals can serve as gatekeepers for the access to further educational opportunities (also) through their discretionary power… including personal theories of justice and fairness" (p. 63). My problem of practice centers around the idea that when grading policies are left up to the discretion of the individual classroom teacher, certain subgroups of students may be disproportionately harmed. The reverse may also be true.

In their study about the implementation of policies related to students with disabilities at a university in the United Kingdom, Wray and Houghton (2019) discussed how university professors balance the needs of outside professional organizations (related to the employment of graduates from the program) with the requirements to make accommodations for disabled students. This is similar to the demands placed on high school by post-secondary institutions to provide grades that rank students to assist in their admissions process.

Overview of Methodology

This study will utilize multiple linear regression analysis to discover the impact of teacher assignment on students' course grades.

Significance of the Research

In the absence of district or school-wide grading policies, teachers utilize a large amount of discretion when developing and enforcing their grading policies. If this level of discretion results in similar students receiving significantly different grades due to the random assignment of students to teachers, then it may be time to re-assess the level of autonomy given to teachers in the area of grading. This does not imply that a top-down approach must be taken. While professional learning communities are growing in popularity in secondary schools, their focus tends to be on the curriculum and student assessment. Introducing differential grading into the conversation could serve as a catalyst for change. Without district or school-level policies in place, which might be resisted by teachers as top-down mandates, grading policy congruence will need to be an organic, teacher-led process.

Organization of the Study

Chapter 2 of this dissertation contains a review of the literature relevant to the topic of differential grading. It begins with an overview of past research on the existence of differential grading in secondary schools and colleges and the conceptual framework that guides this study. It also includes a brief section on the sources of grading policy development, including teachers' own experiences in school and official and unofficial on-the-job training. Finally, I cover the

potential impacts of differential grading on students. Chapter 3 addresses my research methodology. My results are presented in Chapter 4 and Chapter 5 contains a discussion of my findings and recommendations for future research.

Chapter Summary

The goal of this study is to determine if differential grading results in disparate outcomes for students when more than one teacher teaches the same course. Previous research has focused on different outcomes based on factors such as students' demographics or ability levels (as measured by standardized test scores). The results of this study may add *teacher* to this list of factors.

Chapter 2: Review of Literature

Grading in the Era of Accountability

When the National Commission on Excellence in Education issued its report *A Nation at Risk* in 1983, it signaled the beginning of a reform movement that swept through the United States education system. This movement led to changes in curriculum and assessment, as well as increased emphasis on cross-curricular skills such as creativity and critical thinking (Lehman et al., 2018). Later, with the passage of the No Child Left Behind Act of 2001, states were forced to develop accountability programs to measure students' academic performance (Swan et al., 2014). In response to this legislation, states developed report cards to assess schools and districts on items such as standardized test scores and graduation rates. Teachers were also held accountable for their students' performance. In 2011, Florida passed a law that created merit pay for teachers based on a value-added model designed to measure their impact on student performance. While the performance of all parties involved (students, teachers, schools, and districts) came under stricter scrutiny, uniform policies were not developed at the district or school level for how to grade students in this new era of accountability (Lehman et al., 2018; Pollio & Hochbein, 2015; Swan et al., 2014).

Pollio & Hochbein (2015) suggested that the increased emphasis on measuring student performance against a set of state or national standards requires grading practices that reflect student mastery of course standards. The lack of uniform and well-defined grading systems that clearly link teacher-issued grades to student progress against these standards means students do

not have an accurate record of their learning and teachers do not have the information necessary to drive their lessons. To compound this issue, there is no agreement in the education community about what comprises a *valid* grade. According to Welsh et al. (2013), these new reporting requirements from the federal government can serve as a lever to change instruction. If teachers make the shift to focusing more on their students' progress in relation to the standards that will be tested (as opposed to behavioral attributes or curriculum that is not part of the standards), their lesson plans will have to adapt. This shift may also result in new grading policies, as it will become more difficult for teachers to justify placing value on things that are not explicitly taught or emphasized in class. While standards-based grading practices have been more common at the elementary school level, the accountability era has led to the development of standards-based grading practices and report cards at the high school level as well (Townsley & Varga, 2018).

Despite the trend towards standardization in many areas of education (e.g., Common Core Standards, curriculum, and high-stakes testing), grading policies are often left up to individual teachers (Alm & Colnerud, 2015; Guskey & Link, 2018; Kunnath, 2017; Proitz, 2013; Zoeckler, 2007). These policies can vary widely from teacher to teacher, school to school, district to district, and state to state. Because teachers' grading practices have a direct impact on student grade point averages, and thus their class rank, inequities can arise when it comes to students' access to college admissions and scholarships. More immediately, different grading practices may have an impact on student performance and self-efficacy from class to class. Teachers may be using flawed or biased reasoning to develop their grading policies and practices. This should

concern everyone impacted by these grading policies, including students, parents, teachers, school and district-level administrators, and college admissions departments. This review of the literature will explore the existence of these variations in grading policies in more depth, as well as examine the potential sources of these variations.

Existence of Grading Policy Variation

While differential grading exists at all grade levels and across disciplines, I have chosen to focus my research on its occurrence at the high school level. Guskey and Link (2018) hypothesized that variance in grading practices would be greater at the higher grade levels because older students were more capable of displaying their knowledge in a wider variety of formats. They suggested that secondary teachers may rely more on homework because they believe that older students can and should learn or practice some course content independently. The existence of policy variation takes several forms: variations between academic disciplines, between districts/schools, between individual teachers, and between subpopulations of students.

Variation Between Academic Disciplines

In addition to differences between grade levels, there has also been evidence that there are disparities between academic disciplines as well. For instance, in their survey of 442 undergraduate faculty members, Barnes et al. (2001) studied the effects of academic discipline on teaching goals and grading practices. The survey used in their study measured faculty members along two scales: 1) *frame of reference* (their beliefs about using norm-referenced versus criterion-referenced grading practices), and 2) *gatekeeping*. Norm-referenced grading

examines a student's performance in relation to the performance of their classmates, whereas criterion-referenced grading looks at a student's performance compared to a set of standards (Barnes et al., 2001). They proposed that "at least some of the between-discipline variability in grade elevations is attributable to measurable differences in beliefs faculty hold about the meaning grades should convey and how grades relate to learning," (p. 456). When teachers hold norm-referenced beliefs, they are more likely to have a system that limits the number of perfect scores; whereas teachers who hold criterion-referenced beliefs do not impose such limits, as all students may master the standard being assessed.

According to Barnes et al. (2001) gatekeeping refers to the belief that educators in higher education contribute to society by sorting students based on performance. A teacher who subscribes to this belief may view it as their role to send only the best students on to the next level within their discipline. Teachers who do not subscribe to this belief may be more focused on helping all students reach proficiency or mastery. Barnes et al. (2001) argued that "gatekeeping, but not frame of reference, scores differed significantly by academic discipline and by primary teaching role" (p. 462). For example, faculty in disciplines such as zoology, physics, and mathematics had "significantly higher mean gatekeeping scores" than faculty in disciplines such as education and fine arts (p. 464). They also noted that these differences between academic disciplines could be related to the level of paradigm development within a discipline. In disciplines where there is "greater consensus on theories and methods", grades are seen as a way to funnel access to further study (p. 464). In disciplines where there is less consensus on theories

and methods and thus greater subjectivity in performance evaluation, faculty are less likely to subscribe to the gatekeeping function of grades. Some examples include education and psychology. In these fields, *student development* was seen as more important than mastery of a well-established body of knowledge.

According to Proitz (2013), "school subjects may be regarded as the basis on which teachers construct frameworks for assessing achievement and developing grading practices (p. 557). Similar to the findings in Barnes et al. (2001), he noted that there was a divide between math and science teachers, who view their subjects as objective and thus leaned more toward criterion-referenced grading, and English teachers, who view their subject as more subjective. However, in his study of Norwegian lower and upper secondary schools, he noticed that Norwegian (language) and math teachers both tended to rely on a narrow range of evidence when calculating their final grades as opposed to science teachers, who considered a broader range of evidence. Zoeckler (2007) also argued that English teachers viewed their courses as more subjective and thus their grades included non-academic factors such as ability and effort. His case study of a high school English department did not address whether the teachers collaborated with each other in order to determine how their moral judgments about students' character and ability impacted their grades and if those impacts were the same from teacher to teacher. These findings are consistent with those of Sabot and Wakeman-Linn (1991), who posited that math and science teachers at the university level tended to be less lenient graders than their counterparts in English and the social sciences.

Physical education teachers are more likely to use personal or professional judgment in their determination of students' grades (Proitz, 2013; Svennberg et al., 2014). In their study about physical education teachers' grading criteria, Svennberg et al. (2014) hypothesized that while there was agreement across teachers in most areas, the differences were often the result of varying interpretations of the standard being graded. They argued that there were disconnects between what teachers considered to be important and what actually mattered in their grading calculations.

Variations between Districts and Schools

While the variations that occur between subjects can be due to the underlying frameworks of those academic disciplines, there is also evidence that variations occur between school districts and individual schools within districts. Moen and Tjelta (2010) studied how students from various college programs in Norway performed in an advanced college program. The application process is centralized and based solely on students' GPAs, work experience, and age. Because of this process, the expectation would be that students with similar admissions scores would perform similarly in the advanced program. The researchers discovered examples where this was not the case. In their study, students from some colleges consistently underperformed their peers despite having the same admission points. They concluded that those colleges were giving these students disproportionately higher grades than their level of knowledge would indicate. They discovered that teachers' grading practices depended on the overall quality of students at their school, with better schools (as determined by the quality of

students that they recruit from the upper secondary schools) having stricter policies. The researchers pointed to the theory of adaptation level to frame what might be causing the differences in scores across colleges. According to this theory, teachers "anchor their judgment of a specific student's performance to the performance of the other students in the class," (Moen & Tjelta, 2010, p. 223). Thus, students at schools that recruit better students will have a higher bar to meet to earn a certain grade. According to the authors, this use of norm-referenced grading results in equally qualified students from better schools receiving lower admission points than students from less rigorous schools.

Rauschenberg (2014) cited a 2009 study that compared grades with end-of-course exam scores and saw "sizable differential grading across districts" in eight subjects (including humanities, mathematics, and science) that had end-of-course exam requirements (p. 5). The study noted mismatches between grades and exam scores, with some districts seeing a lot of students earning high grades despite not earning a high exam score, while other districts saw a lot of students earning mediocre grades despite having high exam scores.

While the focus of this study is on differences at the classroom level, these studies indicate that the problem may also be occurring across schools and districts. Given that secondary students may be in competition with students from other schools and districts for a finite number of scholarships and college admissions slots, these variations in practice are also concerning. In addition, students who change schools may be impacted if these differences exist.

Variations Between Individual Teachers

In addition to variance between schools and districts, studies have shown that grading practices vary between teachers within the same schools as well. Alm and Colnerud (2015) noted that as students move between classrooms throughout the day, they may also move between different grading practices. They argued that a frequent source of unfair grading practices, as indicated by teachers describing their experiences as students, was the application of personal rules (p. 139). Another theme that emerged from the teachers' reported experiences was the overemphasis on one assessment to determine their final grade. While research has shown that there tends to be variation between academic disciplines in the weighting of achievement factors such as summative test scores as part of students' final grades, this variation can also occur between individual teachers in the absence of a department or school policy.

As an example of this, Guskey and Link (2018) provided evidence that "teachers at each grade level varied considerably in the weights they assigned to different elements" (p. 308). They noted no statistically significant correlation between teachers' years of experience and the weights that they assigned to different elements, which seems to contradict previous research on this topic. They argued that high school teachers gave more weight to "major compositions and examinations, laboratory projects and homework" and that elementary school teachers gave more weight to "formative assessments, exhibits of student work, and classroom observations" (p. 308). The authors of this study expressed concern about the use of formative assessments in determining students' final grades since they are designed to measure progress during the

learning process and should act more as an opportunity to provide feedback rather than a summative measure of achievement. For example, a student may perform poorly on a classwork assignment or a quiz but do well on the final test after receiving feedback or additional instruction. If teachers average these grades together, then the lower classwork and quiz grades will bring down the student's overall grade even though they have mastered the material by the end of the unit.

The inclusion of non-academic factors in students' final grades varies from teacher to teacher and further complicates the formula. According to Guskey and Link (2018), some researchers have argued that the inclusion of multiple factors is rooted in "teachers' belief that 'academic achievement' conceptually includes behaviors that support and promote learning… The weight attached to these behaviors in determining grades has been shown to vary greatly among teachers, however, even those with similar teaching assignments within the same school" (p. 305). Similarly, Kunnath (2017) argued that teachers attempted to balance grading based on academic achievement (outcome-based) with grading based on student work ethic (process-based).

