2024

An Analysis of Teacher Characteristics and Perceptions of Grading Practices on Student Grade Outcomes

Michael B. Meechin
University of Central Florida

AN ANALYSIS OF TEACHER CHARACTERISTICS AND PERCEPTIONS OF GRADING PRACTICES ON STUDENT GRADE OUTCOMES

by

MICHAEL B. MEECHIN
B.A. Westfield State University, 2003
M.ED. Stetson University, 2010

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the Department of Educational Leadership and Higher Education
in the College of Community Innovation and Education
at the University of Central Florida
Orlando, Florida

Summer Term
2024

Major Professors: Daniel Eadens and Larry Walker

ABSTRACT

The purpose of this study was to examine the relationship, if any, between teachers' grade distributions and other moderating factors (i.e., teacher perceptions of grading practices as indicated on the Teacher Perceptions of Grading Practices Survey (TPGPS), grade level taught, enrollment in teacher preparation program, and grading practice methodology). A Pearson correlation was used to research statistical significance in the relationship of a teacher's perceptions of grading practices as measured by the TPGPS and their grade distribution (as measured by an average of scores, calculated on a 4.0 scale, in one course over one academic quarter). A one-way ANOVA was used to research a statistically significant relationship between the grade level taught and grade distribution. The same methodology was used to study relationships between grade distribution and enrollment in a teacher preparation program and the grading methodology used in practice (i.e., traditional, standards-based, hybrid). Findings from this study resulted in no statistical significance in the relationship between grade distribution and any of the moderating factors, with the exception of grade level taught. School leaders and policymakers may benefit from learning that factors traditionally used to guide the development of supports in grading practices have no statistically significant relationship to a teacher's grade distribution. This study leaves space in the research community to explore what factors, if any, impact a teacher's grade distribution in our work in grading and assessment and the implications for practice and policy in grading reform.

To all public school students in the United States

who do not have access to quality education, yet persist. I see you.

ACKNOWLEDGEMENTS

I had put off this journey for as long as I could. I had no idea how much I would need to lean on an incredible village of people to persist through this experience. To my committee chairs, Dr. Dan Eadens and Dr. Larry Walker, thank you for your patience, transparency, guidance, and support in completing this work. To my committee members, Dr. William Gordon and Dr. Scott Fritz, thank you for the feedback and encouragement. Dr. Fritz, thank you for tapping my shoulder and encouraging me to follow this path.

Thank you to all the educators who provided input for this study. We must move education forward with discourse about our most crucial reform topics. You have profoundly impacted the body of research on grading and assessment.

To Cohort 12, I appreciate being able to take this journey beside you all. Those late-night text messages and encouragement meant more than you will ever know.

To my mom, you instilled in me a work ethic to persist through any barrier in my way. In contrast, that persistence did not always manifest while I was a student in my early years. I am hopeful that my quest to be a lifelong learner more than makes up for it.

Lastly, I thank my kids, Jack and Kennedy, for being patient and forgiving when I needed to miss a soccer game or cheerleading competition to achieve my goals. I hope I served as a model to work hard and aim high – you can accomplish anything. My wife, Angela, I am who I am because I am beside you. You have provided an opportunity for me to pursue my dreams, and you are steadfast in your support. I would be nothing without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Background of the Study

Since Edgeworth (1888), scholars have presented their findings on the effectiveness of grades and the inconsistencies seen among educators in reporting student learning outcomes. A great trust exists between the parent and teacher that what is reported as a grade accurately represents student learning aligned to outcomes. Grades and the cumulative grade point average students build may lead to academic honors for a student and open doors to college admissions and incentives in the form of financial assistance. This may lead to a path of remediation, intervention, and even dropout.

Stanley and Baines (2001) remind us of the power of the grade, "The grade is the cornerstone for communication among parents, students, administrators, and teachers. If the cornerstone is faulty, the entire structure may eventually collapse" (p. 227). This point is emphasized by Guskey (2006), who describes the hodgepodge effect in which teachers reduce a student's academic performance into a single course grade, which makes that grade meaningless.

Furthermore, over a century ago, the research highlighted the inconsistencies in a student's grade. Starch (1913) discussed where 142 teachers of English graded two English papers, and scores on the paper ranged from 64 to 98 on one paper and 50 to 98 on the other. This was not isolated to English, as Starch found a similar disparity in scores in mathematics. In Starch's study, over 30 different scores were assigned to one paper, and there was a point distribution over the range of 40 points (Link, 2018). Teachers use many factors to determine and assign a student a grade, and students do not know what defines the grade. Our schooling

experiences speak to the variables that have composed the grades reported at the end of the quarter, semester, and year marks. Historically, it has been accepted that teachers use various measures, from academic performance data to other factors not necessarily aligned with educational outcomes (Brookhart et al., 2020).

Brimi (2011) repeated Starch's study almost 100 years later. In this study, the 73 teachers who participated received training in scoring writing. The teachers then scored one piece of writing, resulting in score distributions ranging from a high of 96 to a low of 50, a distribution of 46 points (Brimi, 2011).

This *hodgepodge effect* is discussed in the research on the use of standardized admissions tests, which "provide a neutral yardstick to assess the performance and promise of students from secondary schools whose course offerings differ widely in variety and rigor" (Buckley et al., 2018, p. 2-3). When determining the measure of academic proficiency defined by a grade, academia is must reference other metrics. Moreover, teachers' grades are not always aligned with state standards, leaving stakeholders wondering where the criterion that makes up the student grade originates (Northern & Petrilli, 2018).

> The Covid Pandemic highlighted some of these shortcomings in grading practice: Changes in grading prompted by the pandemic also forced district leaders to consider the meaning of grades and precisely what they communicate. Many began to recognize that when teachers combine evidence on student achievement with data on homework completion, class participation, punctuality in turning in assignments, and other work habits, the grade becomes a confusing amalgamation that is impossible to interpret. (Guskey, 2021, para. 13)

Furthermore, the research from a century ago reminds us how stagnant any reform efforts have been in producing better outcomes; as Finkelstein (1913) remarks, “When we consider the practically universal use in all educational institutions of a system of marks, we can but be astonished at the blind faith that had been felt in the reliability of the marking systems” (p. 1).

Brookhart (1993) emphasized the central idea of why grades are so important: “Grading is important to study because of the centrality of grades in the educational experience of all students” (p. 139). Every student in every school, from pre-kindergarten to post-secondary, will experience being assigned a grade.

Statement of the Problem

The integrity and significance of student grades are increasingly being called into question. Stanley and Baines (2001) discuss the difficulty in understanding the meaning of a grade, “A student’s final grade does not always simply reflect academic performance. Instead, grades now serve as a potpourri of inappropriate purposes including, but not limited to, self-esteem boosters, public relations, rewards, and vehicles to increase college funding for students” (p. 1372). The variety of variables that make up a student's grade leads to significant reliability, validity, and bias issues that potentially impact student grades positively and negatively. Erickson (2010) reported on the improvement of grading practices. Teachers in the study gave over ten different reasons a student could receive a particular grade – none of which discussed whether the student had proficient knowledge of the course content (Erickson, 2010). Students' grades become a less meaningful measure of what a student has learned (Arum & Roska, 2011; Brimi, 2011; Brookhart et al., 2020; Guskey, 2022; Link, 2023).

Schools make data-based decisions regarding course placements, remediation, or interventions based on inaccurate data reported as grades. For example, students access to more rigorous coursework may be determined by their previous course grades, or the previous course grades determine enrollment in a remedial reading or mathematics course (Grissom et al., 2015; Kalgorides et al., 2013; Luschei & Jeong, 2018). The reliability of the measurement of student grades is under scrutiny, “The grades teachers use to describe students’ performance in school and record on report cards have long been identified by the measurement community as prime examples of unreliable measurement” (Brookhart, 1993; Stiggins et al., 1989, as cited in Guskey & Link, 2018, p. 1).

Link (2018) found that teachers generally agree on the value of grading for encouraging good work and providing students with feedback. However, the research also revealed anxieties regarding the subjective nature of traditional grading methods. As one teacher in the study stated, "There's always a little bit of subjectivity, especially when it comes to things like participation or effort" (Link, 2018, p. 127). This subjectivity can lead to concerns about fairness and potential bias to influence grades.

Furthermore, research by Liu (2018) suggests that teacher workload associated with grading can be a significant factor influencing their perceptions. The study found that "teachers, particularly those with large class sizes, reported feeling overwhelmed by the sheer volume of grading required" (Liu, 2018, p. 412). This can lead to frustration and a desire for more efficient grading practices that maintain quality feedback.

Based on the literature, there is a hodgepodge of grading practices and factors impacting student academic performance, leaving all stakeholders with inaccurate representations of

student learning (Arum & Roska, 2011; Brimi, 2011; Brookhart et al., 2020; Guskey, 2022; Link, 2023). For this study, select teacher characteristics and their perceptions of grading practices are of particular interest, specifically to examine if there is an impact on student grade distributions.

Purpose of the Study

This study examines the influence of teacher perceptions of grading practices on their students' grade distributions, the effect of teacher characteristics on students' grade distributions, and the extent to which a teacher's grading system used in practice results in differing distribution outcomes.

Upon analysis, teacher characteristics (e.g., grade level taught, completion of a traditional university-based teacher preparation program, and grading methodology used in practice) and a teacher's perceptions of grading practices can inform the work of school-based leaders, school districts, and university teacher preparation programs to understand better how grading practices and beliefs may impact students' grades. For this reason stakeholders can better align professional learning, leadership strategies, and instructional coaching to support research-based effective grading methodologies.

Given the inherent complexity of the grading system in the United States, clarifying grading and assessment practices requires adding clarity and a better understanding of how teacher characteristics and perceptions interact with grade outcomes.

Grades have an impact on almost all K-12 students. The significance of this study may have broad implications as most students will experience being assigned a grade in their K-12

enrollment (Guskey, 2022; Guskey & Bailey, 2001; Kalgorides et al., 2013; Link & Guskey, 2022; Reeves, 2011).

Significance of the Study

Student grades, utilized in educational institutions nationwide from kindergarten through twelfth grade, frequently fail to accurately reflect their academic proficiency (Brimi, 2011; Link, 2023). Decisions regarding a student's future, access to rigorous coursework, or college acceptances are predicated on these grades (Guskey, 2022; Guskey & Bailey, 2001; Kalgorides et al., 2013; Link & Guskey, 2022; Reeves, 2011). For example, schools make master scheduling decisions based mainly on grades from previous coursework (Luschei & Jeong, 2018). If research indicates that grades frequently incorporate criteria unrelated to academic comprehension, a student's placement in remedial courses of access to advanced learning opportunities may be determined by inaccurate assessments of their academic ability (Kalgorides et al., 2013; Luschei & Jeong, 2018).

This study aims to generate data that will illuminate any potential relationships between specific teacher characteristics, a teacher's perceptions of grading practices, and the resulting grade distributions. The field must understand how these factors impact student grade outcomes (Link, 2018). Further research should utilize both perception data regarding grading and teacher characteristics, in conjunction with their impact on grade distributions, to gain a deeper understanding of how grading practices affect students (Link, 2018; Liu et al., 2008).

Definition of Terms

100-Point Scale

The 100-point scale will be referenced throughout this study and refers to the traditional grading scale of 0-100 points. Integers under 60 will be considered a failing score on the traditional scale (Guskey, 2013).

Complex Systems Theory

This study used the Complex Systems Theory to guide the development of research questions. This theory assumes that systems become more complex and more challenging to scale. The theory also calls on the variables that make us a system to be looked at rather than independent of one another (San Miguel, 2023).

Grade

Throughout this study, the term *grade* will reference a student's final mark on an assignment or at the end of a course (Guskey, 2013).