Even when teachers share similar beliefs about grading, the end results may still vary. Zoeckler (2007) argued that all of the teachers within one high school English department believed that character development was part of their role. However, they may not have been aware of how that role impacted their individual grading practices. As a group, they acknowledged that these use students' character to determine final grades when students were on

the border between two grades, but they did not address how these judgments may impact the grading of individual assignments, and thus snowball into a larger impact on their final grade. In addition to this blind spot, there is also the concern that the judgment of character, ability, and effort may vary from teacher to teacher. The teachers in this case study expressed that it was important for students to understand the meaning being the grades they were given, but there was no evidence that they thought it was important to calibrate their practices given the subjective nature of teachers' judgment. As part of his study on grading influences and practices, Kunnath (2017) surveyed teachers and asked to what extent their grades were based on 17 different practices. *Grade distributions of other teachers* ranked second to last. Also ranking low was *formal or informal school or district policy or grading distributions* (tied for 13th out of 17). The lack of calibration between teachers forces students to adapt to different policies as they move from class to class and grade level to grade level.

Summarizing his findings, Proitz (2013) addressed the difference between teachers who used a *standardized approach* (math and science) and those who use a *continually negotiated approach* (language and arts and crafts), noting that when student performance measures are open to interpretation "negotiations between colleagues [are] important to reach an agreed-upon standard" (p. 569). If these negotiations do not result in agreed-upon standards of measurement that all teachers within a given subject will use, the result can be grades that are not comparable from teacher to teacher and from student to student.

While the impacts of differential grading on students are lessened if teachers within a subject are using similar assessments, the differences that occur between grade levels may cause harm if students are unfamiliar with the changing methods as they move from one grade level to the next. For example, students entering into Advanced Placement courses in high school (that tend to weigh tests heavily due to the nature of the courses) may be shocked to find that they cannot support their final grade with process grades such as completing homework assignments or submitting their notebooks for a check on notes and organization.

Variations Between Subpopulations of Students

Teachers often include non-performance factors (e.g., participation, effort) into their grading calculations. These factors can be used to apply what Alm and Colnerud (2015) refer to as the "principle of need" in order to help struggling students whose circumstances may impact their ability to perform at grade level (p. 136). In his study of students across North Carolina who were enrolled in Algebra I and English I, and who also took an end-of-course exam in those subjects, Rauschenberg (2014) described his findings as:

> student characteristics are stronger predictors of differential grading than teacher, school, or district characteristics. Female, Limited English Proficient, and 12th grade students earned statistically significant higher grades than other students in all subjects, holding test scores and student, teacher, and school characteristics constant. Low-income students, in contrast, earn lower grades than other students in both subjects. (pp. 5-6)

Once again, due to the "social contract of educational endeavors" referred to previously (Carifio & Carey, 2009, p. 35), the existence of different outcomes for subpopulations of students enrolled in the same courses is cause for concern. If grades are used to track students into future courses, then the lower grades for low-income students may result in a self-fulfilling prophecy that impacts their self-efficacy. Rauschenberg (2014) also observed that students with exceptionalities and LEP students received grades that were higher than their peers. While acknowledging that these higher grades may be the result of the additional services available to these students, he also stated that teachers may be compensating students by "artificially improving grades" (p. 15). While this sentiment may stem from good intentions, it may set those students up for failure in future courses if their next teacher does not have the same grading philosophy. If grades are used to make decisions about future educational opportunities (e.g., college admissions, scholarships, internships), students who receive a boost may be at an unfair advantage.

Variations Between Ability Levels of Students

There have been several studies that focused on teachers' grading practices in relation to the ability level of their students. Kunnath (2017) noted that teachers who taught upper-level courses felt more pressure to issue higher grades. However, they also believed that grades in these classes should reflect mostly academic factors to maintain the rigor of the program (e.g., Advanced Placement, Gifted and Talented Education) and to prepare students for college. Teachers believed that students in standard-level classes needed more flexibility in the grading of

individual assignments and weighted their grading categories (e.g., tests, classwork, homework) differently to include more nonacademic factors. The weighting of assessment categories for the teachers in this study ranged from 30% to 90%, although there was evidence that some PLC groups agreed on common categories and weights (p. 81).

As part of his study examining how grading practices differ across academic disciplines, Proitz (2013) observed that disciplines differed in their treatment of stronger and weaker students. Math and arts and crafts teachers were least likely to consider things other than performance. Science teachers were more likely to consider the progress of weaker students. Norwegian (language) teachers were the most likely to consider non-academic factors for weaker students. Physical education teachers reported relying on participation for stronger and weaker students.

Whether it is from student to student, classroom to classroom, subject to subject, or school to school, there is ample evidence of the existence of differential grading. The implications of these grades can be far-reaching (e.g., graduation, college admissions, scholarships), which makes their validity an important thing to consider for teachers and administrators.

Correlation Between Grades and Standardized Test Scores

At the heart of the discussion about differential grading is the question: *what constitutes a valid grade?* It can be argued that any teacher-assigned grade is valid as long as the teacher consistently applies their written policies and those policies are communicated to and understood

by their students. However, this does not take into account that grades have implications for students outside of the classroom in which they are assigned. The correlation between teacher-assigned grades and standardized test scores has been widely studied, with the underlying implication that a stronger relationship here is better. This assumes that standardized test scores are an accurate measure of students' knowledge and that teachers whose assigned grades are highly correlated with these scores are providing more valid grades.

The goal of this paper is not to take a stance on this particular issue. The difference between students' teacher-assigned grades and their standardized test scores is being used as a variable to measure the difference from teacher to teacher of the same course. If differences exist, they have implications for the students in those classrooms even if the underlying test scores are not the best measures of their knowledge.

Student Subsets

Welsh et al. (2013) reported that teachers' grades were more consistent with test scores for students in the *meets* the standards category than for the *falls far below* category. This could be because the study was done in a high-performing district, and thus teachers had more exposure to students at the *meets* and *exceeds* levels. The inconsistencies between grades and test scores for the *falls far below* category could be the result of teachers compensating for low test scores with other grading categories that measure non-academic factors.

Correlation by Course

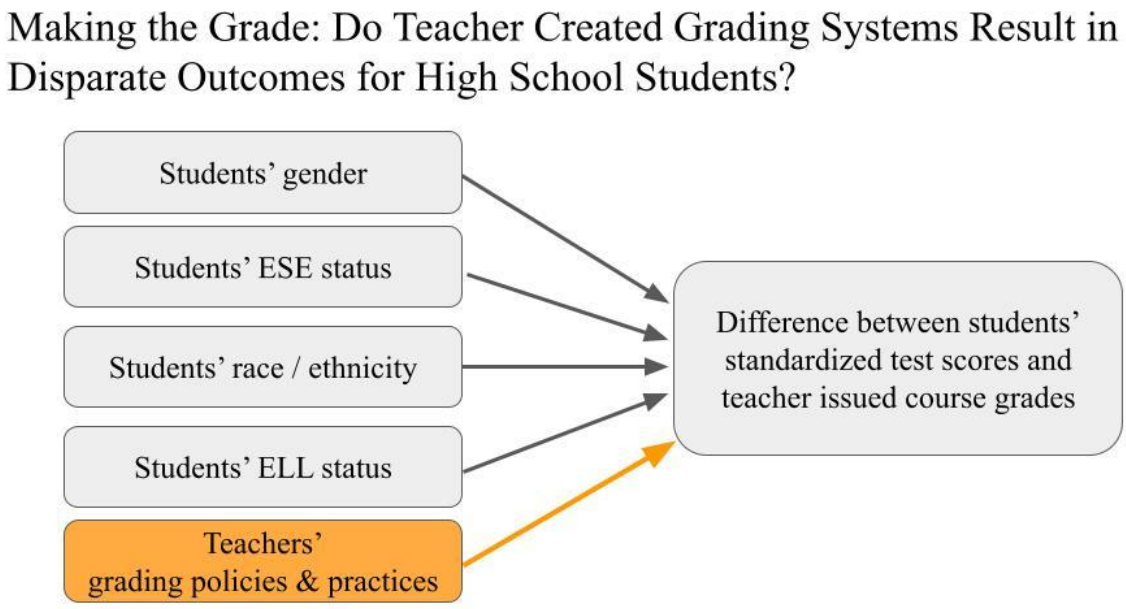
Welsh et al. (2013) discovered that "teachers generally assigned grades below state test scores in mathematics and above state test scores in reading and writing" (p. 30). The relationship between scores and grades was weakest in writing. The authors partially attributed these findings to inconsistencies in how teachers measure proficiency in the classroom versus how they are measured on standardized tests. Rauschenberg (2014) discovered a 19 percentage point gap between the percentage of students who failed the Algebra I end-of-course exam (34.6%) and the percentage who received a failing grade in the course (15.6%). He concluded that "the larger gap indicates that teachers may anchor grades around a certain distribution to some extent, regardless of ability or performance" (p. 9). This is also consistent with previous findings that teachers may compensate lower-performing students by including non-performance factors in their final grades.

Study Conceptual Framework

The existence of grading policy variation has been well documented as described in the last few sections. There has also been significant research in the area of disparate outcomes for students because of these variations. What is largely missing from the literature is a quantitative analysis of the impact of this variation at the individual teacher level. At the secondary school level, courses are often taught by multiple teachers and students are often randomly assigned to those teachers. The aim of this study is to build upon existing research and examine the impact that the teacher has on students' final course grades after controlling for demographic

characteristics and academic achievement as measured by stadardized test scores. Variables in gray (students' gender, ESE status, race/ethnicity, and ELL status) have been examined in previous research. The conceptual framework of this study (Figure 1.1) shows how the individual teacher fits into the overall model.

Figure 1.1



Implications for Theoretical Framework

Readers who are not educators may wonder why variations in grading policy exist and are allowed to continue. While there are numerous reasons why these variations exist, many fit under the umbrella of distributed decision-making that occurs when formal, systemwide policies and procedures do not exist. Viewed through the lens of street-level bureaucracy, teachers'

day-to-day actions are based on their need to make decisions that balance the need for efficiency and accuracy. These two goals can be in conflict with each other and a spectrum of outcomes can result. In Chapter 5, I will discuss the possible implications of this autonomy for students, teachers, and school administrators.

There may also be arguments for why these variations in grading policies *should be* allowed to continue. While these areas are not the focus of this study, it is important to understand the sources of grading policy development in order to understand the larger systems in which these decisions by teachers are being made.

Sources of Grading Policy Development

Differences in teachers' beliefs about grading may come from several sources, including their own experiences as students, their teacher training programs, formal on-the-job training, and their interactions with colleagues that can be considered informal on-the-job training.

Teachers' Own Experiences in K-12 School

Barnes et al. (2001) suggested that "beliefs about teaching goals and beliefs about grades are embedded in the cultures of the academic disciplines" (p. 465). This is backed up by Proitz (2013), who observed that "teachers' socialisation and professional development" is influenced by their academic discipline (p. 556). The seeds of these beliefs were planted when they were students, potentially absorbing and adopting their teachers' methods as their own. Barnes et al. (2001) suggested that if faculty were more aware of the origins of their grading philosophies,

"they would very likely find it easier to develop and stick with defensible grading schemes" (p. 465).

Teacher Training Programs

The existence of training for pre-service teachers is not universal. Guskey and Link (2018) acknowledge that "other researchers suggest these varying grading practices result at least in part from the lack of formal training teachers receive on grading and reporting" (p. 305). In their study, they expected to find differences in teachers' grading policies depending on their years of experience due to changes in teacher education programs, but they found no statistically significant correlations between years of experience and the evidence that teachers choose to assess students and calculate final grades. What complicates this matter further is the potential inclusion of both formative and summative measurements in those final grades. Citing a report by the Swedish government, Alm and Colnerud (2015) noted that teacher training programs often fall short in preparing their students for the important task of grading. They go on to suggest that this lack of training can lead to teachers using their own experiences in school as a driver of their classroom policies (p. 133).

Another possible reason that newer teachers are not implementing grading practices significantly different from their more experienced colleagues could be that while they are benefiting from the most up-to-date training in curriculum and instruction, the professors teaching those courses may still be using more traditional grading practices (Guskey & Link,

2018, p. 314). This disconnect means that pre-service teachers are not experiencing firsthand new grading and assessment practices as students and therefore will not see their value firsthand.

On-The-Job Training & Experiences

While teachers may enter the profession with ideas about how they will grade their students - originating from their own experiences as students or from the instruction they received in their training - those ideas may evolve as teachers settle into their role and are exposed to the school culture around grading and the ideas of their colleagues.