Grade Distribution

Grade distribution is the number of As, Bs, Cs, Ds, and Fs a teacher has in a course at the end of a marking period (Reeves, 2010).

Hodgepodge Grading

The literature refers to hodgepodge grading as grades that include “attitude, effort, and achievement (Brookhart, 1991, p. 36).

Hybrid Grading Systems

Hybrid grading systems will be understood as teachers using parts of standards-based grading and traditional grading systems (Guskey, 2021).

Nontraditional Teacher Training

Nontraditional teacher training refers to teachers who have yet to go through or enroll in a traditional teacher education school. This group should also not be considered to have completed student teaching (Link, 2023).

Standards-Based Grading Systems

"Standards-based grading is the name attached to grading systems in which students' achievement and progress in school are evaluated based on their proficiency in meeting articulated learning standards" (Tomlinson & McTighe, 2006).

Traditional Grading Systems

The traditional grading system refers to the widely used 0–100-point scale in schools and classrooms. This system uses a traditional distribution of grades, with 90-100 representing an A, 80-89 representing a B, 70-79 representing a C, 60-69 representing a D, and an F grade falling between 0 and 59 (Reeves, 2010).

Traditional Teacher Training

Traditional teacher training refers to teachers who have enrolled in a teacher education school and received formal training on best practices and pedagogy. This should be considered a 4-year college or university teacher program with student teaching experience (Link, 2023).

Theoretical Framework

Since Edgeworth's seminal work in 1888, which highlighted the multifaceted nature of student outcome measurement, research on grading has remained steadfast in its pursuit of comprehensively understanding the various factors influencing student achievement. However,

recent research has focused mainly on the measures that make up a grade in isolation (i.e., dropping the use of the zero, switching to a 4-point scale, and introducing rubrics or scales) (Brookhart et al., 2020; Reeves, 2011; Reeves, 2004).

Grading systems are complex. Consider a student's final grade in a mathematics course. This grade might be influenced by factors such as exam performance, homework assignments, class participation, projects, and possibly factors like attendance or extra credit opportunities. Each of these components interacts with one another to shape the overall grade the student receives in the course. How can we understand the relationships between these variables and their combined impact on determining a student's final grade? This study looks through the lens of Complex Systems Theory; according to San Miguel (2023), “A good definition is that Complex System is composed of many interacting units showing emergent properties that cannot be understood in terms of the properties of the individual isolated components” (p. 2). Practitioners face a significant challenge in enhancing grading practices due to the system's inherent complexity, which prohibits isolated examination of individual components. For instance, one teacher might employ a total points system while another might utilize percentage scores.

Furthermore, within these scoring systems, teachers may grade for completion, another may provide a curve, and another may award extra credit (Guskey, 2022; Link, 2023). The system that persists lacks coherence and structure. It requires interpreters to navigate a multitude of measures, seeking meaning amidst complexity (Reeves, 2004).

Drawing from chaos theory, which asserts that problems stem from the entire system rather than isolated parts, Briggs and Peat (1999) underscore the importance of considering the

broader context in addressing issues rather than viewing them as mechanical problems with isolated solutions.

> Chaos theory teaches us that we are always a part of the problem and that particular tension and dislocation always unfold from the entire system rather than from some defective "part." Envisioning an issue as a purely mechanical problem to be solved may bring temporary relief of symptoms, but chaos suggests that in the long run, it could be more effective to look at the overall context in which a particular problems manifest itself. (Briggs & Peat, 1999, p. 160-161)

Chaos theory principles applied to grading systems suggest that isolated reforms, such as the implementation or removal of a zero on a 100-point scale, may yield unintended consequences if not considered within the context of the entire system (Briggs & Peat, 1999).

Through the lens of Complex Systems Theory, we can better understand the challenges faced by grading reform initiatives in gaining traction over the past century. Practitioners and researchers are looking for the "defective part" that Briggs and Peat (1999) outline in their work. Attempting to fix or support one area of grading (e.g., grading for completion vs. authentic grading practices supported by feedback) will not allow for the outcomes desired. Instead, building professional learning that supports this work requires looking at the grading systems in a school or district.

For grading practices to become more effective, we leverage complex systems science, which considers systems with many components and how they are all related (Siegenfeld & Bar-Yam, 2020). For example, society may have a high-level understanding of what a "grade" is;

however, even with that common understanding, there is no consistency in how each school or system determines those grades.

San Miguel (2023) pointed to systems with many interacting *units*, such as those that make up our grading systems. With so many variables, grading systems are examples of systems with interactions at many different scales: classroom level, school level, district level, and state level. Understanding grades as complex systems can help educators better assess their significance (San Miguel, 2023).

A complex systems perspective can be applied to grading systems to categorize practices according to their alignment with a common standard, internal consistency, and overall coherence (see Figure 1). In a correlated system, all the parts influence each other. For example, a district-wide grading policy requires all teachers' behaviors to be outlined so that all can understand. The system depends on clear expectations, schools implementing it at the building level, and teachers at the classroom level. A school district establishes a general framework, but individual schools and teachers act as independent agents within the system. These agents interact with the framework and each other, influencing their interpretations and applications. This interaction leads to variations in grading practices across schools and classrooms. The autonomy granted to teachers at the classroom level further amplifies this complexity.

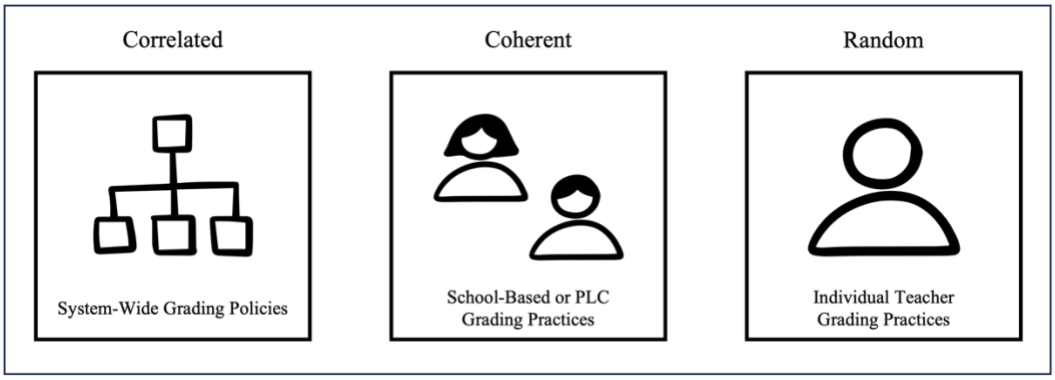


Figure 1

Complex Systems Theory Applied to Grading Practices

School districts establish a general framework similar to the "coherent" system depicted in Figure 1. This framework outlines how grades should be determined and used. However, individual schools and teachers act as independent agents within the system. These agents interact with the framework and with each other in various ways. For instance, schools might develop their own interpretations of the district's guidelines, and teachers might share best practices or attend workshops that influence their grading practices. As Weidman (2017) emphasizes, a key characteristic of complex systems is "the emergence of novel properties from the interaction of the parts" (p. 451). In the context of grading, this translates to the emergence of variations in grading practices across schools and classrooms. The autonomy granted to teachers at the classroom level further amplifies this complexity, as they can adapt the grading system to their specific teaching philosophies and student needs.

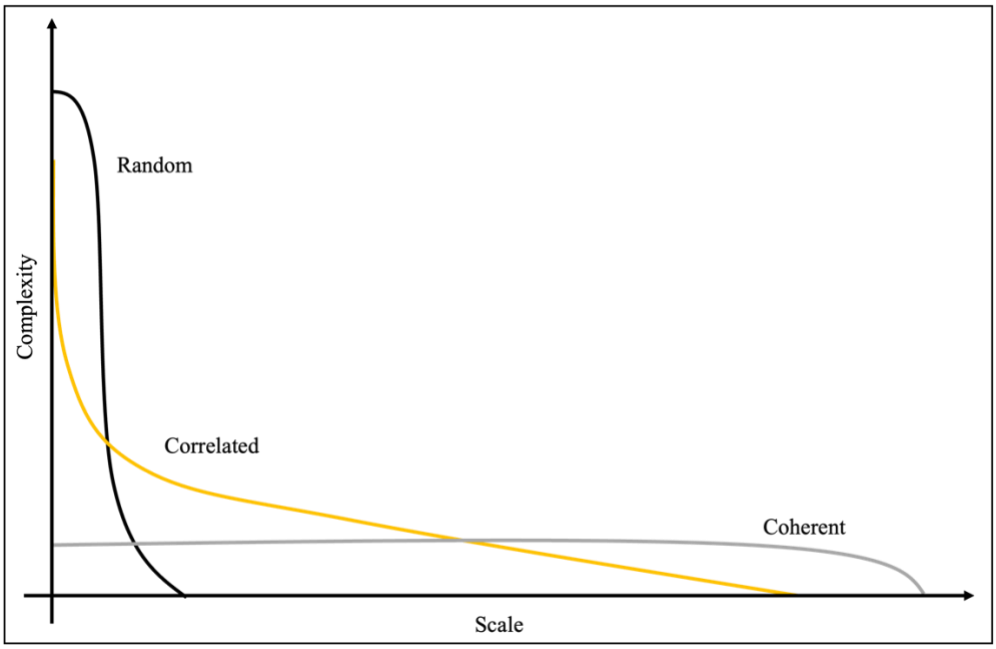


Figure 2

Complexity Profiles for Random, Coherent, and Correlated Systems

In conclusion, applying a Complex Systems Theory lens to grading systems highlights the inherent challenges in achieving uniformity. While a central framework provides a foundation, the system is further shaped by the interactions between schools, teachers, and the broader educational context. This interplay between structure and agency contributes to the emergent complexity observed in grading systems, where seemingly minor variations at the school and teacher level can significantly impact how grades are applied. Understanding these complexities is crucial for designing and implementing effective grading practices that promote student learning and development.

Research Questions

This study will use a quantitative methodology and correlational design to determine the relationship, if any, between the following research questions.

1. What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?
    a. H_0 There is no relationship between teacher perceptions of grading practices and their grade distributions.
    b. H_A There is a relationship between teacher perceptions of grading practices and their grade distributions.
2. What relationship exists, if any, between teacher characteristics and their grade distributions?
    a. H_0 There is no relationship between teacher characteristics and their grade distributions.
    b. H_A There is a relationship between teacher characteristics and their grade distributions.
3. What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based), or hybrid (combination) grading practices and their grade distributions?
    a. H_0 There is no relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.

b. H_A There is a relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.

To explore these connections, this study will use a quantitative methodology and a correlational design. The research questions will center on how different factors influence teachers' grading practices and, in turn, how those practices affect student grades. The first question investigates the relationship between teachers' perceptions of grading and the distribution of grades in their classes. The second question looks at how teacher characteristics influence grading patterns. Finally, the third question examines how the type of grading approach used (traditional, standards-based, or hybrid) relates to student grade distributions.

Limitations of the Study

The limitations of this research study are as follows:

1. The survey instrument was sent to practitioners all over the United States. This response data could vary based on school, district, or state policies.
2. The survey instrument responses will be limited to middle and high school level teachers; therefore, results may not be generalizable to all grade levels K-12.
3. Correlational studies do not imply causation, and while the results may indicate that variables have a relationship, it does not mean that one variable causes a change in another variable (Asamoah, 2014).

Delimitations of the Study

The delimitations used by the researcher were established to allow for a better understanding of the relationship between the elements outlined in the research questions for this study. The researcher only sought classroom teachers at the middle and high school levels to answer the questions provided in the instrument. The instrument will allow for other school-based personnel who may respond to have their data disregarded.