Guskey and Link (2018) point out that while newer teachers may receive training in assessment and grading practices related to the current "high accountability climate", they:

> are receiving little on-the-job guidance about how to teach and assess standards aligned to high-stakes testing and evaluations... Newer teachers also may be complying with the pre-established grading norms of their more experienced colleagues or prescribed grading policies within their school or district, thereby fostering consistency in teachers' grading practices over time. (p. 314)

Proitz (2013) found that math and science teachers were less likely to collaborate around grading as compared to Norwegian and arts and crafts teachers. Since math and science teachers were more likely to view their subject as objective, they were less likely to need the opinions or interpretations of their peers. Norwegian and arts and crafts teachers, who were more likely to use a norm-referenced approach, reported more difficulties around grading. These teachers were more likely to collaborate to calibrate their grading practices.

Kunnath (2017) found that teachers were concerned about how their colleagues' grading practices may differ from their own and how they perceived colleagues whose practices may result in grade inflation. He also found that teachers were willing to include non-academic factors or make adjustments to academic factors (e.g. allowing retakes, exempting assignments) based on their desire to meet individual students' needs and due to pressure from students and administrators. This is consistent with the findings of Zoeckler (2007), whose interviews with English teachers revealed that younger teachers adjusted the weights of their grading categories in order to limit failures and thus avoid criticism from administrators.

These observations support the idea that teachers make ground-level decisions about grading based on the need to balance accuracy with the need to have a system that meets the demands of other stakeholders such as parents and administrators. Grading policies can also be a convenient tool to manage student behavior in the absence of other strategies. If student behavior is not separated from student achievement, then their final grades become difficult to interpret and use to make decisions about future course placement. Given the variety of formal and informal training that teachers receive, it is not surprising that differential grading occurs in the classroom. The following is a review of several studies that have looked at the prevalence of this phenomenon at the high school and college levels.

Impact of Differential Grading on Students

Research on students' perception of differential grading at the high school level is sparse. In their study of college students' perceptions of grading fairness, Gordon and Fay (2010)

discovered that students' sense of fairness hinged more on their professors' teacher practices as opposed to their grading policies. While more research should be done in this area, students' lack of awareness of the impacts of differential grading does not negate their existence. The impacts of differential grading on students can be divided into two categories: immediate impacts on performance in the classroom and future impacts on scheduling and college admissions.

Immediate Impacts

According to Gusky and Link (2018), it is common for teachers to use cognitive and noncognitive factors when calculating students' grades. They also determined that "10-20% of the weight used in determining students' grades is derived from non-cognitive factors such as class participation, work habits and neatness, effort, punctuality in turning in assignments, etc." (p. 309). The use of non-cognitive factors can mean that students are being graded based on their behaviors, as opposed to their mastery of the course content. Kunnath (2017) suggested that teachers who included nonachievement factors such as effort were doing so out of consideration for the effects that students' grades would have on their futures. While these teachers may be using differential grading to adjust to individual students' needs for additional motivation or support, these practices can lead to grade inflation and thus an inaccurate measure of their academic performance. The author noted that some teachers believed that considering a student's ability in their grade provided them with a more accurate grade, despite the fact that ability is a non-achievement factor. It was unclear whether the inclusion of ability serves to raise or lower

students' grades. While grading based on effort can serve as a boost to low-performing students, how do these teachers handle a high-performing student (high ability) who puts forth little effort?

Rauschenberg (2014) cited studies that show that differential grading may occur "if certain types of students exert different amounts of effort depending on teacher characteristics" (p. 3). One study showed that students performed better with a same-gender teacher. Another study showed that Black students performed better when they had a Black teacher. This phenomenon is not a one-way street. Rauschenberg (2014) also noted that "racial, gender, and other stereotypes of student performance also may influence how a teacher issues grades" (p. 3). When students exhibit different levels of effort depending on their teacher and those teachers grade partially based on behaviors like effort, students can be adversely affected.

While including non-cognitive factors in students' grades can have negative effects, the intent of these policies can sometimes be attributed to teachers' desire to motivate students. Carifio and Carey (2009) analyzed the policy of assigning minimum grades through the lens of several motivation theories pulled from the field of educational psychology, citing the lack of research in this area. In their findings they cite a survey of high school teachers which showed that "80% of teachers who self-reported assigning minimum grades do so on a "student-by-student" basis, indicating these teachers believe some students more worthy of the minimum grade than others" (Carifio & Carey, 2009, p. 33). In their analysis of attribution theory, they state that "as for causal conditions that exist outside the student's control, attribution theory suggests that these factors must be stable if they are to have the preferred effects. Thus,

grading practices must be perceived as reliable and predictable if they are to have the intended and desired result of augmenting student expectancies" (p. 29). The authors argued that the fact that students react differently to grading policies (their tendency to approach success versus avoid failure, their beliefs about their locus of control, and that some may avoid putting forth the effort to mask their perceived lack of ability) adds a layer of complexity to grading based on teachers' perceptions of students' effort.

Proponents of policies such as minimum grades (avoiding the use of zeros on a 1-100 scale) believe that they "keep students motivated, hopeful, confident, and optimistic… There may be certain sub-populations of students who could benefit greatly from minimum grading practices and policies, particularly in certain ("high risk") courses and subject matters" (Carifio & Carey, 2009, p. 23). In addition to the immediate impacts that grading policies may have on students' classroom performance, the existence of varying policies also affect students in subsequent educational settings.

Future Impacts

Guskey and Link (2018) listed several important decisions that are made about students based on their grades, including grade-level promotion, honor roll, enrollment in future classes, and college admissions (p. 304). They also expressed concerns about what they called 'score pollution' (p. 314). When teachers use a combination of formative assessments (such as homework and quizzes) designed to measure a student's level of proficiency during the learning process and summative assessments (such as end-of-unit tests) designed to measure proficiency

at the end of the learning process, students' final grades do not represent their true mastery of the course content. When grades are used by students, parents, guidance counselors, and outside stakeholders to make educational decisions, this misrepresentation of their mastery can be problematic. The inclusion of non-achievement factors makes grades difficult to interpret and may result in grade inflation that misleads students, parents, and colleges (Alm & Colnerud, 2015; Gordan & Fay, 2010; Kunnath, 2017; Minaya, 2020; Proitz, 2013; Zoeckler, 2007). Millet (2018) argued that lenient grading practices are associated with less reliable grades. Given the importance of grades as indicators during the college admissions process, the author suggests that "internal benchmarking may provide situational awareness, reduce variability across instructor and disciplines, and improve grading reliability" (p. 1524).

According to Rauschenberg (2014), "students who receive artificially higher grades than other students… may have an advantage in college admissions. Furthermore, these students may need additional remedial work to relearn material in college, which has costly implications for the students and the state" (pp. 17-18). Moen & Tjelta (2010) argued that the differences in grading practices between some colleges were so great that students from colleges with less strict practices entering into advanced programs would potentially need to spend up to four semesters to catch up to their peers admitted from colleges with stricter practices. These "error admission" students "have lower GPA, higher failure rate, more examination attempts and less progress" than other students (p. 235).

When differential grading occurs at the college level, it can influence students' choice of degree program. As outlined in Minaya (2020), scholars have proposed that "differences in grading standards across fields may influence [students'] major choices, and STEM fields are associated with higher and more rigorous grading standards than non-STEM fields" (p. 944). Students in her study were more likely to choose a STEM major after a grading policy change at two Florida universities reduced the difference in grades between STEM and non-STEM courses. Existing STEM majors were also less likely to drop out. It is concerning that differential grading policies can influence such an important decision as a student's college major, and more importantly, some policies may be creating artificial barriers to entry into certain fields of study. Sabot and Wakeman-Linn (1991) also suggested that variations in grading leniency across disciplines can influence students' choice of majors.

Gaps in the Current Literature

Since the passage of the No Child Left Behind Act of 2001, researchers have started to take a closer look at the relationship between 1) student grades and standardized test scores, and 2) student grades and student demographics. In several studies, the Pearson correlation coefficient (or similar measure) was between 0.4 - 0.6 for grades and test scores (Bowers, 2011; Drexlerová et al., 2019; Welsh & D'Agostino, 2013; Westphal et al., 2020). This indicates a moderate correlation between teacher-assigned grades and students' standardized test scores. Of particular interest was a German study that had a coefficient of 0.4 (Westphal et al., 2020). The

authors noted that this was lower than what is typically found in studies done in the United States. However, they did not offer an explanation for the difference.

The results were mixed when it came to whether the gap between grades and test scores tended to be positive or negative. Angelo and Balcão Reis (2021) observed that teachers' grades were consistently higher, while Beswick et al. (2005) observed that they tended to be lower. Bowers (2009) noticed that the gap was smaller in core subjects as opposed to electives, but Welsh and D'Agostino (2013) reported mixed results. In their study, teacher-assigned grades in math were generally lower than the associated standardized test, while grades in reading and writing were generally higher. The impact of whether teachers err on the side of a positive or negative gap is not entirely clear, although Gershenson (2020) noticed that teachers with more rigorous grading practices saw their students' end-of-course exam scores increase by significantly greater amounts than teachers with lower grading standards.

Several of the studies proposed that the variance of teacher-assigned grades was greater within the classroom as opposed to between classrooms (Martinez et al., 2009; Welsh & D'Agostino, 2013). Aside from comparing teacher-assigned grades to standardized test scores, most of these studies were more focused on other variables (non-academic, behavioral attributes) and their relationship with grades and test scores. Studies done at a macro level (using data from a state or national database) did not parse out the specific differences between teachers. Studies done at the school level had a better chance of addressing this issue, although most did not.

Guskey and Link (2018) acknowledged that while their study tracked what elements teachers were choosing to measure with their grading policies, they did not look into why those elements were chosen. They also recommend that longitudinal studies tracking new teachers during their first few years in the classroom might provide more insight into if/how their practices change over time and what factors may cause those changes. I believe that researching the *why* might help uncover any implicit biases that teachers may have regarding their students and/or their courses. Tracking new teachers over time could provide helpful insights for those responsible for ongoing professional development at the school and district levels.

Carifio & Carey (2009) suggested that more research is needed into the effects of minimum gradings practices on student behavior. They point out the need to look into how no-zero policies are already being implemented, which is difficult due to the lack of district-wide and/or school-wide policies on this and other grading issues. Rauschenberg (2014) noted that his study did not distinguish between the causes of differential grading. While the effects of differential grading are of immediate concern to students, understanding the causes may help teacher training programs and ongoing school and district professional development efforts address this issue.

My goal is to add to the existing research, which has tended to focus on the relationship between students' grades and other factors attached to the students themselves (e.g., prior academic achievement and demographics). My research will focus on the *teacher factor*, a variable that lies entirely outside of the students themselves. To do this, I will analyze the difference between

teacher-assigned grades and students' standardized test scores and examine how that difference varies from teacher to teacher. The foundation for this study lies in previous research about the relationship between grades and test scores.

Chapter Summary

The existing literature points to the existence of differential grading across academic disciplines, districts and schools, and individual classrooms. Research focusing on subpopulations of students has shown that differential grading also occurs within individual classrooms. What is missing from the literature is how differential grading impacts students when they are divided up among multiple teachers within the same course. The purpose of this study is to discover if the *teacher* factor is a significant predictor of students' grades after controlling for students' demographics and academic achievement as measured by the standardized tests associated with the courses being examined.

Chapter 3: Methodology

The purpose of this study is to measure how much the *teacher factor* influences differences between student course grades and standardized test scores after controlling for demographic characteristics. I do not propose to be able to completely isolate the teacher factor, as there are a large number of variables that are not included in the data that influence student achievement in the classroom. However, given that students are randomly assigned to teachers' classes, the absence of these variables likely does not negate the idea that significant differences in grades between teachers are cause for concern.

Pilot Study Results

In my pilot study, I used a hierarchical linear regression model that analyzed the impact of students' demographics, prior academic achievement (as measured by their Biology EOC scores), and *teacher* on their teacher-assigned first quarter grades in Chemistry I Honors.

The relationship between students' first quarter grades in Chemistry I Honors and their Biology EOC achievement level scores was investigated using the Pearson product-moment correlation coefficient. A preliminary analysis was conducted to ensure no violation of the assumption of normality. The results are shown in Table 1. There was a medium, positive correlation between the two variables, $r = 0.358$, $n = 134$, $p < 0.001$. This analysis was conducted again using students' raw scale scores instead of their achievement levels, which may lump too many scores together that reflect very different levels of knowledge. There was a medium, positive correlation between the two variables, $r = 0.375$, $n = 137$, $p < 0.001$.

Table 1
Results of Pearson Product-Moment Correlation Examining Students' First Quarter Grades for Chemistry I Honors and their Biology EOC Achievement Level Scores

Variable		Posted Grade	Biology EOC Achievement Level
Posted Grade	Pearson's r	-	
	p-value	-	
Biology EOC Achievement Level	Pearson's r	0.358***	-
Biology EOC Raw Scale Score	Pearson's r	0.375***	-

***p < 0.001

The positive correlation between students' Biology EOC scores (a proxy for prior academic achievement) and their Chemistry I Honors course grades is to be expected. An independent samples t-test was conducted to examine whether random assignment of students to a teacher for this course could have resulted in a disproportionate number of high-scoring or low-scoring students for a particular teacher. The results are shown in Table 2. The difference in average scale scores between teachers was not significant.

Table 2
Results of Independent Samples T-Test Examining Biology EOC Scale Scores by Teacher

	Teacher 1		Teacher 2		t (131)	p	Cohen's d
	M	SD	M	SD			
Biology EOC Scale Score	424.78	18.18	423.06	21.88	0.492	0.623	0.085

Hierarchical linear regression was used to assess the ability of one control variable (*teacher*) to predict students' first quarter grades for Chemistry I Honors, after controlling for the influence of *gender*, *ESE status*, *race/ethnicity*, and *prior academic achievement* as measured by students' Biology EOC achievement level scores. The results are shown in Table 3. Preliminary analyses were conducted to ensure no violation of the assumptions of normality, linearity, multicollinearity, and homoscedasticity. *Gender*, *ESE status*, *race/ethnicity*, and *prior academic achievement* were entered as Step 1, explaining 14.8% of the variance in students' grades. After the entry of the *teacher* in Step 2, the total variance explained by the model as a whole was 46.7%, F (12, 121) = 10.709, p = <.001. In the final model, only two variables were significant: *teacher* and *gender*, with *teacher* recording a higher beta value (B = 11.672, p < .001) than *gender* (B = 3.459, p < 0.05). After accounting for demographics and prior achievement, the average grade of students assigned to Teacher 2 was 11.67 points higher than those assigned to Teacher 1. This is a difference of just over 1 letter grade on the grading scale (70-79 = C, 80-89 = B, 90-100 = A).

Table 3
Hierarchical Regression Results for Students' First Quarter Grades for Chemistry I Honors by Teacher, controlling for Gender, Race/ethnicity, ESE status, and prior achievement

Variable	B	SE B	Adjusted R^2	ΔR^2
Model 1			0.148	0.218***
Intercept	97.183***	9.904		
Gender	2.290	1.723		
ESE	0.233	2.889		
Race: Black or African American	-1.183	2.172		
Ethnicity: Hispanic	-1.576	2.529		

or Latino				
Race: American Indian or Alaskan Native	1.203	6.972		
Race: Asian	4.830	4.094		
Race: Native Hawaiian or Other Pacific Islander	6.182	9.962		
Biology EOC Level 2	-13.512	10.680		
Biology EOC Level 3	-17.141	9.922		
Biology EOC Level 4	-13.656	10.125		
Biology EOC Level 5	-7.346	9.972		
Model 2			0.467	0.297***
Intercept	86.085***	7.938		
Gender	3.459*`	1.369		
ESE	0.583	2.285		
Race: Black or African American	-1.758	1.719		
Ethnicity: Hispanic or Latino	-2.140	2.001		
Race: American Indian or Alaskan Native	7.111	5.556		
Race: Asian	2.496	3.249		
Race: Native Hawaiian or Other Pacific Islander	-0.737	7.920		
Biology EOC Level 2	-9.055	8.462		
Biology EOC Level 3	-12.671	7.864		
Biology EOC Level 4	-8.479	8.030		
Biology EOC Level 5	-2.041	7.911		
Teacher	11.672***	1.356		

Note. Model 1 includes Gender, ESE, Race: Black or African American, Ethnicity Hispanic or Latino, Race: American Indian or Alaskan Native, Race: Asian, Race: Native Hawaiian or Other Pacific Islander, Biology EOC Level 2, Biology EOC Level 3, Biology EOC Level 4, Biology EOC Level 5

*p < .05

***p < .001

While these results cannot necessarily be generalized beyond this small sample size, they did point to the need for more research in this area and potential variables to include in an expanded model.

Current Study

Research Participants/Sample

Course grades, standardized test scores, and demographics were collected for all students attending a high school in Northeast Florida during the 2018-2019 school year. This school was chosen because all data needed for the model were available, including student demographics, grades, and test history. I collected data on all students to find courses that would fit the assumptions of my statistical model and be representative of different domains. The school is a mid-sized suburban high school with approximately 1,600 students. The student population was 54.8% White, 21.1% Black, 14.5% Hispanic, and ~10% other minoritized students. 18% of students were labeled ESE and 5.2% were labeled ELL. The school has historically been a 'B' or 'C' school as measured by Florida's school grading system.

Research Design

I sought to examine differences between groups using pre-existing data, therefore a causal-comparative design was chosen. For the most part, students are randomly assigned to teachers by the student information system when the master schedule is created. While I cannot rule out that some students may have been purposely placed with a certain teacher, that is not the

norm for the majority of students. This random assignment provides a naturally occurring situation where students are divided up into different groups that can be compared to each other.

Research Variables

Dependent Variable in the Model

The dependent variable for the model is the difference between students' teacher-assigned course grades and their standardized test scores (i.e., *difference score*). To calculate this difference score, several discrete variables were available, including scale scores and achievement levels for the state's: Algebra I end-of-course exam (henceforth referred to as an EOC), Geometry EOC, Biology EOC, U.S. History EOC, and the Florida Standards Assessment for English Language Arts. These achievement levels are reported by the state and are the result of a conversion from their raw scale score to a 1-5 scale (Florida Department of Education, n.d).

Students' course grades were computed from the average of their teacher-assigned quarterly grades. Students' final grades typically include a midterm and final exam in addition to their quarterly grades. The two exam grades were left out of the calculation for the purposes of this study for two reasons: 1) those grades were not easily accessible in the student information system, and 2) those grades may be the result of a school or district-level common assessment meant to prepare students for state-level standardized tests and are therefore not under the control of the classroom teacher.

The official grading scale is 0 - 100, although teachers can assign grades greater than 100. I was unable to access their gradebooks to see how these grades were derived, but even this access would not tell the full story. If a student was assigned a grade of 92 on a test, there is no way to tell if that grade represents that student's achievement on the original test or if it is the result of 1) a retest because of a lower original grade, 2) points added for test corrections, or 3) points added for some form of extra credit.

The grades and test scores were converted to Z-scores (i.e. normal scores) because the scale and distribution of each variable were not similar.

Independent Variables in the Model

Students' gender. This dichotomous variable was coded as 0 = male and 1 = female. While this school's student information system also classifies students as male or female, this ignores the presence of students who identify as non-binary or identify as a gender other than that assigned at birth. There is a plethora of research in the area of gender and school performance (as measured by teacher-assigned grades and standardized test scores), but there has not been much done in the area of non-binary or transgender students.

Students' race/ethnicity. This discrete variable was coded as 0 = White, non-Hispanic, 1 = Black, non-Hispanic, 2 = Hispanic, 3 = Multiracial, and 4 = Asian or Pacific Islander. These classifications were based on information provided by the students' parents or guardians when they registered for school.

Students' ESE status. This dichotomous variable was coded as 0 = not ESE and 1 = ESE. Students were classified as ESE if they had an Individualized Education Program. The ESE classification also includes students identified as gifted.

Students' ELL status. This dichotomous variable (English Language Learner) is coded as 0 = not ELL and 1 = ELL. This school district had several different classifications for English Language Learners, including those who tested out of the program and were on a monitored status but did not receive formal services from the school. For this study, only those who were currently receiving services were coded as ELL.

Students' teacher. This was coded as a discrete variable. This particular school contained a self-contained ESE unit composed of four teachers who taught multiple subjects within their subject area. For example, the English teacher taught English 1, 2, 3, and 4 (all grade levels). Due to the size of the unit, each teacher typically had one of each course and had less than 10 students in each of those courses. This was in contrast to a general education or inclusion (mixed of students with and without IEPs) teacher, who typically taught just 1-2 different courses and had an average of 25 students in each class period. (Teachers at this school taught 5 class periods and had 1 planning period.) I chose not to include the self-contained unit teachers in this study due to the small number of students that they taught and the unique nature of that program.

Data Collection Procedures

I collected my data by running reports in the district's student information system. I collected data for every student in the school for the 2018-2019 school year. I ran the same report for each 9-week grading period to calculate their yearly average. The school district no longer uses this program. While most of this information was moved over into the new system, my ability to run reports in the new system and fill in any gaps in my data was limited.

I chose courses that met the following criteria: 1) had more than one teacher, 2) had more than 50 total students, and 3) had a standardized test taken by those students that could be reasonably tied to their performance in that course. For example, the Algebra 1 EOC can be reasonably used as a measure of academic achievement for Algebra 1. The Florida Standards Assessment for English Language Arts can be reasonably used as a measure of academic achievement for English 2.

Four courses were chosen for this analysis: Biology, U.S. History, English 2, and Geometry. These courses were chosen because they are required courses that have a state assessment associated with them. The standard and inclusion version of each course was included. Each of these courses has at least one upper-level version (Honors and/or Advanced Placement) that students could opt into.

Research Questions

1) How do teacher-issued course grades relate to students' scores on standardized tests within the same course?

2) How do students' demographic characteristics (gender, race/ethnicity, ELL status, and ESE status) and what teacher they were assigned to affect the difference between students' course grades and their standardized test scores?

Data Analysis

Before running the analyses for research questions one and two, I had to follow several steps to make sure that my data was ready to enter into JASP, an open-source software program similar to SPSS Statistics.

Once a course was chosen, students who were not in a teacher's class for at least 3 out of the 4 quarters were not included. There may be instances where a student was enrolled in a teacher's class for at least 3 quarters, but their 3rd quarter grade was so low (less than 10 on the 0-100 scale) that it can be assumed that the student was no longer regularly attending the class but was not unenrolled until the 4th quarter. These students were also not included in the analysis. Additionally, students that did not have a score for the standardized test associated with the course were not included in the analysis.

Descriptive statistics were run to look for patterns in the data, specifically the differences in the student populations of each teacher within the course being examined. For example, whether one teacher tended to have more students with higher standardized test scores or whether one teacher tended to have more students classified as ESE or ELL.

To answer Research Question 1, students' standardized test scores and teacher-issued course grade will be converted to z-scores and then the difference between them will be

calculated. A positive number represents a course grade Z-score that is higher than the corresponding test score Z-score. A negative number represents a course grade Z-score that is lower than the corresponding test score Z-score.

To answer Research Question 2, I will use a multiple linear regression model. The Z-score differences calculated in question one will be the dependent variable and the other student characteristics in my conceptual framework will be the independent variables.

Limitations

The use of convenience sampling to select the school being studied means that the results may not be generalizable beyond the teachers or courses examined. The teachers of the courses were not interviewed, nor were their syllabi available for examination. Therefore, there is no way of knowing their specific grading policies and practices or how students interacted with the teachers in those classrooms.

Since I am working with pre-existing data, I am limited in the number of control variables that I can use to account for the differences in students' grades other than the teacher factor (control variables are gender, race/ethnicity, ESE status, ELL status, and academic achievement as measured by standardized test scores). Other factors that can account for the variability of students' grades (e.g. student motivation, access to outside help such as tutoring, etc.) were not accounted for in this study.

Chapter Summary

The goal of this study is to use a multiple linear regression model to determine if a student's teacher affects the size of the difference between their teacher-assigned course grade and their level of academic achievement (as measured by standardized test scores). While the results may not be generalizable beyond the courses or school included in the sample, they may point to the need for more research in this area.

Chapter 4: Results

The courses chosen for this analysis had student populations that closely mirrored the school's total population, although it should be noted that male, Black, and Hispanic students are overrepresented in all four courses and ELL students are overrepresented in three out of four (see Table 4). Each course had an Honors level option that students could enroll in based on test scores and teacher recommendations. U.S. History also had an Advanced Placement option.