Assumptions of the Study

The assumptions of this research study are as follows:

1. Grades are only one indicator for stakeholders to decide a student's academic performance.
2. Grades are composed of many factors, including elements that may not measure a student's academic performance.
3. Most stakeholders value grades.
4. Practitioners who participate in this study want their students to be successful.
5. Practitioners participating in this study may need more professional learning to receive effective grading practices.
6. Practitioners often need more guidance from the school, district, or state to implement their grading systems.

Organization of the Study

This research study has been organized into five chapters. Chapter One gives an overview of the statement of the problem, the purpose of the study, and the significance of the study as it

relates to the field of education. The theoretical framework, Complex Systems Theory, also outlines the lens through which one should read this study. Also covered in Chapter One are the research questions, limitations, delimitations, and assumptions. Chapter Two reviews the literature relevant to this study, which was used to guide the work outlined therewithin. The literature review includes the historical context of grading, research-based effective grading practices, research on grading perceptions, and why grade distributions matter. Chapter Three describes the methodology, including the research design, instrumentation, data collection, data analysis, and validation and credibility. A discussion of the results from the methodology will be included in Chapter Four, with Chapter Five concluding with the findings from the study and how one can further the research from these results.

Summary

Grading reform efforts aim to move beyond simply informing teachers about effective grading practices. By understanding how teacher perceptions of grading influence student achievement (as measured by grade distributions), this research seeks to develop targeted professional learning opportunities. Ultimately, these efforts aim to create a more consistent and equitable grading system that fosters deeper student learning, not just reflects grades.

CHAPTER 2
REVIEW OF THE LITERATURE

As educators grapple with the ever-increasing demands of their profession, a growing body of research sheds light on the complexities of grading practices. This research highlights how these complexities can create significant burdens for educators. The literature is organized into four parts: (a) historical context of grading in the United States, (b) research-based effective grading practices, (c) research on perceptions of grading, and (d) why grade distributions matter. The historical context of grading is essential to understanding our grading practices in the United States and where they go. Exploring research around effective grading practices is critical to understanding the benefits for teachers and students. A cornerstone of effective pedagogy, assessment practices, particularly grading, have garnered significant scholarly attention in recent years. This surge in research reflects a growing recognition of the complexities inherent in grading practices and how these complexities can pose substantial challenges for educators. Lastly, grade distribution will be discussed, including the research on the normal distribution, standards-based grading systems, and the role of grading policies in shaping grade distributions. This study examines the relationships between teacher perceptions of effective grading, teacher characteristics, grading systems used, and student grade distributions.

Historical Context of Grading Inconsistencies

A persistent challenge in contemporary education is the enduring reliance on standardized assessments and their corresponding grading practices. This has resulted in a concerning lack of change in student grades over the past two decades, suggesting a potential disconnect between these assessments and actual learning outcomes. According to Guskey

(2002), the overreliance on narrow assessments, such as multiple-choice tests, restricts the curriculum and limits student learning opportunities. These assessments often focus on rote memorization and fail to measure a student's ability to apply knowledge, solve problems, or think critically. A curriculum solely focused on preparing students for such tests may not equip them with the well-rounded skills they need for college, careers, and life.

Edgeworth (1888) discussed the consequences of various issues, such as grading inconsistencies, focusing on awarding levels of distinction to graduates and other critical academic merits. Edgeworth's research occurred around the same time many of the nation's most prestigious institutions began ranking student performance.

As early as the late 1700s, Yale University began to rank student performance using four categories, leading to the grade point average scale still in use today. In the late 1800s, Harvard University started classifying students into five grading tiers, with the lowest failing the course (Durm, 1993). However, in a seminal study, Starch and Elliot (1912) discussed issues in their findings regarding the validity of marks being assigned by teachers to indicate a student's academic performance on an outcome. The outcomes of the Starch and Elliot study damaged the use of percentage scores on a 100-point scale that there was a shift in the education field to move toward a 5-point scale determined by the masses to be considered "fairer" (Vatterott, 2015). However, Brimi (2011) replicated the research results from the Starch and Elliot study found almost 100 years later. There was virtually no difference between the variability seen in Starch and Elliot and that of Brimi 100 years later.

Our grading and assessment systems have transitioned from meaningful feedback systems in the 19th century, where student progress was commonly presented to parents orally

during a home visit. However, these oral reports of student progress soon made their way to more standardized forms of grade reporting to manage the increasing demand for evaluating student progress (Guskey & Bailey, 2001). This point is illustrated by Kirschenbaum et al. (1971), where research findings reported that the number of high schools in the United States grew from 500 in 1870 to 10,000 in 1910. In addition, the number of students in public schools increased from 6.8 million to 17.8 million during the same period.

Elementary schools were able to maintain more narrative descriptions of student learning; however, high schools needed to shift to percentage scores because narrative grading practices were deemed to be too time-consuming. This shift in practice eliminated the specific information that families could use to determine what their students knew and could do (Farr, 2000).

By the mid-1900s, more than 80% of schools had begun to adopt the A-F grading scale, the system most commonly used today (Grindberg, 2014). Most recently, there has been a shift to grades based on standards from achievement or standards-based grading (SBG). The SBG methodologies ensure that academic achievement is reported separately from work habits and other nonachievement factors (Guskey & Bailey, 2010).

The historical issues presented regarding grading practices outlined in this literature review were disrupted when the COVID-19 pandemic hit in 2020. “The onset of the COVID-19 pandemic last spring forced schools to pivot to a virtual learning model, which immediately exposed the vulnerabilities of traditional grading and reporting models” (Guskey, 2021, para. 1). As calls for grading reform intensify, the need to develop and implement more meaningful assessment practices tailored to the virtual learning environment becomes increasingly critical.

Table 1

Key Studies on Inconsistencies

Study	Sample	Findings	Method
Brimi (2011)	Seventy-three high school English teachers scored one paper after receiving training on grading writing.	Range of 46 points	Descriptive statistics
Starch & Elliot (1912)	One-hundred and forty-two English teachers scored two papers using procedures from their own schools.	Range of 34 points on paper A, range of 49 points on paper B	Descriptive statistics

Research reveals significant inconsistencies in grading practices throughout history, making it difficult to compare grades across time directly. Standards for what constitutes an "A" or a "B" have demonstrably changed. As Pauline Cahen (2019) points out, "[historical studies of student performance on standardized tests show] a clear upward trend in scores over time, even when accounting for changes in the tests themselves" (p. 12). This suggests that grades reflect not just a student's absolute knowledge or skills but their performance relative to the expectations of their specific era. Grading has also always been impacted by subjectivity. James Popham (2001) highlights the inherent biases that teacher judgment can introduce and how factors like race, gender, or socioeconomic background can distort the picture of student achievement. The methods used to assess learning have also evolved. Standardized tests have become more prevalent, while alternative assessments like portfolios are gaining traction. This shift reflects a continuous debate about the most effective ways to measure student learning (Stiggins & Chappuis, 2005). Understanding these historical inconsistencies is crucial when interpreting

grades and emphasizes the need for ongoing efforts to create a fairer and more objective assessment system.

Research-Based Effective Grading Practices

Research-based effective grading practices and their use in schools and classrooms will be essential to improving more accurate student grades. To empower teachers to make grades more meaningful, we need to leverage research on assessment reliability and validity.

Effective grading practices go beyond simply assigning a letter or number and instead serve as a tool to enhance learning. As highlighted by Marzano (2007), research emphasizes clear learning goals that emphasize their role in uniting curriculum, instruction, and assessment. As O'Connor (2007) suggests, this feedback should be specific and actionable to guide students toward improvement. Instead of averaging past performance, focusing on a student's current understanding encourages a growth mindset (Association for Supervision and Curriculum Development, 2012). Furthermore, research by Edutopia (2023) suggests moving away from traditional averaging practices that can obscure the accurate picture of a student's learning journey. Finally, effective grading considers individual needs through differentiated assessments, as Strobel Education (2023) suggested, to cater to various learning styles and abilities. By incorporating these elements, grading can foster a learning-centered environment where students are empowered to improve their understanding of the subject matter.

Reliability Issues

Grading reliability assumes that students with the same academic performance or proficiency levels should receive the same grade. Consistency is called into question, “The

reliability of an assessment refers to the consistency of results. The most basic interpretation generally references test-retest reliability, characterized by the replicability of results" (Importance of Validity and Reliability, 2023). To ensure that grades accurately reflect student achievement and allow for meaningful comparisons among peers, grading systems should be designed to promote consistency and reliability.

In efforts to improve the accuracy and fairness of grading, some educators are exploring alternative assessment methods that move beyond traditional A-through-F letter grades. Guskey (2019) found that using the 100-point scale can cause probable errors of plus/minus up to six points in either direction. Having teachers attempting to differentiate between a 92 and an 83 on a 100-point scale leaves room for interpretation (Brimi, 2011; Guskey, 2019; Link, 2023). This is further emphasized in Reeves' research, "To insist on the use of a zero on a 100-point scale is to assert that work that is not turned in deserves a penalty that is many times more severe" (Reeves, 2004, p.325). These items are not addressed in our schools of teacher education to the depth that they are needed, and there are severe consequences to a student's grade when a teacher with little to no understanding is assessing work (Reeves, 2004).

Even so, the research suggests teachers can improve the reliability of their students' grades by using as many sources of information demonstrating student academic work as possible. For example, Haladyna (2019) stated that if you only have one data source, you will introduce random error to the equation, which will impact the reliability of the student's grade. In addition to adding multiple measures of student academic performance, teachers can bring additional specificity to their grading practices. Creating instructions that provide specific steps to complete the assessment, writing questions that are tightly aligned with the content that has

been taught, and leveraging colleagues to give feedback on the questions being asked can help to bring additional reliability to the assessments that teachers write. (Schillingburg, 2016).

From Bloom's work (1968), education research has been clear that students need to know what is expected from them to meet academic learning goals. Teachers must develop targets for learning and ensure that those targets are transparent to the students. In addition, the feedback students receive needs to guide them toward learning (Guskey, 2022). Hattie and Timperley (2007) added that grades often lack feedback and need to provide information in the form of feed-forward, where students know what improvements to make to demonstrate their learning.

However, this work is difficult for teachers to execute. Still, the research is clear that defining criteria for what we need from students is essential in improving reliability in grading practice (Liu, 2008). Even though the work is rigorous, the more teachers can do to clarify grading criteria for their students before an assessment is given, the more impact this will have on learning outcomes (Brookhart et al., 2020; Brimi, 2011).

Guskey (2019) argues that clarity of criteria is now an essential foundation for good assessment but that identifying and describing clear criteria remains a challenge for many teachers.

> Assessment theory has made great strides in the last century, and clarity of criteria is now an essential foundation for good assessment. However, our work with teachers suggests that identifying and describing clear criteria is one of the most challenging things teachers strive to do. (Guskey, 2019, para. 11)

Zhang (2023) examined the rater variability and reliability of constructed response questions in the New York high-stakes English and mathematics test, highlighting that single-

rater scores can lead to unreliable results even in high-stakes assessments that have adhered to rigid design standards.

Bias impacts grade reliability, “we find evidence of a persistent racial bias in teachers’ grading, with Black students receiving lower grades than White students with similar achievement levels” (Ho et al., 2018, p. 175). We see the same with gender in Carothers’ (2017) research, “across all subjects and grade levels, teachers reported holding higher expectations for boys’ mathematics performance and provided boys with more positive feedback in mathematics than girls” (p. 142). Feldman (2018) discusses the role of unconscious bias in grading, “unconscious bias can have a significant impact on how teachers perceive and interact with students, and it can also influence grading decisions (p. 23).

We must design professional learning programs and grading policies that support teachers in developing specific criteria that allow students to know precisely what is needed from them on assessments. If we can better understand teachers' perceptions of grading practices, we can develop professional learning programs that better meet their specific needs in developing more reliable assessments.

Validity Issues

Haladyna’s (2019) research defines validity as a truthful and accurate grade representation. Ensuring that a grade is a valid measure of student learning is essential in grading practice.

However, there are many threats to validity when discussing grading practices (Liu, 2008). Misrepresentation is common in K-12 grading and assessment practices. Misrepresentation occurs when a learning outcome benefits from a more complex approach to

assessment, but the instructor uses a more straightforward form of evaluation, such as multiple-choice (Haladyna, 2019). For example, relying on multiple-choice exams to assess scientific knowledge can be a form of misrepresentation. This approach doesn't fully capture a student's understanding. A more effective way to assess learning would be through performance-based assessments (Haladyna, 2019).

Teacher education programs may place a greater emphasis on curriculum and instruction, leaving less room for in-depth training on assessment practices. "Most university teacher-education programs focus on curriculum and instruction, with less attention given to assessment" (Link, 2023, para.10). Teachers spend about a third of their careers engaging in grading and assessment work; however, the focus of teacher preparation programs is on curriculum and instruction (Link, 2023).

Teacher perceptions of grading practices may be linked to student grade distributions. Teacher preparation programs may need to be redesigned to focus more on grading and assessment practices while also informing school districts about how to best prepare their current teaching staff for grading and assessment work (Will, 2021).

In their research, Reeves and Feldman (2020) have identified that teachers must define the purpose of their grades. We can inform our professional learning systems to support teachers and shift toward research-based best practices (e.g., clear descriptions of learning, inaccuracy of averaging, and identifying deficiencies) for grading and assessment (Reeves & Feldman, 2020).

The validity issues in grading practice are further outlined by Guskey and Link (2018), who found high school teachers gave more weight to "major compositions and examinations, laboratory projects and homework," and elementary teachers gave more weight to "formative

assessments, exhibits of student work, and classroom observations" (p.308). The authors further expressed concern over formative assessment being relied on for grading instead of measuring student progress during the learning process. Guskey and Link (2018) express concern in their writing that if homework grades where a student performs poorly are averaged with a final exam score where the student does well, this can bring down the overall student grade even though the student demonstrated mastery. Guskey (2017) also addresses the disconnect between curriculum and standardized tests, "the pressure of high-stakes testing can pressure teachers to narrow the curriculum, with teachers focusing on the specific skills and knowledge that will be tests on standardized tests" (p. 42).

Research also questions grades as a good predictor of college success, "High school grade point average (GPA) remains the single best predictor of freshmen year college grades, even when controlling for the effects of standardized test scores" (Impedovo et al., 2001, p.32).

Fairness Issues

Are the current grading systems that are widely used in schools fair? Fairness is relevant when we examine the body of research on what measures make up a student's grade. While lesson planning and pedagogical strategies are essential, "better grading can provide clearer and more accurate information on students' learning that can be used as a basis for making improvements" (Link & Guskey, 2022, p. 408). Developing the targets that students are to demonstrate is essential to better understanding the purpose of grades within a school or district.

The purpose of grades in schools is a topic of debate. Some argue that grades are a necessary tool for measuring student achievement, while others believe that they can be counterproductive.

> For grades to be assigned reasonably, schools must agree on their purpose. We see the primary purpose of grades as communicating current student achievement to whomever has the need and right to know that information, including students. This means that all students must have equal opportunity to learn and to show what they know, understand, and can do (O'Connor et al., 2018, para.5).

The research suggests that to implement grading practices effectively; schools need to have a shared understanding of the purpose of grades and how they should be used. This can be achieved through collaborative work and professional development focused on grading practices (O'Connor et al., 2018). Setting a mission and vision that guide the practice is effective in implementing grading. When guiding principles are in place, and teachers are supported through professional learning, the impact can be powerful (O'Connor et al., 2018).

One significant concern lies in the subjectivity inherent to traditional grading. Research by the National School Boards Association (2020) highlights how factors like participation can disadvantage certain groups. The study found that conventional grading "can exacerbate existing achievement gaps and contribute to a system that is not equitable for all students" (National School Boards Association, 2020). This is because socioeconomic background or cultural differences can influence participation styles, unfairly impacting a student's grade.

These disparities in grading practices are further compounded by biases that educators may unconsciously hold. According to research by the American Psychological Association (2019), implicit biases can significantly influence teachers' perceptions of student performance and behavior. For example, students from minority backgrounds may be perceived as less engaged or disruptive, which can result in lower participation grades and an overall reduction in

their academic evaluations. This phenomenon is particularly troubling as it not only skews academic records but also affects students' self-esteem and motivation. The study emphasizes that "addressing implicit biases in educational settings is crucial to fostering an equitable learning environment for all students" (American Psychological Association, 2019).

Moreover, the traditional grading system often fails to account for the diverse learning styles and strengths that students bring to the classroom. A report by the Education Trust (2021) underscores the importance of developing more inclusive assessment methods. The report advocates for a shift towards competency-based assessments, which allow students to demonstrate their knowledge and skills in various ways rather than being limited to standardized tests and conventional participation metrics. The report states, "Competency-based assessments provide a more comprehensive and equitable measure of student achievement by recognizing the diverse ways in which students learn and demonstrate understanding" (Education Trust, 2021). By embracing such alternative assessment strategies, educators can better support all students, ensuring that grades reflect true learning rather than superficial indicators of participation and compliance.

Efforts to address this concern have led to exploring "equitable grading" practices. However, critics argue that these methods, which may de-emphasize points and averages, can have unintended consequences. Opponents express concern that a focus on participation over mastery can lead to grade inflation and a need for clear expectations for students. This suggests that removing subjective elements is not the answer, and focusing on proper understanding is crucial (National School Boards Association, 2020).

Studies by Guskey (2018) advocate for clear and well-defined grading criteria and rubrics. When teachers have a shared understanding of what constitutes good work, grading becomes fairer and less subjective. “When grading criteria are clear and specific, teachers are less likely to be influenced by personal biases or idiosyncratic preferences when assigning grades (Guskey, 2018, p.48).

Fairness also comes up in the equity debate. According to Patterson (2003), “Fair does not mean equal; yet, when it comes to grading, we insist that it does” (p. 572). Treating all students the same when assigning due dates and expecting all students to learn at the same pace does not meet the needs of all learners (O’Connor et al., 2018). Other researchers have noted the potential for similar disparity. Barberis and Buchowicz (2015) highlight the role of educators in shaping educational opportunities. They argue that educators' "discretionary power...including personal theories of justice and fairness" (p. 63) can influence access to further educational opportunities. This is particularly concerning when considering grading practices, which can be subjective and vary greatly between educators.

McNeil (2000) explores how high-stakes testing, a practice often heavily reliant on grades, can disadvantage students from low-income backgrounds. She argues, "The tests themselves are biased against low-income students and students of color, in part because these students are more likely to have had unequal educational experiences…" (McNeil, 2000, p. 188). This highlights the importance of clear and consistent grading practices within schools, ensuring educators are empowered to assess student achievement fairly and not perpetuate inequalities.

There are severe consequences to grading practices that could be deemed *unfair*. In Baltimore City Schools, a lawsuit alleged that the school did not review students' transcripts and

provide the necessary interventions for students to graduate on time (Bowie, 2021). The basis for the case was that unfair grading practices were limiting student access to academic opportunities. Student families were not able to determine, based on student grades, where the student was academically, and when it came time for graduation, many students that the school still needed to meet the graduation requirements. This case resulted in multiple suspensions of staff members within the school district (Bowie, 2021).

Another case brought suit against Fayette County Public Schools over the fairness of grades assigned in a virtual class during the Pandemic (Six, 2022). When we consider the consequences of grades that students are assigned, the seriousness of considering whether the system as it is designed is fair is all too real in our work. As previously discussed, major decisions regarding a student's academic progress are grounded in these measures.

Considering the multifaceted purposes of grades, the importance of a shared understanding among educators about how grades should be used is essential.

> Grades are the primary basis for making important decisions about students. They determine whether students are promoted from one grade level to the next. They also assess honor roll status and enrollment in advanced or remedial classes, and they factor into special education services and college or university admissions. (Link, 2023, para. 3)

Given the multiple roles that grades play, such as determining course placement and college admissions, the importance of educators having a shared understanding of how grades should be used is critical. Over 1,800 colleges and universities are now "test-optional," putting an incredible weight on a student's grade point average (GPA) composed of individual teacher

grades (Jimenez, 2023). Post-pandemic changes will significantly affect student outcomes for college admissions, especially if grading practices for K-12 are in question (Jimenez, 2023).

The issue of fairness continues as the educational system recovers from the pandemic. Some students lived in homes where virtual learning could be a reality; however, some went to homes where their learning stood still (Reeves & Feldman, 2020). Instead of relying solely on grades, teachers can use a variety of assessments to get a more accurate picture of student learning (Reeves & Feldman, 2020).

The common thread throughout this literature review is the continued need to better support and train our teachers in effective grading and assessment practices. In conclusion, the research underscores the need for a multifaceted approach to grading that prioritizes fairness without compromising academic rigor. Educators are tasked with developing practices that assess student learning accurately, provide growth opportunities, and remain unbiased regardless of a student's background.

Research on Professional Learning

Current research emphasizes the importance of well-designed professional learning programs in equipping teachers with effective grading practices. These programs should move beyond simply teaching teachers "how to grade" and instead focus on fostering a deeper understanding of the purpose and complexities involved in assessment.

One key aspect of effective professional learning is the emphasis on clear learning goals. As Guskey (2018) states, "Teachers need to understand the specific learning goals they want students to achieve before they can design effective assessments" (p. 42). Professional learning

programs that involve collaborative development of learning objectives and rubrics can be particularly beneficial. As advocated by Arter and McGaw (2013), this collaborative approach allows teachers to share best practices and develop a shared understanding of assessment expectations. One study by Arter and McGaw (2013) found that "when teachers worked together to develop clear learning targets and well-designed rubrics, they were more likely to make consistent judgments about student performance" (p. 218). This collaborative approach promotes consistency and fosters a sense of ownership and accountability among teachers.

Effective professional learning programs also delve into the issue of bias in grading. Research by the National School Boards Association (2020) highlights how traditional grading methods can disadvantage students from specific backgrounds. Professional learning programs can equip teachers with strategies to mitigate bias, such as utilizing anonymous grading or focusing on specific criteria outlined in rubrics. By acknowledging the potential for bias and providing tools for its reduction, professional learning programs can promote fairer and more equitable grading practices.

Furthermore, research suggests the importance of ongoing support and follow-up. Effective PL extends beyond a one-time workshop. Providing opportunities for teachers to implement new strategies in their classrooms, share experiences with colleagues, and receive ongoing coaching can significantly enhance learning and implementation.

Current research underscores the critical role of well-designed PL programs in fostering effective grading practices. By emphasizing clear learning goals, collaborative rubric development, strategies to reduce bias, and ongoing support, PL programs can empower teachers to utilize grading as a powerful tool to improve student learning and growth.

Research on Perceptions of Grading

The research has shown significant differences in perceptions based on the level of grade that a teacher instructs. Link (2018) found variability between the zero for incomplete work at the elementary level and those teachers in the middle and high school levels. This research is vital in developing the survey instrumentation so that additional items can be added to differentiate the results further.

Like others that have focused on teacher perceptions, this study aims to inform the development of more effective professional learning and teacher preparation programs, which will bring greater clarity to our practice.