Table 4

Sociodemographic Characteristics of School and Individual Courses

Characteristic	School		Biology		U.S. History		English 2		Geometry	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Male	928	53.6	118	59.3	114	60.0	129	60.3	136	58.6
Female	803	46.4	81	40.7	76	40.0	85	39.7	96	41.4
ELL Status										
No	1,643	94.9	187	94.0	172	90.5	205	95.8	216	93.1
Yes	88	5.1	12	6.0	18	9.5	9	4.2	16	6.9

Sociodemographic Characteristics of School and Individual Courses

Characteristic	School		Biology		U.S. History		English 2		Geometry	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
ESE Status										
No	1,411	81.5	154	77.4	145	76.3	167	78.0	182	78.4
Yes	320	18.5	45	22.6	45	23.7	47	22.0	50	21.6
Combined Race/Ethnicity										
White, Non-Hispanic	940	54.3	91	45.7	82	43.2	102	47.7	112	48.3
Black, Non-Hispanic	379	21.9	55	27.6	51	26.8	61	28.5	65	28.0
Hispanic	245	14.2	37	18.6	43	22.6	37	17.3	38	16.4
Multi-racial	112	6.5	12	6.0	11	5.8	11	5.1	15	6.5
Amer. Indian or Alaskan Native	4	0.2	1	0.5	0	0.0	1	0.5	0	0.0
Asian or Pacific Islander	51	2.9	3	1.5	3	1.6	2	0.9	2	0.9

Research Question 1

How do teacher-issued course grades relate to students' scores on standardized tests within the same course?

For each course, descriptive statistics were used to determine how teachers compared to each other based on the Z-score differences between their assigned course grades for students and those students' standardized test scores. A positive Z-score difference indicates that the teacher tended to assign grades that were relatively higher than their students' standardized test scores. A negative Z-score difference indicates that the teacher tended to assign grades that were relatively lower than their students' standardized test scores. The following is a breakdown of the results by course.

Biology

As shown in Table 5, Biology teachers were split on the issue of positive and negative course grade and standardized test score differences. On average, Teachers 1 and 3 had negative differences while Teachers 2 and 4 had positive differences. While Teacher 1 had the average closest to zero, they also had the widest range, with minimum and maximum values far greater than the other three teachers. Teacher 2 had the largest average positive difference and Teacher 3 had the largest average negative difference, but both had the smallest standard deviations and ranges. This implies that their grading practices were the most consistent across students relative to standardized test scores. Relying solely on one measure of central tendency cannot give a holistic picture of the accuracy or consistency of teachers' grading practices.

Table 5
Z-Score Differences by Biology Teacher

	Z-score Difference			
	1	2	3	4
Valid	28	14	60	97
Mean	-0.117	0.452	-0.575	0.235
Std. Deviation	1.455	0.914	0.914	1.047
Variance	2.118	0.836	0.836	1.097
Range	6.450	3.020	4.680	5.860
Minimum	-3.020	-1.040	-2.890	-2.630
Maximum	3.430	1.980	1.790	3.230

When broken down by Biology EOC Achievement Level, all four teachers were fairly consistent in whether they tended to have a positive or negative difference within each level (see Table A1). All four tended to have positive differences with students who scored a Level 1 and

two out of four also had positive differences with students who scored a Level 2. All four teachers tended to have a negative difference with students who scored a Level 3, 4, or 5.

U.S. History

Table 6 shows the measures of central tendency for U.S. History teachers. As with Biology, the teacher with the mean closest to zero, Teacher 2, also had the widest range in Z-score differences. U.S. History teachers varied widely in their means, with Teacher 1 having, on average, a larger negative difference and Teacher 3 having, on average, a larger positive difference. All three teachers had similar standard deviations.

Table 6
Z-Score Differences by U.S. History Teacher

	Z-score Difference		
	1	2	3
Valid	63	98	29
Mean	-0.799	0.265	0.906
Std. Deviation	0.826	0.896	0.991
Variance	0.682	0.802	0.981
Range	4.160	5.130	4.190
Minimum	-2.930	-2.500	-0.420
Maximum	1.230	2.630	3.770

When broken down by U.S. History EOC Achievement Level, U.S. History teachers were not as consistent as their Biology peers. However, they did follow the general pattern of

having positive differences with lower-scoring students and negative differences with higher-scoring students (see Table A2).

English 2

Table 7 shows the data for English 2 teachers. Like Biology, they were split on the issue of positive versus negative differences. On average, Teachers 2 and 3 had negative differences while Teachers 1 and 4 had positive differences. Teacher 1 had the average closest to zero and the smallest range.

Table 7
Z-Score Differences by English 2 Teacher

	Z-score Difference			
	1	**2**	**3**	**4**
Valid	67	52	64	31
Mean	0.340	-0.636	-0.389	0.903
Std. Deviation	0.811	1.215	1.186	1.121
Variance	0.658	1.476	1.406	1.257
Range	3.950	6.110	6.650	4.960
Minimum	-1.680	-3.880	-5.180	-1.670
Maximum	2.270	2.230	1.470	3.290

When broken down by FSA ELA Achievement Level, English 2 teachers mirror their peers in Biology. (see Table A3). Teachers tended to have a positive difference with students who scored a Level 1 or 2 and a negative difference with those who scored a Level 3 or above.

Geometry

Compared to their peers in Biology, U.S. History, and English 2, Geometry teachers had the smallest range of average Z-score differences (see Table 8). Teachers 1 and 2 tended to have small negative differences while Teacher 3 had a modest positive difference. Their ranges were in line with those found in the other three courses.

Table 8
Z-Score Differences by GeometryTeacher

	Z-score Difference		
	1	2	3
Valid	120	71	41
Mean	-0.131	-0.179	0.680
Std. Deviation	0.566	1.166	1.096
Variance	0.320	1.359	1.201
Range	3.050	5.450	4.870
Minimum	-1.550	-2.130	-1.830
Maximum	1.500	3.320	3.040

When broken down by Geometry EOC Achievement Level, Geometry teachers tended to have a positive difference with students who scored a Level 1 and a negative difference with students who scored a Level 2, 3, or 4. The size of the negative difference increased as the achievement level increased (see Table A4). Teacher 1 followed this pattern. Teacher 2 had negative differences across all levels, with the smallest difference occurring with students who scored a Level 1. In other words, while Teacher 2 did not follow the overall pattern of having a positive difference for students who scored a Level 1, those students did see the smallest negative difference as compared to their peers who scored a Level 2, 3, or 4. Teacher 3 had 41 students, 40 of which scored a Level 1 or 2. Both lower levels tended to have a positive difference, with Level 1 students receiving a significantly larger difference. There was a negative difference for the one Level 4 student assigned to Teacher 3.

Summary of Findings for Research Question 1

Similar patterns of the differences between teacher-assigned grades and standardized test scores emerged across all four courses. In each course, there was a two-two or two-one split between teachers who tended to have a positive difference versus those who tended to have a negative difference. However, these descriptive statistics only tell part of the story. I noted earlier that teachers in all four courses tended to have positive differences for lower-performing students and negative differences for higher-performing students. If a teacher within a course has a greater percentage of higher or lower-performing students compared to their same-course colleagues (e.g., Biology to Biology or Geometry to Geometry), that could account for the skew

in their average Z-score difference. Research Question 2 addresses this issue by using multiple linear regression analysis and looking at the part and partial correlations of the independent variables in the model.

Research Question 2

How do students' demographic characteristics (gender, race/ethnicity, ELL status, and ESE status) and what teacher they were assigned to affect the difference between students' course grades and their standardized test scores?

Table 9 provides a summary of the significant regression coefficients from the multiple linear regression analysis performed for each course. Courses were examined separately because previous research has indicated that grading beliefs and practices may differ across academic disciplines. Students' ESE status was significant in only one course. Students' gender and ELL status were significant in two of the four courses. Students' race/ethnicity was significant in three of the four courses. Students' assigned teacher was significant in all four courses. (See Appendix A for complete tables.)

Table 9

Summary of Significant Coefficients in Multiple Regression Analysis (All Courses)

Model	Biology	U.S. History	English 2	Geometry
Gender (1)		0.008**		0.001**
Combined Race/Ethnicity (2)	0.016*		0.053*	0.011*
Combined Race/Ethnicity (3)			0.012*	
Combined Race/Ethnicity (4)				
Combined Race/Ethnicity (5)				
Combined Race/Ethnicity (6)				
English Language Learner (1)		0.046*	0.044*	
ESE Student (1)				0.007**
Teacher (2)	0.039*	<0.001***	<0.001***	

Summary of Significant Coefficients in Multiple Regression Analysis (All Courses)

Model	Biology	U.S. History	English 2	Geometry
Teacher (3)		<0.001***	<0.001***	0.045*
Teacher (4)		-		-

*p < 0.05
**p<0.01
***p < 0.001

The following sections contain the results of each individual course analysis. All models used to analyze Research Question 2 contained the same five independent variables listed in Table 6 and the dependent variable of Z-score difference that was discussed in Research Question 1.

Biology

Multiple linear regression was used to assess the degree to which the independent variables (students' gender, ELL status, ESE status, race/ethnicity, and assigned teacher) explained the difference between students' teacher-assigned course grade and their score on the standardized test score associated with that course (see Tables A5 and A6). The total variance explained by the model was 17%, F (11, 187) = 3.47, p < 0.001. Two variables were significant in the model: students' race/ethnicity and their assigned teacher. Black students, when compared to their white peers, (B = 0.44, p < 0.05) and those assigned to Teacher 2, when compared to their peers assigned to Teacher 1, (B = 0.77, p < .05) had positive Z-score differences (see Table A7).

The part and partial correlations for each independent variable are reported in Table 10. The independent variable with the largest unique contribution to the total variance explained by

the overall model was the students' assigned teacher. This variable accounted for 9.8% of the variance, which represents 58% of the total variance explained by the model. Students' race/ethnicity accounted for 4% of the variance.

Table 10

Part And Partial Correlations for Biology Coefficients (Biology)

Model		Partial	Part
H_1	Gender	0.078	0.071
	Combined Race/Ethnicity	0.214	0.200
	English Language Learner	0.042	0.039
	ESE Student	0.105	0.096
	Teacher	0.324	0.313

Note. The intercept model is omitted, as no meaningful information can be shown.

The same analysis was performed for the other three courses: U.S. History, English 2, and Geometry.

U.S. History

For U.S. History, the total variance explained by the model was 39%, F (9, 180) = 12.63, $p < 0.001$ (see Tables A8 and A9). Three variables were significant in the model: students' gender, ELL status, and their assigned teacher. Female students ($B = 0.35$, $p < 0.05$) English Language Learners ($B = 0.45$, $p < 0.05$) and students assigned to Teacher 2 ($B = 1.01$, $p < 0.001$) and Teacher 3 ($B = 1.61$, $p < 0.001$) all had positive Z-score differences (see Table A10).

The part and partial correlations for each independent variable are reported in Table 11. The independent variable with the largest unique contribution to the total variance explained by

the overall model was the students' assigned teacher. This variable accounted for 27.3% of the variance, which represented 70.5% of the total variance explained by the model.

Table 11

Part And Partial Correlations (U.S. History)

Model		Partial	Part
H_1	Gender	0.197	0.157
	Combined Race/Ethnicity	0.175	0.139
	English Language Learner	0.148	0.118
	ESE Student	0.136	0.108
	Teacher	0.556	0.523

Note. The intercept model is omitted, as no meaningful information can be shown.

English 2

For English 2, the total variance explained by the model was 29%, F (11, 202) = 7.34, $p < 0.001$ (see Tables A11 and A12). Three variables were significant in the model: students' race/ethnicity, ELL status, and their assigned teacher. Black (B = 0.33, $p < 0.05$) and Hispanic (B = 0.54, $p < 0.05$) students, as compared to their white peers, had positive Z-score differences. English Language Learners (B = 0.77, $p < 0.05$) also had positive Z-score differences. Students assigned to Teacher 2 (B = -1.04, $p < 0.001$) and Teacher 3 (B = -0.75, $p < 0.001$) had negative Z-score differences as compared to their peers assigned to Teacher 1 (see Table A13).

The part and partial correlations for each independent variable are reported in Table 12. The independent variable with the largest unique contribution to the total variance explained by the overall model was the students' assigned teacher. This variable accounted for 17.2% of the variance, which represented 60.1% of the total variance explained by the model.

Table 12

Part And Partial Correlations (English 2)

Model		Partial	Part
H_1	Gender	0.115	0.097
	Combined Race/Ethnicity	0.209	0.181
	English Language Learner	0.141	0.121
	ESE Student	0.095	0.081
	Teacher	0.441	0.415

Note. The intercept model is omitted, as no meaningful information can be shown.

Geometry

For Geometry, the total variance explained by the model was 21.3%, F (9, 222) = 6.67, $p < 0.001$ (see Tables A14 and A15). Four variables were significant in the model: students' gender, race/ethnicity, ESE status, and their assigned teacher. Female ($B = 0.38$, $p < 0.05$) and ESE students ($B = 0.54$, $p < 0.05$) both had positive Z-score differences. Black students ($B = 0.34$, $p < 0.05$) had positive Z-score differences as compared to their white peers. Students assigned to Teacher 3 ($B = 0.44$, $p < 0.05$) had positive Z-score differences as compared to their peers assigned to Teacher 1 (see Table A16).