Brookhart (2016) found that most teachers are not well-trained in grading and assessment practices, leading to very little use of research-based best practices in grading.

> Without improved assessment training and a better understanding of teachers' beliefs about grading and its role in student success, school leaders and policymakers may continue to have limited knowledge about the challenges current grading practices pose. Some of these grading struggles may significantly affect the quality of teaching and learning offered in K-12 schools and may ultimately prevent many students from attaining their educational goals (Link, 2018).

Guskey (2018) takes it a step further, stating, "Teachers who participated in professional development programs focused on effective grading practices reported feeling more confident in their grading decisions" (p.78). Brookhart (1991) discusses that this lack of training results in a hodgepodge effect, with grades comprising many factors. In other words, a grade may be representative of factors beyond academics. Improving our training through professional learning

and teacher preparation programs will be essential to remedy many of the grading issues in our K-12 schools today. The instrument used in this study, the Teacher Perceptions of Grading Scale, disaggregates data in five areas: Importance, Usefulness, Student Effort, Student Ability, and Teacher grading Habits.

Importance and Usefulness

According to Elkins (2016), most research around learning begins with Bloom and his taxonomy, which builds toward higher-level cognitive operations. However, Bloom also published work around his findings that starting all students at the same point and expecting them to arrive at their learning simultaneously needed to be revised. Guskey (2007) discussed how Bloom stated that learning needs to be designed around the individual student to close learning gaps for students. This system requires additional time for some students to master their learning. Our grading and assessment systems often do not consider this type of differentiation. When we move all students along in their learning, the outcomes are bleak, as highlighted by Arum and Roska (2011), they found college students had little to no learning take place from the start of their first year to the end of their sophomore year, "We observe no statistically significant gains in critical thinking, complex reasoning, and writing skills for at least 45 percent of the students in our study" (Arum & Roska, 2011, p. 36). Brookhart's (2017) research discussed a focus on learning, "Many teachers reported using grades not only to evaluate learning, but also to motivate students and communicate learning expectations" (p.21).

Grading is intended to assess where a student is concerning achieving proficiency on an academic outcome or standard. "Teachers often expressed concerns about the inherent subjectivity of grading, particularly in open-ended assessments" (McMillian, 2018, p.242). This

is why learning and learning alignment matter, “When the stated criteria are inconsistent with how the grading is done, this affects the learning-teaching process since the assessment is sending out a different message regarding what is important to learn” (Svennberg et al., 2014).

Proponents of grades argue that they provide valuable feedback and motivation for students. A study by the University of Michigan (Michaelsen, 2020) found that 80% of students surveyed based their self-worth on academic performance. Good grades can boost confidence and a sense of accomplishment, encouraging students to continue striving. Additionally, grades are often a factor in college admissions and scholarship opportunities (ASCD, 2012). As ASCD (2012) states, "Grades can also be a factor for consideration in an honor society" (p. 6). This can give students access to further education and potentially more significant career options.

However, critics argue that grades can have negative consequences. A study by Marco Learning (n.d.) highlights concerns that a focus on grades can overshadow the true purpose of education – learning itself. Students fixated on achieving high marks may prioritize memorization and test-taking strategies over a genuine understanding of the material. As the study states, "most student evaluations are composed of both evaluative feedback, which judges student work, and descriptive feedback, which provides information about how a student can become more competent" (Marco Learning, n.d., para. 3). Ideally, grades should complement descriptive feedback to guide improvement, not solely measure achievement.

Current research suggests that grades play a complex role in education. While they can motivate students and provide a benchmark for achievement, they should not be the sole focus of learning (Marco Learning, n.d.). Effective educators strive to utilize grades alongside other

assessment methods to create a holistic picture of student progress and foster a love of learning beyond the classroom (ASCD, 2012).

Student Effort and Ability

The relationship between student effort, ability, and grading practices is a complex topic with ongoing research. While traditional methods often aim to reflect effort and ability in a single letter or number, current research suggests a more nuanced approach may be necessary.

One concern is the potential for bias in traditional grading. Subjective elements like participation can disadvantage students from specific backgrounds, as highlighted in a study by the National School Boards Association (2020). The study found that "students of color and students with disabilities are more likely to receive lower grades in courses that emphasize subjective criteria" (National School Boards Association, 2020, p. 4). This raises questions about whether grades accurately represent a student's understanding of the material or are influenced by factors beyond their control.

Alternative grading practices, such as standards-based grading, aim to address these concerns by focusing on mastery of specific learning objectives. However, critics argue that these methods may need to account for individual student differences sufficiently. A study by the Center for Applied Special Technology (CAST, 2018) emphasizes the importance of considering factors like learning styles and disabilities when assessing student effort. As CAST (2018) states, "Universal design for assessment (UDL) ensures that all students have the opportunity to demonstrate their knowledge and skills" (p. 2). This suggests that a one-size-fits-all grading approach may not be practical for all learners.

The research around effort and ability further emphasizes the “hodgepodge” effect of grading student outcomes. Stiggins, Frisbie, and Griswold (1989) found that 50% of teachers considered student ability when assigning a grade, and 86% looked at a student’s effort in assigning a grade to a student's work. Cross and Frary (1999) launched a large-scale study examining over 7,000 middle and high school teachers in a single school system. The study compared actual grading practice with teacher perceptions of grading. The study reported that 25% of teachers indicated that they raised student scores based on the student’s efforts (Cross & Frary, 1999).

This pattern in the literature remains consistent with Brookhart (1994), suggesting that teachers use a variety of factors when grading. A later study by McMillan and Nash (2000) found that teachers viewed effort as increasing achievement, leading to higher marks for non-academic measures. The same was true in their study of elementary school teachers, who found that academic achievement was the most critical factor in determining a student's grade. Still, the majority included behavior, effort, participation, and extra credit in the grade.

All the studies found in the literature remained consistent in that teachers use non-academic measures, which include effort and ability, extra credit, ability to work in a group, attendance, behavior, and conduct, to determine a student's grade. Understanding these factors is essential to understanding their influence on a teacher’s grading habits.

In conclusion, current research underscores the need for grading practices that acknowledge the interplay between student effort and ability while minimizing bias. Educators are called upon to develop assessment methods that are clear and fair and provide opportunities for all students to demonstrate their understanding of the material.

Teacher Grading Habits

Studies reveal that teacher grading can be influenced by many factors beyond a student's understanding of the material. A study by Brookhart (2017) identified factors such as student behavior, prior performance, and even neatness as potential influences on grades. The study found that "teachers are more likely to give higher grades to students who exhibit positive behaviors and to lower grades for students who exhibit disruptive behaviors" (Brookhart, 2017, p. 12). This raises concerns about how much grades reflect learning and the potential for bias within the grading process.

However, research also suggests that many teachers actively seek ways to improve their grading practices. A study by Arter and McGaw (2013) found that professional development programs focused on clear learning objectives and rubrics can lead to more consistent and reliable grading. As Arter and McGaw (2013) state, "When teachers use clear learning targets and well-designed rubrics, they are more likely to make consistent judgments about student performance" (p. 218). This highlights the importance of ongoing professional development in fostering effective grading practices. Historically, teachers have rewarded good students with good grades and punished bad students with bad grades. Teachers see grades as controlling classroom behavior (Kohn, 1999).

Some contend that grades have such a negative impact on students that they should be abolished altogether. “In a world overflowing with information, our schools should focus on developing a love of learning and the ability to think critically. Grades, however, narrowly define success and squelch both curiosity and passion. It’s time to ditch grades and embrace a system that fosters deep understanding and a lifelong love of learning” (Vatterott, 2015). Teachers

should use feedback as the driving force behind learning, and summative assessment should be eliminated (Barnes, 2018; Kohn, 1994; Spencer, 2017). Guskey (2019) states that there are a few essentials that must be considered regarding the effects of grades versus comments:

1. What is the nature of the assessment (e.g., multiple-choice, performance-based)?
2. What is the content area?
3. What is the age or grade level of the students?
4. What are the student's previous academic backgrounds and experiences?
5. What is the student's economic background?
6. What is the nature of the comments to be given as feedback?
7. What does the student believe about failure (self-efficacy)?
8. What is the interaction between the comment and the grade?

These questions are only sometimes addressed in a school of teacher education when discussing grading and assessment practices (Link, 2023).

Other teacher perceptions about students' ability to re-assess were addressed in a recent Education Week poll, where *the chance to redo assignments* ranked 11 out of 24 for teachers polled. However, the same poll given to students ranked the redo as the top choice for motivating them to improve their learning (Heubeck, 2023). The argument against this process of reassessing work is that it is time-consuming. Re-learning takes time (Heubeck, 2023).

Re-assessment has a tremendous impact on student learning, yet the battle continues. Those in favor argue for reducing test anxiety and allowing students to demonstrate learning on a more fluid timeline. This is juxtaposed by arguments for not motivating students and developing poor academic habits (Guskey, 2023).

According to Brookhart, Guskey, McTighe, and William (2020), teachers' beliefs about learning, assessment, and the purpose of grading guide their grading practices. Attempting to change grading practices does not change the underlying assumptions of the practitioner (Brookhart et al., 2020). Our professional learning must support teachers in understanding the intent of the grade, "Beginning at the earliest levels, teachers must help students and their families understand that grades do not reflect who you are as a learner, but where you are in your learning journey" (Guskey, 2022). When families and students see that grades reflect their current academic performance, they understand that knowing where they are in their learning is essential to their academic success (Guskey, 2022).

Then there is workload: "Teachers reported feeling pressure to complete grading quickly due to large class sizes and limited planning time, which could impact the quality of the feedback provided" (Wiliam and Black, 1988, p.8). Brookhart (2013) echoes the stresses of grading: "Research suggests that the pressure of grading can contribute to teacher stress and burnout" (p.43).

In conclusion, research on teacher grading habits paints a multifaceted picture. While factors beyond learning can influence grades, there's also a growing movement towards more objective and standards-based assessment. By implementing clear learning goals, utilizing rubrics, and engaging in professional development, educators can continue to refine their grading practices to ensure they accurately reflect student learning and promote a fair and equitable learning environment.

Research on Grade-Level Differences in Grading Practices

Research suggests that grading practices can differ significantly based on the grade level a teacher instructs. Elementary educators often emphasize formative assessment and provide students with descriptive feedback to guide improvement. A study by McMillan (2018) found that elementary teachers "frequently used formative assessment practices like exit tickets and observations to monitor student learning and adjust instruction accordingly" (p. 123). This focus on formative assessment allows teachers to identify student needs early and provide targeted support.

However, as students progress through the grades, grading practices tend to shift towards a more summative approach, relying on points and percentages to represent achievement. A study by Popham (2001) observed that "high school teachers were more likely to utilize traditional grading methods that emphasize summative assessments like tests and projects to determine final grades" (p. 42). This may be due to increased pressure on standardized testing and college admissions at higher grade levels.

While there are valid reasons for these differences, some researchers advocate for a more balanced approach across all grade levels. By incorporating formative and summative assessment elements throughout a student's educational journey, educators can provide a more comprehensive picture of learning and development.

Research on the Role of Teacher Preparation Programs on Grading Practices

Research offers a mixed perspective on the role of teacher preparation programs in shaping effective grading practices. While some studies highlight gaps in current training, others

acknowledge positive developments in equipping future educators with essential assessment skills.

A potential concern lies in the adequacy of training provided by teacher preparation programs. A study by Hemmeter et al. (2008) surveying faculty members from institutions with teacher preparation programs revealed concerns that graduates often need more training in utilizing evidence-based assessment practices, including effective grading methods. The study found that faculty members reported graduates "often struggled to design assessments that accurately measured student learning objectives" (Hemmeter et al., 2008, p. 142). This suggests a potential need for more comprehensive training within teacher preparation programs.