The part and partial correlations for each independent variable are reported in Table 13. The independent variable with the largest unique contribution total variance explained by the overall model was the students' gender. This variable accounted for 3.7% of the variance, which represented 17.4% of the total variance explained by the model. Students' assigned teacher accounted for 2.4% of the variance, which represented 11.3% of the total variance explained by the model.

Table 13

Part And Partial Correlations (Geometry)

Model		Partial	Part
H_1	Gender	0.211	0.192
	Combined Race/Ethnicity	0.183	0.166
	English Language Learner	0.101	0.090
	ESE Student	0.180	0.163
	Teacher	0.172	0.155

Note. The intercept model is omitted, as no meaningful information can be shown.

Summary of Findings for Research Question 2

While at least one category of student demographics was found to be significant in each of the courses analyzed, the only factor that was significant in all four courses was the assigned teacher. In Chapter 5, I will compare these findings to those of previous studies and discuss their significance in the broader context of teacher-driven grading policies.

Chapter 5: Discussion

The results of this study are congruent with those found in previous studies about differential grading. Additionally, new findings emerged about differences across teachers within the same course.The purpose of this study was predominately to determine if a student's randomly assigned teacher would have a significant impact on their course grade after accounting for their demographics and level of academic achievement as measured by the standardized test associated with the course being analyzed. In all four courses, this teacher factor was significant.

This analysis is aided by my firsthand knowledge of this school site. As a teacher and an assistant principal at this site, I had the opportunity to observe teachers and their grading practices. This personal connection with the subject matter of this study allows me to place the data into context when needed.

Connections to Previous Research

Most of the previous research in this area has not focused on the impact of the teacher as an individual. If teachers were considered, it was typically as a group (e.g., English teachers compared to math teachers). The main focus had instead been on comparing groups of students based on demographic characteristics such as gender and race/ethnicity. The following is a synthesis of how the results of this study compare to the existing literature.

Variations Between Academic Disciplines

As discussed in Chapter 2, previous researchers have contended that math and science teachers tend to view their subject matter as more objective than subjective and tend to be less lenient graders compared to their English and social sciences peers (Barnes et al., 2001; Proitz, 2013; Sabot and Wakeman-Linn, 1991; Zoeckler, 2007). Math and science teachers are also more likely to be gatekeepers, viewing it as their role to sort students and select only the best to move on to the next level within their discipline. These beliefs are in line with criterion-referenced grading policies. This frame of reference is likely to result in teachers who base students' grades on their level of proficiency on a set amount of objective course standards. If these are the beliefs of most of the teachers in math and science, then we would expect to see similar teacher-assigned grades across teachers after accounting for student demographics and prior academic achievement.

The results of the geometry model in this study are in line with these assumptions. Of the four subjects analyzed, the geometry teachers had the smallest range of mean Z-score differences between individual teachers. The math department in this school had an unwritten rule that summative assessments should be worth 70% of students' overall quarterly grades. While this was not an official policy enforced by the administration, it was generally adhered to with slight variations from teacher to teacher. Retake policies may have varied from teacher to teacher, as well as other grading-related policies such as test corrections and acceptance of late work. The other three departments (Science, English Language Arts, and Social Studies) did not have a

similar policy. Even loose adherence to this informal grading policy could help explain why students' assigned teacher ranked four out of five in the part and partial correlations for Geometry.

Biology teachers had the second smallest range of mean Z-score differences, which confirms previous findings that have tended to show math and science teachers on the strict side of the gatekeeping spectrum. Strict graders (i.e., gatekeepers) would be less likely to include non-academic factors in their grading systems, as their main focus is measuring how students perform against a set of standards. It would make sense that teachers with this outlook would have grading systems that are more in line with their same-subject peers.

U.S. History teachers had the largest range, followed closely behind by English teachers. Previous research (Zoeckler, 2007) has shown that English teachers are more likely to include non-academic factors in their grading systems, which could account for the wider range. Factors such as participation, effort, and organization are subjective, and thus more likely to be assessed and recorded differently from teacher to teacher. This could account for the greater amount of variation in student grading outcomes from teacher to teacher.

One previous study (Welsh et al., 2013) noted that math teachers were more likely to assign grades that were lower than test scores, while English teachers were more likely to assign grades that were higher. This was not the case in this study. The English teachers were evenly split on the issue of positive versus negative Z-score differences. Two of the Geometry teachers had negative differences and the third had a positive difference. Geometry Teacher 1 had

approximately half of the students in the sample and their negative average was minimal (-0.131). However, the sample size in this study was small (four English teachers and three math teachers).

Variations Between Subpopulations of Students

Previous research (Rauschenberg, 2014) has shown that female students, students with exceptionalities, and students with limited proficiency in English are more likely to receive higher grades than their peers after holding test scores and other factors constant. The explanation given for this is the inclusion of non-academic factors in teachers' grading systems that are designed to compensate for perceived barriers that those students may be facing that could impact their academic performance.

The results of this study showed similar results. Gender was significant in the U.S. History and Geometry models, with female students receiving, on average, a positive Z-score difference. Students' ELL status (indicating limited proficiency in English) was significant in U.S. History and English 2, with English Language Learners receiving, on average, a positive Z-score difference. Students' ESE status was significant in Geometry, where students with exceptionalities received, on average, a positive Z-score difference. In all four courses, one teacher was the designated ESE teacher. This designated ESE teacher taught the majority of the students with individual education plans (IEPs), so the Z-score differences for those students may be masked by the fact that the *teacher* factor was also significant.

Students' race/ethnicity was significant in Biology, English, and Geometry, with Black students receiving, on average, a positive Z-score difference in all three courses. Hispanic students received, on average, a positive difference in English 2. In the previous research referenced in Chapter 2 on differential grading based on student demographics (Alm & Colnerud, 2015; Rauschenberg, 2014), students' race or ethnicity was either not examined or not significant.

When a positive gap occurs (a student's teacher-assigned grade is higher than their standardized test score), it could be the result of grades that overrepresent students' academic achievement, standardized test scores that under-represent their achievement, or a combination of both. If teachers are inflating grades by including non-academic factors in order to compensate for perceived barriers faced by these students, then the social contract that exists between teachers and the consumers of these grades (e.g., students, parents, colleges) is in question. If these gaps between course grades and standardized test scores are the result of teacher biases that stem from low expectations for certain groups of students, the problem becomes much broader and more systematic.

Variations Between Ability Levels of Students

Welsh et al. (2013) noted that teachers' assigned grades were more accurate (as measured by their correlation with test scores) for students who scored higher on standardized tests as opposed to those who scored lower on these tests. This could be because teachers may include non-academic factors into their grading systems in order to allow for students with lower test

scores to have other paths to increase their grades. However, those non-academic factors may also serve to boost the grades of students with higher test grades if these students are more likely to perform well on these non-academic tasks (i.e. participation, organization, effort).

All of the courses in this study were standard-level. In Biology, U.S. History, and English, teachers averaged a positive Z-score difference with students who scored a Level 1 or 2 on the standardized test associated with their course and a negative Z-score difference for students who scored a Level 3, 4, or 5. In Geometry, teachers averaged a positive difference for students who scored a Level 1 and a negative difference for students who scored a Level 2, 3, 4, or 5. The positive differences for lower-scoring students parallel the results of previous research (Rauschenberg, 2014).

For all four subjects, the differences were largest at the two extremes. In other words, students' scores at the high and low ends on the standardized test for each subject were more likely to have a grade that was not close to their test score. This implies that the grading systems of teachers of these standard-level courses were more likely to produce grades closer to their standardized test scores for students who are proficient (Level 3) and less so for students who perform far above and below the average. Students with lower test scores had, on average, grades that were increased by their teachers' grading policies. This could be the result of the inclusion of non-academic factors or classroom assessments that do not match the rigor of the associated standardized test. Students with higher test scores had, on average, grades that were lower than

their test scores would predict. This could be the result of students performing well on academic assignments but not on non-academic tasks.

New Findings

While the existence of differential grading down to the level of the individual teacher has been studied, it has not been meaningfully studied by using *teacher* as an independent variable in a quantitative analysis of grades as compared to test scores. Previous studies have described the variations in teachers' individual grading policies (to what extent they exist, why they exist, and how they might be viewed as unfair), but the actual impact on students has not been examined outside of theoretical assumptions. The results of this addition to the existing literature are described below.

This study was able to build upon previous work by quantifying the impact of teachers' individual grading policies on student outcomes. At least one teacher was a significant variable in all four models. In three of the four models, the *teacher* variable had the highest partial correlation. The coefficients for the teachers in each course are listed in Table 14.

Differences across teachers in U.S. History and English were significant at the .001 level, while teachers in Biology and Geometry were significant at the .05 level. This confirms previous findings about differences in grading across academic disciplines. This study shows that despite the propensity of math and science teachers to grade on a more objective basis, there are still significant differences across teachers of the same course within those disciplines. While there are multiple factors that can explain different outcomes for students after controlling for

demographics and standardized test scores, these results indicate that students' assigned teacher can play a significant role in determining their grades.

Table 14
Teacher Coefficients for All Subjects

	Unstandardized	Standard Error	t	p
		Biology		
Teacher (2)	0.769	0.369	2.082	**0.039***
Teacher (3)	-0.248	0.283	-0.879	0.381
Teacher (4)	0.519	0.269	1.931	0.055
		U.S. History		
Teacher (2)	1.041	0.141	7.404	**<0.001****
Teacher (3)	1.607	0.212	7.579	**<0.001****
		English 2		
Teacher (2)	-1.041	0.198	-5.267	**<0.001****
Teacher (3)	-0.751	0.186	-4.027	**<0.001****
Teacher (4)	0.389	0.259	1.5	0.135
		Geometry		
Teacher (2)	-0.16	0.134	-1.194	0.234
Teacher (3)	0.437	0.217	2.013	**0.045***

* $p < 0.05$
**$p < 0.001$

Student scheduling at this school site was done mostly by the student information system. Students' course requests were entered and a master schedule was built by guidance counselors and administrators based on course requests. The student information system then placed students into courses that match their requests. The particular class period selected for a particular course depends on students' overall requests. For example, if a student has requested a course that is only offered once during the day, then the system will place their other classes

accordingly. This method does not produce a complete schedule for all students. Some students, particularly those with multiple requests for courses that are only offered once or twice during the day, have to be scheduled by hand. The overall result is a mostly random assignment of students to teachers of the courses that they requested. Students did not get a say in choosing a particular teacher within a course.

The exception to this random assignment is the placement of ESE students - those with individual education plans (IEPs) that require them to be placed with an ESE-certified teacher. In each of the courses examined in this study, one of the teachers per course was ESE-certified. This school district uses an inclusion model, which requires that there be at least one non-ESE student in each ESE class. (Ideally, the mix of ESE and non-ESE students would be closer to 50-50 in this model.) Therefore, the students with IEPs were not placed randomly. The non-ESE students placed in the ESE-certified teachers' classes may or may not have been picked at random. School counselors may select non-ESE students whom they believe could benefit from the additional instructional supports that occur in ESE classes. A larger school - or a school with more ESE allocations - would be more likely to have more than one ESE-certified teacher per subject and therefore have random student assignments within the inclusion classes.

Given that students taking a particular course are typically assigned to teachers of that course at random, the results of this study should be concerning to students, teachers, and administrators. Before looking at the implications of differential grading, it is important to put these results in the context in which grading policies are developed.

Theoretical Framework

According to the theory of street-level bureaucracy, teachers develop grading systems that must balance the desire to accurately report students' level of knowledge with the need to have an efficient and easy-to-understand reporting system. Given the limited amount of time that teachers have without students (one 50-minute planning period, and 50 minutes of contract time after students leave for the day at this school site) and the total number of students assigned to secondary teachers (5 class periods of approximately 25 students each at this school site), teachers must rely on assessments that are easy to check and provide results that are easy for students and parents to understand.

Teachers often do this with little guidance from above, as is the case with the teachers in this school. There were no school or district-level grading policy guidelines in place to put any parameters on teachers' grading policies. The official grading scale was 0 to 100, although teachers could assign grades greater than 100. The fact that teachers could assign grades outside of the official 0 to100 scale is further evidence of the discretion teachers have when developing and implementing their grading policies.

The theory of street-level bureaucracy also suggests that teachers may rely on stereotypes as part of their effort to balance efficiency with accuracy. *Participation* or *employability* grades are an efficient way to assess students, but they are based on the assumption that activity in the classroom correlates with depth of knowledge of the material being assessed. Multiple choice tests are also an efficient way to assess students, but more time-consuming methods of

assessment (e.g., essays, class discussions, oral presentations) may be a more accurate way to distinguish between different levels of understanding.