There are also positive developments in this area. Research by the National Council on Teacher Quality (NCTQ, 2020) found that many teacher preparation programs are incorporating performance-based assessments into their curriculum. This approach, as the NCTQ (2020) states, "allows teacher candidates to demonstrate their ability to use assessments to inform instruction and improve student learning" (p. 12). These programs provide future educators with hands-on experience designing and utilizing practical assessments, including fair and informative grading practices.

Research on Grading Methodologies

Traditional grading typically relies on a point-based system, where assignments and tests contribute to a final letter grade. This approach can be subjective, as highlighted by Guskey (2018), who argues that traditional grading often "blurs the lines between effort, behavior, and actual learning" (p. 48). Students may receive lower grades due to factors outside their control,

such as participation or neatness, which may not accurately reflect their understanding of the material.

In contrast, standards-based grading focuses on mastery of clearly defined learning standards. Students receive grades based on their ability to demonstrate proficiency in specific skills or knowledge areas, often reported via letter grades alongside descriptive feedback. This approach, as advocated by Marzano and Kendall (2007), allows for "greater clarity about learning expectations and promotes a growth mindset among students" (p. 12). Students can see areas for improvement and have opportunities to re-demonstrate mastery before a final grade is assigned.

Research also acknowledges the potential drawbacks of standards-based grading. Critics argue that a de-emphasis on points and averages can lead to grade inflation and a lack of clear communication regarding student progress (Guskey, 2018). Opponents also express concern that focusing solely on mastery can create a binary system where students are labeled "mastered" or "not mastered," potentially overlooking the learning process.

Why Grade Distributions Matter

Traditionally, a normal distribution (bell curve) was considered ideal, with most students clustered around the average grade and fewer receiving very high or low marks. “Many teachers and administrators still hold the misconception that grades should be distributed in a normal curve” (Angela, 1993, p.132). Research by Benjamin Bloom (1981) challenged the normal curve, arguing that with practical teaching, most students can achieve high levels of learning. “If grading reflects true learning, then the normal curve is not an appropriate model for grades”

(Bloom, 1981, p. 4). Current research emphasizes the importance of clear learning objectives, well-designed assessments, and practical feedback to promote student learning rather than adhering to a specific grade distribution. “The focus should be on using grades to communicate effectively with students about their progress and to guide future instruction” (Guskey, 2018, p.62).

Black and Latinx students underperform white students, and in the literature, this is attributed to the Pygmalion Effect (Gleeson, 2019). This is the *belief*, consciously or unconsciously, that a student will perform poorly. The students inherit these beliefs about themselves, lowering academic performance (Gleeson, 2019). In his research, Gershenson (2015) discusses findings from studies that indicate that non-black teachers have lower educational expectations for black students. This emphasizes the need for further research on how teachers' perceptions could impact their grade distributions (Gershenson, 2015).

The pandemic has impacted K-12 education in ways that cannot be entirely determined. There are learning deficits from the shift to online schooling, discussed in EdWeek (2020b), where 95% of schools closed their doors and shifted to online learning. Not all students had access to the tools and support needed to thrive in this setting. Research suggests that the COVID-19 pandemic significantly negatively impacted student grades. Students are making slower academic progress compared to pre-pandemic years. This learning gap is particularly concerning in reading and mathematics, where students fell behind by an estimated 4.1 and 4.5 months, respectively (Education Week Research Center, 2023). Another study from the Brookings Institution (2023) highlights a troubling trend. The research indicates that the opportunity to learn gap between students from low-poverty and high-poverty schools widened

significantly during the pandemic, particularly in mathematics. Research suggests that effective strategies include extended learning opportunities, high-quality tutoring, and a focus on foundational skills to get students back on track (Education Week Research Center, 2023).

Another fallout from the pandemic has been a shift away from SAT and ACT scores in college admissions. With the change to "test-optional," colleges and universities rely on the high school grade point average, which is of higher importance (Jimenez, 2023). Furthermore, Hodara and Cox (2016) found that in a review of 57 community colleges, 33% of students required remedial English and 59% remedial mathematics.

With the combination of eased grading practices during the pandemic (EdWeek, 2020a) and the reliability of those grades, we are admitting students on the pretense of their academic readiness in college and university.

This study attempts to determine whether grade distributions are impacted by a teacher's perceptions of grading practices, helping practitioners determine whether teachers using effective grading practices impact grade distributions positively or negatively. The teacher characteristics are of interest as the research has shown that grading inflation occurs at higher rates at schools with higher levels of affluence (Gershenson, 2020). This study attempts to use data, such as school affluence, to determine if there are higher grade distributions across teachers in schools with similar characteristics. With already outlined issues with grading practices grounded in the research, is that exacerbated even more for students who are in underserved areas? We also see students that are in urban areas being influenced by factors that are not focused on academic performance, "Teacher perception of urban high school students being prepared for class, arriving on time and submitting homework assignments have been found to better correlate with

grade point average than overall academic performance" (Steward et al., 2008). We are reminded by this section of the literature review that we consistently see the need to provide a system of effective grading practice supported by professional learning in schools and teacher preparation programs.

Summary of the Literature Review

Throughout the review of the literature, grading and assessment practices in schools have been, and continue to be, a significant issue of reliability, validity, and fairness when it comes to understanding a student's academic proficiency. The literature will help to guide the answers to the research questions outlined in Chapter 1. The literature will inform this researcher's theory that a teacher's perceptions of research-based effective grading practices impact the distribution of their students' final grades. If this is the case, we must use the data from this study to inform our preparation of professional learning programs to support teachers in implementing more effective grading and assessment practices.

CHAPTER 3
METHODOLOGY

Introduction

Grading and assessment practices and reform efforts are complex (Guskey, 2022; Guskey & Bailey, 2001; Kalgorides et al., 2013; Link & Guskey, 2022; Reeves, 2011). This gap in the research and the need for more consensus among scholars indicates that there is work to be done on teacher perceptions of grading practices and their impact on student grade outcomes.

This study aims to unpack information that could assist practitioners and policymakers in better understanding the relationship between teacher perceptions of grading practices and the impact on student grade distributions. The study also used a survey instrument to examine teachers' perceptions of grading in five domains, including (1) Importance, (2) Usefulness, (3) Student Effort, (4) Student Ability, and (5) Perceived Self-Efficacy of the Grading Process. A supplemental domain examined self-reported grade distribution sampling of final grades from an academic quarter. Grade distributions were self-reported by respondents for one class period or subject from one academic quarter of the school year. The respondent reported this as the total number of As, Bs, Cs, Ds, and Fs for their students' final grades for the marking period. The researcher then calculated an average using a traditional grade-point-average calculation of A=4, B=3, C=2, D=1, and F=0.

This study and its research questions were to add to the body of research in grading and assessment practices. This study employed a quantitative research design to determine if there were statistically significant differences between a teacher's perceptions of grading practices and their grade distribution; if there were statistically significant differences between teacher

characteristics (i.e., grade level taught, enrollment in a teacher preparation program vs. no enrollment in a teacher preparation program) and their grade distribution; and if there were statistically significant differences between the teacher grading model (i.e., traditional, standards-based, hybrid) and their grade distribution.

Research Design

This quantitative method and correlational design study aims to determine the degree and direction of the relationships between variables. In this case, it will identify relationships, if any, between teachers' perceptions of grading practices and the student grade distributions of those teachers. In addition, the study seeks correlations between teacher characteristics and the grade distributions of students. Teacher characteristics included grade level taught and enrollment in a teacher preparation program vs. non-enrollment in such a program. In addition, the grading system the teacher used was examined for a relationship to grade distribution.

First, a descriptive analysis was conducted to exhibit mean grade distribution scores and survey perception scores in each of the five domains previously outlined for each teacher based on grade level taught, enrollment in a teacher preparation program, and the grading model used to report student grades.

To examine the relationship between the Teacher Perceptions of Grading Practices Survey (TPGPS) domain score and their grade distribution, the researcher calculated a Pearson *r* correlation. The continuous variable for the Pearson r correlation was the grade distribution Average. To attain the grade distribution average, the researcher collected self-reported final letter grades for students in one course for one academic quarter to supplement the TPGPS

instrument. These letter grades were converted to a traditional grade-point average score (e.g., A = 4, B = 3, C = 2, D = 1, and F = 0). A grade distribution average was then calculated by totaling the letter grade sum and dividing it by the total number of grades reported. The TPGPS is included in Appendix A. All other relationships (e.g., grade level taught and grade distribution average, teacher preparation program enrollment and grade distribution average, and grading practices and grade distribution average) were examined using a one-way ANOVA.

A survey design was adopted to gather descriptive and comparative data to describe the teacher participants in the study. Descriptive research is foundational when there are correlations to be studied (Lobo, 2005). Participants were given the TPGPS instrument, validated in previous studies (Liu, 2004; Liu, O'Connell, & McCoach, 2006, 2008). Written permission was received from Dr. Liu to use and revise the survey instrument (see Appendix B).

Research Questions

Research questions 1 through 3 were developed to guide this study:

1. What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?
2. What relationship exists, if any, between teacher characteristics and their grade distributions?
3. What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based), hybrid (combination) grading practices, and their grade distributions?

Population and Sampling

This study used voluntary response sampling to capture as extensive a sampling as possible from teachers across the United States. Limitations of previous studies have been tied to the small sample size of respondents and not providing results representative enough of the education system. Recruitment to research has traditionally been challenging and often results in a small sample for a particular study. This will slow the progress of valuable research and impact the outcomes (Wertheimer, 2013). To increase the sample size for this current study, social media platforms were leveraged to target survey respondents and cast a broader representation and sampling of data. The connectedness of the social media platforms serves as a benefit for survey distribution, "In many cases, these networks can be accessed with relative ease, particularly when an initial participant was recruited using social media (but even if not), and networked individuals may share characteristics relevant for study eligibility" (Gelinas et al., 2017, p. 14).

As of the 2021-22 school year, the last year data was reported, there were 3.6 million teachers in the United States (National Center for Education Statistics, 2024). With a sample size such as this, G*Power suggests a sample size of 112 participants to determine a Pearson's correlation coefficient of r = .30 with 90% power (alpha = .05, one-tailed). For the ANOVA, a population of 3.6 million and a confidence level of 95% with an interval of +/- 5%, G*Power suggested a sample size of 84 for two groups and 102 for three groups to examine significant effects. A decision was made only to examine large effects as the required sample to examine medium and small effects would not have been possible within the examined population. G*Power confirmatory analyses for this study can be found in Appendices C - E.

This study used a self-selection sampling methodology. As previously stated, the TPGPS instrument was published online, and participants could choose to participate in the data collection. While the sample size of collected data was small (n = 165), it meets the parameters for sample size outlined in the previous paragraph.

Instrumentation and Validity

An online survey called the TPGPS was adopted to examine relationships between teacher responses and the grade distributions of their students. “This instrument was designed to measure teachers’ perceptions of the importance and usefulness of different grading practices, their perceptions of student effort and ability and the influence these have on grading practices, and teachers’ personal grading habits” (Liu et. al., 2006, p. 2).

The survey instrument consisted of Likert Scale-style questions, fill-in-the-blank, checkbox, and multiple-choice questions and was administered online using Qualtrics in September 2023. The instrument asked participants to rate their answers using a rating scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree). In addition, behavioral questions were included, asking teachers to consider factors used in assigning final grades.

An additional question regarding the self-reporting of grade distributions for a marking period was added to the instrument. This question and the instrument were piloted with five teachers not included in the survey to consider the flow and comprehension of what was being asked.