This theory also explains that workers at the ground level of an organization have a lot of discretion in their day-to-day decisions, as it would be difficult to develop a policy manual that would cover every possible scenario. Despite having written grading policies, teachers often use their discretion when deciding when to follow their own procedure and when to make exceptions based on students' individual situations and needs. These exceptions may or may not be applied equally to all students.

The traditions of each academic discipline may also influence teachers' decisions about their grading policies. In the absence of any school wide policies, teachers may rely on pre-established norms within their departments. These norms may be influenced by the policies that teachers were exposed to during their own higher-ed experiences.

The high degree of autonomy that teachers have when developing their grading policies and the variety of factors that they can choose to assess have led to a system where grading policies can vary widely from teacher to teacher. Teachers, influenced by their own education and the practices specific to their academic fields, are often reacting to the needs of the system (an easy-to-manage, easy-to-understand gradebook) as opposed to developing policies that stem from best practices and reflect a common vision of assessment and reporting.

My interpretations of these data show that this autonomy may lead to disparate outcomes for students. The implications of those disparate outcomes for students, as well as for teachers and school leaders, are discussed below.

Implications for Practice

For Students

The grades that students receive can have immediate and far-reaching effects, with varying degrees of importance depending on the individual student. A low quarterly grade can prevent a student from being placed on the A/B Honor Roll. A low yearlong average can prevent a student from being eligible for the next level or Honors level of subsequent courses. For example, students may be required to have a C or higher in Geometry in order to move on to Algebra 2. They may need to have an A or B in Geometry order to get into Algebra 2 Honors. Two students with similar demographics and test scores may end up on different academic trajectories because they were assigned to different Geometry teachers. While schools may take test scores into consideration when placing students, they may be more apt to trust a teacher-assigned grade since it is assumed to be the result of a year's worth of observations and assessments as opposed to a one or two-day standardized test.

While neither grades nor standardized tests can completely capture a student's level of mastery in a particular course, the cumulative effect of consistently giving students grades that are lower than their test scores would predict or giving students grades that are higher than their test scores would predict may be that students are labeled as *good* or *bad* students based on these

grades. These labels may have implications for students' sense of self-worth and create a self-fulfilling prophecy as their future teachers embrace these labels.

Grades also have a cumulative impact on students' academic trajectories. Students' course grades are converted into a points system (A = 4, B = 3, C = 2, D = 1, and F = 0) and averaged together to calculate their grade point average (GPA). GPAs are used to determine class rank and a minimum of a 2.0 (C-average) is required to receive a diploma in the state where this study was conducted. Given the skewed nature of the 0-100 grading scale, very small differences in teachers' grading practices can have large impacts on their students. In the school district where this study was conducted, the letter grades A-D only encompass 10 points each (A = 90-100, B = 80-89, C = 70-79, and D = 60-69) while an F encompasses 0-59. A one point difference in students' grades - an 89 versus a 90 - results in a 3.0 GPA versus a 4.0 GPA.

Students' GPAs can heavily impact college admissions and access to scholarships. While applications typically allow students to highlight other aspects of their lives (e.g., work/volunteer experience, athletics, and other school-based extracurricular activities), GPAs are an efficient and easy-to-understand method for sorting students. The social contract that exists between secondary schools and institutions of higher education depends on the accuracy of the grades that are used to calculate students' GPAs. This social contract is becoming more important as more colleges and universities shift away from using standardized tests such as the SAT and ACT as a major factor in their admissions process.

Given that teacher-issued grades tend to be less accurate at the extremes, high-performing students competing for labels such as summa cum laude or valedictorian/salutatorian may be at a greater risk of having their grades influenced by their random assignment to a particular teacher.

While a teacher's syllabus may reflect that they are including non-academic factors in their grades, that information is typically not communicated to interested parties that are outside the school system. When colleges are evaluating students for admissions and scholarship purposes, they do not have access to the information that would tell them how a student's grade was calculated. If these grades are influenced by teachers who have significantly different grading policies, then students may gain or lose access to future educational opportunities based on their random assignment to one teacher or another.

For Teachers

While teachers are the ones assigning grades, they can also be impacted by them. Teachers' reputations among students, peers, and administrators can be influenced by the perceived fairness of their grading policies and outcomes. Teachers who consistently issue grades that are higher than their student's test scores as compared to their peers may be viewed as having an easy class and be preferred by students. Teachers who consistently issue grades that are lower than their student's test scores may be viewed as having a tough class and receive complaints from students. Significant variations among teachers within a subject or department may also cause strife that can impact teachers' ability to work together as a team.

Teachers are also impacted by the grading policies of their current students' previous teachers. If a student received grades in a previous course that was inflated by non-academic factors, that student may not be prepared for the level of rigor of the current course. The student may also feel frustrated if their new teacher does not include those same non-academic factors in their grading system. In contrast, if a student received a grade in a previous course that was lower than what their test score would predict, they may be placed in a subsequent course that is too easy for them. This student may grow frustrated and require additional enrichment to not grow bored in class.

For School Leaders

School-based administrators need to be aware of any issues that cause disparate outcomes for students, whether teacher assignment is random or purposeful. If differential grading standards result in grades that do not have a predictable correlation with academic achievement across all students, then any decision made based on those grades can adversely affect students. Students may lose access to accelerated programs if their grades are too low. Parents may be unaware of their students' academic deficiencies if grades are too high (in relation to other measures of achievement).

While school leaders may be hesitant to implement top-down uniform grading policies that take away teachers' autonomy in their classrooms, they owe it to all stakeholders to ensure that students' grades are fair and accurate. I will describe a possible approach to meeting these

goals in the Recommendations for Future Research section below. However, the limitations of this study must first be addressed.

For Teacher Preparation Programs

It is important for teacher preparation programs to include grading practices as part of their core curriculum so that new teachers are prepared to enter the classroom with research-based practices. This may prevent them from being overly influenced by the practices already in existence at their school.

Given the amount of people entering the classroom without degrees in education (especially at the secondary school level where a degree in a content area may be considered a good substitute for a degree in education), states and school districts should consider including lessons on grading in their alternative certification programs.

Limitations

One of the limitations of this study is that neither grades nor standardized test scores are a completely accurate measure of students' knowledge of a given subject. If a teacher has an average positive Z-score difference, it could be because they have more evidence with which to assess a student's level of knowledge as opposed to what is measured by one standardized test. Teachers could also be assessing standards (academic and non-academic) that are not covered by the test. I do not propose to attest to the accuracy of either measure. This study is grounded in the real-world implications of differential grading. Its purpose is not to address the legitimacy of the grading system as a whole.

Another limitation of this study is the method by which students' final grades are calculated in most school systems. Students' performance is typically averaged across grading periods to come up with a final grade. For the purposes of this study, students' quarterly grades were averaged to determine their final grades. Mid-term and final exam grades were not taken into account. While this is the traditional method for calculating students' grades, it does create a problem when trying to determine the accuracy of those grades in relation to students' standardized test scores. If a student earned a 60 for the first quarter, a 70 for the second quarter, an 80 for the third quarter, and a 90 for the fourth quarter, their final average grade would be a 75. However, the student has clearly shown growth over the course of the year. If this student scores a Level 4 or 5 on their end-of-course exam, their teacher-assigned grade will appear to under-report their academic achievement despite the fact that the teacher assigned a grade of 90 for the fourth quarter.

Another limitation is the author's inability to determine how the quarterly grades were calculated by each teacher (i.e. weighting of categories, grade recovery options, extra credit). These grades are pre-existing data points that cannot be investigated any further without interviewing the teachers who assigned them or obtaining copies of their syllabi. Individual teachers within a course could use completely different grading systems but still have results that are similar to their peers. This could happen if students who score higher on standardized tests are also better at performing any non-academic tasks that their teachers may include in their grading system and students who score lower on these tests tend to be worse at performing these

tasks. The outcomes would be high test scores/high grades and low test scores/low grades, which would result in Z-score differences close to zero.

Individual teachers could also use similar grading systems but have results that are different from their peers. This could be the result of any number of factors, including teachers' adherence to their own policies and the relationship that teachers have with their students. Students may be willing to work hard for one teacher but put in minimal effort for another.

Finally, the variance explained by each model ranged from 17% to 39%. Other factors may explain the remaining variance, such as student motivation and access to academic help outside of the classroom. These factors were not addressed in this study.

Recommendations for Future Research

This study creates a new lens through which to analyze differential grading, but it only represents one school site. More research is needed to see if the *teacher* factor is significant in other settings. Different class sizes should also be considered. Teachers with a smaller student load may not have to lean as heavily on efficiency when developing their grading policies, and thus be able to use measures that more accurately assess their students' level of academic performance.

While the courses in this study were directly tied to a standardized test that served as a proxy for academic achievement, this analysis should also be done with courses that do not have an associated test. In the school district where this school site is located, all standard-level courses contribute evenly to students' GPAs. For example, a C earned in an elective course and a

C earned in a core content course impact GPA in the same way. (In this district, Honors level courses are given an extra .5 point and Advanced Placement courses are given an extra 1.0 point on the 0-4 scale.) Some colleges also require students to submit a core class GPA to avoid this problem.

If this phenomenon is widespread, then the next step should be action research that includes teachers in developing possible solutions. Depending on the level of autonomy that has previously existed within a school site or district, there may be considerable resistance from teachers to the idea of adopting common policies. This resistance might be softened if teachers are able to see firsthand the outcomes that result from their current practices. The Professional Learning Community model used by this school district lends itself to this type of research. Given that grading is a daily exercise that occurs at the ground level of education, it stands to reason that solutions should be developed from the ground up.

This work may begin as an independent study for individual teachers in order to give them the opportunity to analyze their own data in a safe environment. Once teachers feel comfortable with their own data, they may be more willing to share and analyze data at the course or department level via collaborative action research projects.

A potential study to prepare for this step would be presenting the results of this study to a focus group of teachers. A series of open-ended questions could be developed to gauge how teachers react to the data and determine how best to present it to maximize teacher buy-in of potential changes to their grading practices.

Conclusions

The results of this study add to the existing body of research regarding differential grading by taking a quantitative approach to what has up until now been mostly a theoretical discussion about the fairness of teacher-developed grading policies. The results point to the need for more research in this area to confirm the findings that the *teacher* factor has a significant impact on student outcomes. If additional research points to a systemic problem, school leaders should consider more professional development in the area of grading and teacher-preparation programs should include it as part of their core curriculum.

References

Alm, F. & Colnerud, G. (2015). Teachers' experiences of unfair grading. *Educational Assessment, 20,* 132-150.

Angelo, C., & Reis, A. B. (2021). Gender gaps in different grading systems. *Education Economics, 29*(1), 105-119. http://dx.doi.org/10.1080/09645292.2020.1853681

Archbald, D., Glutting, J., & Qian, X. (2009). Getting into honors or not: An analysis of the relative influence of grades, test scores, and race on track placement in a comprehensive high school. *American Secondary Education, 37*(2), 65-81.

Barberis, E., & Buchowicz, I. (2015). Creating accessibility to education: The role of school staff's discretionary practices. *European Education, 47*(1), 61-76. doi:10.1080/10564934.2014.1001264

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

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*(1), 116.

Bizup, J., Booth, W. C., Colomb, G. G., Fitzgerald, W. T., & Williams, J. M. (2016). *The craft of research* (4th ed.). Chicago: The University of Chicago Press.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47*(5), 609-629. https://10.1108/09578230910981080

Bowers, A. J. (2011). What's in a grade? the multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research & Evaluation, 17*(3), 141-159.

Carifio, J., & Carey, T. (2009). A critical examination of current minimum grading policy recommendations. *High School Journal, 93*(1), 23-37. doi:10.1353/hsj.0.0039

Creswell, J. W. & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Los Angeles: SAGE Publishing.

Czibor, E., Onderstal, S., Sloof, R., & van Praag, C. M. (2020). Does relative grading help male students? Evidence from a field experiment in the classroom. *Economics of Education Review, 75.*

Drexlerová, A., Šeďová, K., & Sedláček, M. (2019). The relationship between grading and teacher judgment. *Journal of Pedagogy / Pedagogický Casopis, 10*(2), 9-35. https://10.2478/jped-2019-0005

Florida Department of Education. (n.d.). *Florida standards assessments scale scores for each achievement level.* https://www.fldoe.org/accountability/assessments/k-12-student-assessment/fsa.stml

Gershenson, S. (2020). End the 'easy A': Tougher grading standards set more students up for success. *Education Next,* (2), 18.

Gordon, M., & Fay, C. (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. *College Teaching, 58*(3), 93-98.

Guskey, T. R., & Link, L. J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303-320. doi:10.1080/0969594X.2018.1555515

Harrits, G. S. (2019). Stereotypes in context: How and when do Street‐Level bureaucrats use class stereotypes? *Public Administration Review, 79*(1), 93-103. doi:10.1111/puar.12952

Knight. M., & Cooper, R. (2019) Taking on a new grading system: The interconnected effects of standards-based grading on teaching, learning, assessment, and student behavior. *NASSP Bulletin*, *103*(1), 65-92. doi.org/10.1177/0192636519826709

Kunnath, J. P. (2017). Teacher grading decisions: Influence, rationale, and practices. *American Secondary Education, 45*(3), 68-88.

Lehman, E., De Jong, D., & Baron, M. (2018) Investigating the relationship of standards-based grades vs. traditional-based grades to results of the scholastic math inventory at the middle school level. *Education Leadership Review of Doctoral Research, 6*, 1-16.

Lewis, D. (2019) Gender effects on re-assessment attempts in a standards-based grading implementation. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*. doi: 10.1080/10511970.2019.1616636

Lipsky, M. & Weatherley, R. (1977). Street-level bureaucrats and institutional Innovation: Implementing special-education reform. *Harvard Educational Review*, *47*(2), 171-197.

MacDermott, R. J. (2013). The impact of assessment policy on learning: Replacement exams or grade dropping. *The Journal of Economic Education, 44*(4), 364-371.

Maag Merki, K., & Holmeier, M. (2015). Comparability of semester and exit exam grades: Long-term effect of the implementation of state-wide exit exams. *School Effectiveness & School Improvement, 26*(1), 57-74.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78-102. https://10.1080/10627190903039429

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.

Millet, I. (2018). The relationship between grading leniency and grading reliability. *Studies in Higher Education, 43*(9), 1524-1535.

Minaya, V. (2020). Do differential grading standards across fields matter for major choice? Evidence from a policy change in Florida. *Research in Higher Education, 61,* 943-965.

Møen, J., & Tjelta, M. (2010). Grading standards, student ability and errors in college admission. *Scandinavian Journal of Educational Research, 54*(3), 221-237. doi:10.1080/00313831003764503

Pollio, M., & Hochbein, C. (2015) The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record*, *117(11)*, 1-28.

Prøitz, T. S. (2013). Variations in grading practice - subjects matter. *Education Inquiry, 4*(3), 555-575. https://10.3402/edui.v4i3.22629

Rauschenberg, S. (2014). How consistent are course grades? An examination of differential grading. *Education Policy Analysis Archives, 22*(92), 1-38.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *The Journal of Economic Perspectives, 5*(1), 159-170.

Sattin-Bajaj, C., Jennings, J. L., Corcoran, S. P., Baker-Smith, E., & Hailey, C. (2018). Surviving at the street level: How counselors' implementation of school choice policy shapes students' high school destinations. *Sociology of Education, 91*(1), 46-71. doi:10.1177/0038040717751443

Steward, R. J., Hill, M. F., Neil, D. M., Pritchett, T., & Wabaunsee, A. (2008). What does GPA in an urban high school actually mean? *Educational Considerations, 36*(1), 11-16.

Svennberg, L., Meckbach, J., & Redelius, K. (2014). Exploring PE teachers' 'gut feelings': An attempt to verbalise and discuss teachers' internalised grading criteria. *European Physical Education Review, 20*(2), 199-214. https://10.1177/1356336X13517437.

Swan, G., Guskey, T., & Jung, L. (2014) Parents' and teachers' perceptions of standards-based and traditional report cards. *Educational Assessment, Evaluation & Accountability*, *26(3),* 289-299.

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research & Evaluation, 18*(2), 153-172.

Townsley, M., & Varga, M. (2018) Getting high school students ready for college: A quantitative study of standards-based grading practices. *Journal of Research in Education, 28*(1), 92-112.

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. (2013) Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues & Practice, 32(2),* 26-36. doi: 10.1111/emip.12009

Westphal, A., Lazarides, R., & Vock, M. (2020). Are some students graded more appropriately than others? Student characteristics as moderators of the relationships between teacher-assigned grades and test scores in mathematics. *The British Journal of Educational Psychology*, e12397. https://10.1111/bjep.12397

Wisch, J.K., Ousterhout, B. H., Carter, V., & Orr, B. (2018). The grading gradient: Teacher motivations for varied redo and retake policies. *Studies in Educational Evaluation, 58,* 145-155.

Wormeli, R. (2018). *Fair isn't always equal: Assessment and grading in the differentiated classroom* (2nd ed). Portland, ME: Stenhouse Publishers.

Wray, M., & Houghton, A. (2019). Implementing disability policy in teaching and learning contexts – shop floor constructivism or street level bureaucracy? *Teaching in Higher Education, 24*(4), 510-526. doi:10.1080/13562517.2018.1491838

Zoeckler, L. G. (2007). Moral aspects of grading: A study of high school english teachers' perceptions. *American Secondary Education 35*(2), 83-102.

Appendix A - Tables

Table A1

Z-Score Difference by Biology EOC Achievement Level (All Teachers)

	Biology EOC Achievement Level				
	1	2	3	4	5
Valid	45	71	67	10	6
Missing	0	0	0	0	0
Mean	0.835	0.028	-0.464	-1.103	-1.010
Std. Deviation	1.298	0.982	0.768	0.770	0.476
Minimum	-2.450	-3.020	-2.890	-2.690	-1.860
Maximum	3.430	1.710	1.230	-0.130	-0.500

Table A2

Z-Score Difference by U.S. History EOC Achievement Level (All Teachers)

	U.S. History EOC Achievement Level				
	1	2	3	4	5
Valid	32	48	57	32	21
Mean	0.966	0.250	-0.104	-0.479	-0.940
Std. Deviation	1.365	1.043	0.700	0.706	0.608
Minimum	-2.500	-2.930	-1.880	-2.210	-2.170
Maximum	3.770	1.770	0.940	0.330	-0.170

Table A3

Z-Score Difference by FSA ELA Achievement Level (All Teachers)

	FSA ELA Achievement Level				
	1	2	3	4	5
Valid	61	62	73	17	1
Missing	0	0	0	0	0
Mean	0.870	0.031	-0.603	-0.988	-1.280
Std. Deviation	1.288	0.989	0.833	0.731	NaN
Minimum	-5.180	-3.460	-3.880	-2.690	-1.280
Maximum	3.290	1.220	0.710	0.290	-1.280

Table A4

Z-Score Difference by Geometry EOC Achievement Level (All Teachers)

	Geometry EOC Achievement Level			
	1	2	3	4
Valid	98	53	69	12
Mean	0.398	-0.215	-0.308	-0.582
Std. Deviation	1.160	0.775	0.415	0.472
Minimum	-2.130	-2.030	-1.160	-1.470
Maximum	3.320	0.880	1.030	0.060

Table A5

Model Summary - Z-score Difference (Biology)

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	1.125
H_1	0.412	0.170	0.121	1.055

Table A6

ANOVA (Biology)

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	42.539	11	3.867	3.473	<0.001*
	Residual	208.222	187	1.113		
	Total	250.761	198			

Note. The intercept model is omitted, as no meaningful information can be shown.
* $p < 0.001$

Table A7

Coefficients (Biology)

Model		Unstandardized	Standard Error	t	p
H_0	(Intercept)	-0.044	0.080	-0.547	0.585
H_1	(Intercept)	-0.594	0.275	-2.162	0.032
	Gender (1)	0.168	0.158	1.065	0.288
	Combined Race/Ethnicity (2)	0.440	0.182	2.422	0.016*
	Combined Race/Ethnicity (3)	0.309	0.221	1.398	0.164
	Combined Race/Ethnicity (4)	-0.037	0.334	-0.110	0.913
	Combined Race/Ethnicity (5)	-0.267	1.070	-0.250	0.803
	Combined Race/Ethnicity (6)	1.062	0.642	1.654	0.100
	English Language Learner (1)	-0.196	0.339	-0.579	0.563
	ESE Student (1)	0.311	0.216	1.438	0.152
	Teacher (2)	0.769	0.369	2.082	0.039*
	Teacher (3)	-0.248	0.283	-0.879	0.381
	Teacher (4)	0.519	0.269	1.931	0.055

[a] Standardized coefficients can only be computed for continuous predictors.

* $p < 0.05$

Table A8

Model Summary - Z-score Difference (U.S. History)

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	1.075
H_1	0.622	0.387	0.356	0.863

Table A9

ANOVA (U.S. History)

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	84.568	9	9.396	12.629	<0.001*
	Residual	133.923	180	0.744		
	Total	218.491	189			

Note. The intercept model is omitted, as no meaningful information can be shown.

*$p < 0.05$

Table A10

Coefficients (U.S. History)

Model		Unstandardized	Standard Error	t	p
H_0	(Intercept)	0.010	0.078	0.130	0.897
H_1	(Intercept)	-1.108	0.146	-7.588	<0.001
	Gender (1)	0.349	0.130	2.689	0.008*
	Combined Race/Ethnicity (2)	0.186	0.155	1.197	0.233
	Combined Race/Ethnicity (3)	-0.030	0.171	-0.175	0.861
	Combined Race/Ethnicity (4)	0.471	0.279	1.686	0.094
	Combined Race/Ethnicity (6)	0.675	0.517	1.306	0.193
	English Language Learner (1)	0.456	0.227	2.014	0.046*
	ESE Student (1)	0.305	0.165	1.843	0.067
	Teacher (2)	1.041	0.141	7.404	<0.001**
	Teacher (3)	1.607	0.212	7.579	<0.001**

[a] Standardized coefficients can only be computed for continuous predictors.

*p < 0.05

**p < 0.001

Table A11

Model Summary - Z-score Difference (English 2)

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	1.203
H_1	0.534	0.286	0.247	1.045

Table 12

ANOVA (English 2)

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	88.078	11	8.007	7.338	<0.001
	Residual	220.414	202	1.091		
	Total	308.492	213			

Note. The intercept model is omitted, as no meaningful information can be shown.

Table A13

Coefficients (English 2)

Model		Unstandardized	Standard Error	t	p
H_0	(Intercept)	-0.034	0.082	-0.407	0.684
H_1	(Intercept)	0.002	0.159	0.015	0.988
	Gender (1)	0.245	0.149	1.639	0.103
	Combined Race/Ethnicity (2)	0.332	0.170	1.949	0.053*
	Combined Race/Ethnicity (3)	0.543	0.215	2.523	0.012*
	Combined Race/Ethnicity (4)	-0.040	0.335	-0.119	0.905
	Combined Race/Ethnicity (5)	0.758	1.057	0.717	0.474
	Combined Race/Ethnicity (6)	0.641	0.749	0.855	0.393
	English Language Learner (1)	0.770	0.380	2.027	0.044*
	ESE Student (1)	0.272	0.200	1.363	0.174
	Teacher (2)	-1.041	0.198	-5.267	<0.001**
	Teacher (3)	-0.751	0.186	-4.027	<0.001**
	Teacher (4)	0.389	0.259	1.500	0.135

[a] Standardized coefficients can only be computed for continuous predictors.

*p < 0.05

**p < 0.001

Table A14

Model Summary - Z-score Difference (Geometry)

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	0.941
H_1	0.461	0.213	0.181	0.852

Table A15

ANOVA (Geometry)

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	43.524	9	4.836	6.669	<0.001*
	Residual	160.994	222	0.725		
	Total	204.518	231			

Note. The intercept model is omitted, as no meaningful information can be shown.

*p <0.05

Table A16

Coefficients (Geometry)

Model		Unstandardized	Standard Error	t	p
H_0	(Intercept)	-0.002	0.062	-0.040	0.968
H_1	(Intercept)	-0.484	0.108	-4.488	<0.001
	Gender (1)	0.375	0.116	3.222	0.001*
	Combined Race/Ethnicity (2)	0.344	0.134	2.556	0.011*
	Combined Race/Ethnicity (3)	0.287	0.166	1.729	0.085
	Combined Race/Ethnicity (4)	0.159	0.235	0.677	0.499
	Combined Race/Ethnicity (6)	0.412	0.610	0.676	0.500
	English Language Learner (1)	0.356	0.236	1.513	0.132
	ESE Student (1)	0.542	0.199	2.730	0.007*
	Teacher (2)	-0.160	0.134	-1.194	0.234
	Teacher (3)	0.437	0.217	2.013	0.045*

[a] Standardized coefficients can only be computed for continuous predictors.

*$p < 0.05$