This study is built on the Complex Theory Framework, meaning that the complexities of grading practices are composed of many individual variables that cannot be considered in isolation. The Complex Theory Framework calls for the data to be considered based on the interaction with other data. The study is designed to enable this research to analyze multiple data sets and their interactions with one another. For example, the study may produce data on teacher characteristics and how they correlate to grade distributions.

Lastly, a section was added to the survey instrument requiring participants to self-report a snapshot of a grade distribution from a previous academic term. It should be noted that the study relies on participants to provide this data using the survey instrument.

The TPGPS has been developed and validated using exploratory factor analysis and confirmatory factor analysis (Liu et al., 2006).

In multivariate statistics, exploratory factor analysis (EFA) is a valuable tool for revealing the hidden patterns within many variables. This statistical method falls under factor analysis and aims to unveil the fundamental connections between the observed variables. Researchers often rely on EFA while constructing scales, like survey scales that consist of questions used to assess specific research subjects. Doing so effectively pinpoints a group of underlying, unobservable constructs that influence the battery of measured variables (Liu et al., 2006).

Moreover, the TPGPS instrument underwent validation through confirmatory factor analysis (CFA). Confirmatory factor analysis is a specialized type frequently applied in social research, particularly education. Its primary purpose is to examine whether the measures of a specific construct align with the researcher's conceptual understanding of that construct or factor.

Confirmatory factor analysis evaluates whether the collected data adhere to a predetermined measurement model (Liu et al., 2006).

Liu (2006) also conducted a reliability analysis on the pilot survey results to validate the instrument further.

> The reliability analysis showed that all factors except factor two (Importance) had a reliability coefficient alpha larger than .7 was used as the minimum acceptance level. The collection of items for factor two needed to include three more items to potentially reach a level of .8 reliability coefficient (See below). Although four constructs of this instrument were initially hypothesized, the exploratory factor analysis (EFA) split the usefulness and importance of grading into two parts, which indicated that this domain was multi-dimensional. Thus, it was more appropriate to treat them as two factors. The usefulness and importance of grading are separate interpretable constructs along the perceptions of the grading practice continuum. The pilot validation study revealed that a five-factor solution better interprets the data than the hypothesized four constructs. (Liu, 2006, p. 68)

With the ease of use of the TPGPS, the goal was to distribute the instrument to as many teachers as possible using the reach of social media platforms.

Data Collection Procedures

The study was approved by the Institutional Review Board (IRB) in fulfillment of the University of Central Florida's protocol. The IRB approved the study on November 28, 2023. On November 29, 2023, the survey was distributed as planned using various social media platforms,

including Facebook, Instagram, LinkedIn, and X, formerly known as Twitter. The survey remained open for responses through December 22, 2023. Approval and closure confirmation from IRB can be found in Appendix F and Appendix G.

The survey was posted to multiple social media platforms and targeted teachers in the United States. An identifying question was added to the study to identify respondents not practicing in the United States. The survey was posted to the LinkedIn social media platform, and the researcher leveraged several professional connections to share their reach in the educational profession. In addition, the Facebook social media platform was used to target several closed Facebook Groups targeting teachers, which included Teacher Professional Development, Teacher Goals: Connected Schools, Teacher's Forum, Teachers Sharing Ideas and Resources, Teachers Throwing Out Grades, Standards Based Learning and Grading, For Teachers by Teachers, The Best Teacher Group in the World, and The Principal's Desk. Lastly, X, formerly known as Twitter social media platform, targeted teachers using hashtags such as #edchat, #teachers, #edreform, and #cpchat. Social media posts were added weekly for one month to all platforms to encourage increased survey participation.

After the data collection period closed, data was downloaded from Qualtrics and cleaned to remove any respondent data that was not complete and did not answer all the questions asked. Any respondents who identified as not teaching in the United States were also removed from the data.

Data Analysis

The researcher used a parametric statistical analysis to investigate the three research questions that guided this study. Confidential Teacher Perceptions of Grading Practices Survey scores, teacher characteristics data, and grade distributions were collected through a Qualtrics survey and analyzed using the IBM Statistical Package for Social Science (SPSS) program.

The items in the TPGPS are divided into five domains: Importance, Usefulness, Student Effort, Student Ability, and Teachers' Grading Habits. A sixth domain requires respondents to self-report the grade distributions for one of their courses by providing the total number of grades from A to F.

This study's quantitative method and design aligned with the research questions listed in Table 2. The TPGPS collected all the quantitative data. The survey was divided into six sections aligned to the research questions.

Table 2

Summary of Research Questions and Analyses

Research Question	Variables	Analysis
What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?	IV: TPGPS Survey Domain Scores DV: Grade Distribution Average	Pearson Correlation
What relationship exists, if any, between teacher characteristics and their grade distributions?	IV: Grade Level Taught IV: Enrollment in Teacher Preparation Program DV: Grade Distribution Average	One-Way ANOVA
What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based) hybrid (combination) grading practices, and their grade distributions?	IV: Teacher Grading Practices DV: Grade Distribution Average	One-Way ANOVA

Research Question 1

What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?

The analysis of research question 1 examined the domain scores of the TPGPS survey. An average domain score was calculated for each of the five domains in the TPGPS survey: (1) Importance, (2) Usefulness, (3) Student Effort, (4) Student Ability, and (5) Perceived Self-Efficacy of the Grading Process. In addition, the grade distribution was converted into a grade-point-average style score for each participant, as previously discussed. The researcher considered the grade distribution score to be the dependent variable throughout. The TPGPS survey domain scores were the independent variables for this analysis. A Pearson *r*-correlation was run to determine whether a relationship existed between the continuous dependent variable of the grade distribution score and the independent variables in the form of the five domain scores from the TPGPS. The Pearson correlation measures the strength and direction of association between variables and can provide data that helps determine the association's strength (Laerd Statistics, 2018).

Meeting Statistical Assumptions for Research Question One

The Pearson *r*-correlation has three major assumptions: that there needs to be a linear relationship between two variables, that there are no significant outliers, and that there is bivariate normality (Laerd Statistics, 2018). The relationship between each TPGPS domain score and the grade distribution average was tested for linearity. If no linear relationship was found, Spearman's rank-order correlation was run. Pearson's correlation analysis also requires checking for outliers in the data. Outliers are data points that do not fit the pattern of the data set (Laerd

Statistics, 2018). Analysis and tests were run with both outliers included and outliers altered if they were found. Both variables were tested for bivariate normality using the Shapiro-Wilk test for normality. The results of these analyses are reported in Chapter Four.

Research Question 2

What relationship exists, if any, between teacher characteristics and their grade distributions?

For research question 2, the data were analyzed with a one-way ANOVA to determine a relationship, if any, between the multinomial independent variable of the grade level taught (e.g., elementary school, middle school, high school), the dichotomous independent variable of enrollment in a teacher preparation program and the continuous dependent variable of the grade distribution score. The one-way ANOVA determines whether there are any statistically significant differences between two or more unrelated groups (Laerd Statistics, 2017a). The dichotomous independent variable of enrollment in a teacher preparation program and the continuous dependent variable of the grade distribution score.

Meeting Statistical Assumptions for Research Question Two

The one-way ANOVA has three major assumptions that must be met. There are no significant outliers, the customarily distributed dependent variable, and the homogeneity of variance with the dependent variable (Laerd Statistics, 2017a). Boxplots were examined to test for outliers. If outliers were detected, tests were run with both the outliners included and altered. Any tests that were used are included in the analysis in chapter four. In addition, Shapiro-Wilk's was used to analyze the data to meet the assumption of normality, and the results are included in

the analysis in chapter four. Levene's Test of Equality of Error Variances was used to analyze the homogeneity of the variances, and any violations are reported in chapter four.

Research Question 3

What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based), or hybrid (combination) grading practices and their grade distributions?

For research question 3, a one-way ANOVA was again used to examine a relationship, if any, between the multinomial independent variable of the grade level taught (e.g., elementary school, middle school, high school) and the continuous dependent variable of the grade distribution score.

Meeting Statistical Assumptions for Research Question Three

The one-way ANOVA has three major assumptions that must be met: that there are no significant outliers, that the dependent variable is normally distributed, and that variance is homogeneous with the dependent variable (Laerd Statistics, 2017b). Boxplots were examined to test for outliers. If outliers were detected, tests were run with both the outliners included and altered. Any tests that were used are included in the analysis in chapter four. In addition, Shapiro-Wilk's was used to analyze the data to meet the assumption of normality, and the results are included in the analysis in chapter four. Levene's Test of Equality of Error Variances was used to analyze the homogeneity of the variances, and any violations are reported in chapter four.

Summary

This study aimed to investigate the relationship, if any, between multiple factors, including teacher perception of grading practices, grade level taught, teacher preparation program completion, and grading methodology and grade distribution. In this chapter, the researcher detailed the research design methodology, including the selection of population and sampling, instrumentation and validity, and analysis procedures. Included in this chapter were reliability and validity data for the instrument, the Teacher Perceptions of Grading Practices Survey. The statistical analysis was also explained, including identifying the variables. The results of the statistical analysis are presented in Chapter Four.

CHAPTER 4
PRESENTATION AND ANALYSIS OF DATA

Introduction

This study was conducted to examine if there is a relationship between teacher perceptions of grading practices, teacher grade level taught, teacher enrollment in formal preparation programs, and grading system and the grade distribution of the students in their course. Data were collected using the Teacher Perceptions of Grading Practices Survey (TPGPS) instrument. Additional data concerning teacher characteristics were collected from the survey. All data were analyzed to determine if any statistically significant relationships existed between a) the teacher's perceptions of grading practices, b) the teacher's grade level taught, c) the teacher's enrollment in a formal educator preparation program, d) the grading system used by the teacher and the grade distribution of the students in their course.

The TPGPS instrument data was analyzed by downloading the data from Qualtrics and removing any data that was not complete. Teachers not actively teaching in the United States were also removed from the data. For each respondent, a survey domain score was calculated by taking the average of responses in each of the following domains: (1) Importance, (2) Usefulness, (3) Student Effort, (4) Student Ability, and (5) Perceived Self-Efficacy of the Grading Process, resulting in a TPGPS domain score. Teacher grade level taught, completion of a formal educator preparation program, and grading system used in practice were all captured as a part of the TPGPS instrument.

For grade distribution, each respondent was asked to self-report the total number of As, Bs, Cs, Ds, and Fs for students in one class from one academic term. If the teacher reported

grades in another format (e.g., E, I, U for early grades, or 1, 2, 3, 4 for standards-based systems), they were asked to convert them to the letter grade reporting options. Each respondent's grade distribution data was then calculated into an average score using a grade point average calculation scale, where A = 4, B = 3, C = 2, D = 1, and F = 0. The following formula was used to calculate the grade distribution average score: sum of all letter grades / total number of letter grades issued = grade point average score.

Chapter Four includes the results of the data analysis and methodology outlined in Chapter Three. The first part of the chapter exhibits the demographic characteristics of TPGPS respondents from this study. The second part of the chapter exhibits findings arranged by research questions that were answered using four quantifiable statistical techniques: (a) descriptive statistics, (b) Pearson r-correlation tests, and (c) one-way ANOVA tests. Each statistical test met the following assumptions for reliability in results: (a) no outliers were identified that skewed the data results, and (b) the assumption of normality was confirmed by the Shapiro-Wilk's test (Laerd, 2018).

Survey Population and Demographic Information

Demographic Summary

This quantitative study was open to all active teachers in the United States. The design was intended to provide a more comprehensive sampling of respondents with different grading practices as a part of their pedagogy. Previous studies using the TPGPS instrument have been limited to one district or a subset of teachers within a district. A narrow scope of practice could influence respondents and data within *a* district; therefore, the reason for a national survey was

not necessarily to increase respondents but to capture a more extensive set of teachers with varying perceptions and practices.

The survey was open via social media platforms already outlined in the previous chapter and was open to respondents for one month. The study yielded one-hundred and sixty-five (165) respondents who attempted to participate; however, once the data was cleaned for respondents outside of the study criteria and who completed all instrument questions, the data analysis includes the responses of one-hundred and twenty-eight (128) teachers. The first section of the TPGPS instrument included teacher characteristic questions, which included gender, grade level taught, years of experience, Title I funding at their place of employment, type of grading system used in practice, and whether they had completed a traditional university teacher educator program. Table 3 includes the demographic characteristics of respondents.

Table 3

Demographic Characteristics of Respondents (n=128)

Characteristics	*n*	%
Gender		
Female	102	79.7
Male	25	19.5
Prefer Not to Say	1	0.8
Grade Level Taught		
Elementary School (PK-5)	26	20.3
Middle School (6-8)	26	20.3
High School (9-12)	74	59.4
Years of Experience		
1-4 Years	13	10.2
5-9 Years	16	12.5
10-14 Years	34	26.6
15-19 Years	20	15.6
20 + Years	45	35.1
Were you trained in a traditional university teacher training program?		
Yes	93	72.7
No	35	27.3

Characteristics	*n*	%
Does your school receive Title I funding?		
Yes	41	32.0
No	87	68.0
Type of Grading System		
Traditional (e.g., 0-100% Scale)	82	64.1
Standards-Based (e.g., 4, 3, 2, 1)	25	19.5
Hybrid (combination of both)	21	16.4

Most respondents were female (79.7%) and worked at the high school level (59.4%). Most had been enrolled in a traditional university teacher educator program (72.7%), and most identified as using a traditional 100-point grading scale in practice (64.1%).

Quantitative Analysis of Results and Descriptive Statistics

Descriptive statistics, means, standard deviations, and effect size were computed to ascertain the levels of agreement between the teacher characteristics and the responses to each survey item from the TPGPS instrument. Table 4 includes the means and standard deviations for the TPGPS items.

Table 4

Means and Standard Deviations for Teacher Perceptions of Grading Practices Survey Items (Elementary n = 26; Middle n = 26; High n = 74)

Factors and Items	Elementary		Middle		High	
	M	*SD*	*M*	*SD*	*M*	*SD*
Importance						
1. Grading is an important criterion for judging students' progress.	2.37	0.93	1.92	0.56	2.20	0.85
2. Grading has an important role in classroom assessment.	2.19	0.83	1.88	0.59	2.08	0.75
3. Grading has a positive effect on students' academic achievement.	2.56	0.97	2.19	0.69	2.46	0.98
4. Grading practices are important measures of student learning.	2.19	0.68	2.08	0.80	2.25	0.86
5. Grading practices are important measures of student achievement.	2.19	0.68	1.96	0.53	2.24	0.82
6. Grading has a strong impact on students' learning.	2.67	1.07	2.31	0.88	2.40	1.10
Usefulness						
7. Grading helps me categorize students' knowledge/ performance as above average, average, and below average.	2.22	0.64	2.15	0.88	2.18	0.73
8. Grading can help me improve instruction.	2.19	0.83	1.73	0.45	1.89	0.58

Factors and Items	Elementary		Middle		High	
	M	*SD*	*M*	*SD*	*M*	*SD*
9. Grading can encourage good work by students.	2.26	0.76	1.92	0.56	2.00	0.64
10. Grading helps me in deciding what curriculum to cover.	2.85	1.26	2.26	1.17	3.07	1.21
11. Grading is a good method for helping students identify their weaknesses in a content area.	2.26	0.90	1.85	0.37	2.04	0.88
12. Grading can keep students informed about their progress.	1.89	0.32	1.77	0.65	1.82	0.56
13. Grading provides information about student achievement.	2.15	0.72	2.04	0.82	2.15	0.75
14. Grading documents my instructional effectiveness.	2.81	1.00	2.96	1.18	2.97	1.14
15. Grading provides feedback to my students.	1.96	0.52	1.92	0.74	1.84	0.68
16. High grades can motivate students to learn.	2.00	0.48	2.00	0.69	2.08	0.69
17. Grades of zero can demotivate students to learn.	2.04	0.81	2.15	1.12	2.16	1.16
Student Effort						
18. I consider student effort when I grade.	2.56	1.09	2.35	0.98	2.47	1.14
19. I give higher report card grades for students who show greater effort.	3.11	1.12	2.81	1.06	2.80	1.21

Factors and Items	Elementary		Middle		High	
	M	*SD*	*M*	*SD*	*M*	*SD*
20. I will pass a failing student if he or she puts forth effort.	2.85	1.17	2.69	1.09	2.97	1.14
21. Grades are based on students' completion of homework.	3.89	1.12	3.23	1.39	3.45	1.22
22. Grades are based on the degree to which students participate in class.	3.19	1.21	3.50	1.17	3.40	1.21
23. Grades are based on students' improvement.	2.78	1.05	2.27	0.78	2.51	0.92
Student Ability						
24. I consider student ability in grading.	2.44	0.85	2.15	0.88	2.64	1.10
25. Grades are based on students' problem-solving ability.	2.30	0.72	2.15	0.73	2.08	0.63
26. Grades are based on students' critical thinking ability.	2.30	0.72	2.12	0.59	2.07	0.58
27. Grades are based on students' independent thinking ability.	2.33	0.83	2.00	0.49	2.12	0.72
28. Grades are based on students collaborative learning ability.	2.52	0.89	2.65	1.02	2.55	0.96
29. Grades are based on students' writing ability (quality of writing, not handwriting skills).	2.26	0.76	2.50	1.07	2.53	1.09
Teachers' Grading Habits						
30. If a student fails a test, I will offer him or her a second chance to take the test.	2.19	0.96	1.92	1.15	2.05	1.00

Factors and Items		Elementary		Middle		High	
		M	*SD*	*M*	*SD*	*M*	*SD*
31.	If a student fails to complete an assignment, I will assign him or her a grade of zero.	2.63	1.21	2.50	1.36	2.56	1.34
32.	If a student fails to complete an assignment, I will subtract grade points progressively until the assignment is turned in.	3.22	1.34	3.38	1.36	3.48	1.28
33.	I often give students opportunities to earn extra credit.	2.44	1.09	2.92	1.32	3.19	1.32
34.	I often look at the distribution of grades for the whole class after I finish grading.	2.07	0.92	2.08	1.06	1.83	1.06
35.	I have my own grading procedure.	2.44	0.97	2.54	1.14	2.27	1.02
36.	I often confer with my colleagues on grading criteria.	2.33	1.07	2.12	0.77	1.96	0.72

Note. The Likert coding for the TPGPS instrument was as follows: Strongly Agree = 1; Agree = 2; Neutral = 3; Disagree = 4; Strongly Disagree = 5

A Kruskal-Wallis H test was run to determine if there were differences in TPGPS scores between the three grade levels taught by teacher respondents: elementary school (PK-5), middle school (6-8), and high school (9-12). Distributions of the TPGPS scores were similar for all groups in the TPGPS, as assessed by visual inspection of a boxplot. Median TPGPS scores were

not statistically significantly different between groups, except for TPGPS item 2, item 23, and item 33, as highlighted in Table 5.

Table 5

Independent-Samples Kruskal-Wallis Test Summary; Grade Level Taught Statistically Significant Differences Among Groups

Item	Total *n*	Sig.[a,b]
2. Grading has an important role in classroom assessment.	128	.012
23. Grades are based on students' improvement.	128	.014
33. I often give students opportunities to earn extra credit.	128	.020

a. The significance level is .050.

b. Asymptotic significance is displayed.

Descriptive statistics among groupings were examined using IBM SPSS Statistics software, version 29, and are exhibited in Tables 6, Table 7, and Table 8. There were no significant differences in responses to TPGPS items among groups when accounting for grade level taught. As demonstrated by the standard deviations in Table 6, there is good agreement among responses to survey items, regardless of grade level taught.

Table 6

Descriptive Statistics of TPGPS Domain Averages by Grade Level Taught

TPGPS Domains	Grade Levels Taught	Descriptives			
		N	*M*	*SD*	*SE*
Importance Domain	Elementary School (PK-5)	26	2.35	.65	.13
	Middle School (6-8)	26	2.06	.54	.11
	High School (9-12)	74	2.28	.65	.08
Usefulness Domain	Elementary School (PK-5)	26	2.24	.33	.07
	Middle School (6-8)	26	2.10	.40	.08
	High School (9-12)	74	2.20	.40	.05
Student Effort Domain	Elementary School (PK-5)	26	2.99	.70	.14
	Middle School (6-8)	26	2.81	.74	.15
	High School (9-12)	74	2.93	.79	.09
Student Ability Domain	Elementary School (PK-5)	26	2.36	.55	.11
	Middle School (6-8)	26	2.26	.55	.11
	High School (9-12)	74	2.33	.48	.06
Teachers' Grading Habits Domain	Elementary School (PK-5)	26	2.43	.41	.08
	Middle School (6-8)	26	2.50	.46	.09
	High School (9-12)	74	2.48	.42	.05

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation; *SE* refers to standard error.

However, when examining the grouping based on enrollment in a traditional university teacher preparation program vs. respondents who took a non-traditional approach, there is more significant variance in agreement within the group in the TPGPS domain of *Student Effort*, as evidenced by the standard deviations in Table 7.

Table 7

Descriptive Statistics of TPGPS Domain Averages by Completion of Traditional University Teacher Preparation Program vs. Non-Traditional

TPGPS Domains	Grade Levels Taught	Descriptives			
		N	*M*	*SD*	*SE*
Importance Domain	Traditional Teacher Prep	93	2.28	.68	.07
	Non-Traditional Teacher Prep	35	2.16	.47	.08
Usefulness Domain	Traditional Teacher Prep	93	2.19	.42	.04
	Non-Traditional Teacher Prep	35	2.19	.27	.05
Student Effort Domain	Traditional Teacher Prep	93	3.00	.77	.08
	Non-Traditional Teacher Prep	35	2.77	.80	.13
Student Ability Domain	Traditional Teacher Prep	93	2.29	.46	.05
	Non-Traditional Teacher Prep	35	2.41	.61	.10
Teachers' Grading Habits Domain	Traditional Teacher Prep	93	2.50	.43	.04
	Non-Traditional Teacher Prep	35	2.42	.43	.07

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation; *SE* refers to standard error.

In Table 8, descriptive statistics are grouped based on the type of grading system used in practice: traditional 100-point scale, standards-based scale, or a combination of both types of systems. The data also show good agreement among groups in all TPGPS domains.

Based on the descriptive statistical analysis results, there is slight variance, if any, among groups when accounting for grade level taught, teacher preparation completion, or grading system used in practice on any of the domains in the TPGPS instrument.

Table 8

Descriptive Statistics of TPGPS Domain Averages by Type of Grading System Used in Practice

TPGPS Domains	Grade Levels Taught	Descriptives			
		N	*M*	*SD*	*SE*
Importance Domain	Traditional System	82	2.25	.60	.07
	Standards-Based System	25	2.41	.71	.14
	Hybrid: Combination of Both	21	2.03	.63	.14
Usefulness Domain	Traditional System	82	2.22	.37	.04
	Standards-Based System	25	2.18	.43	.09
	Hybrid: Combination of Both	21	2.09	.38	.08
Student Effort Domain	Traditional System	82	2.82	.72	.08
	Standards-Based System	25	3.12	.76	.15
	Hybrid: Combination of Both	21	3.17	.96	.21
Student Ability Domain	Traditional System	82	2.30	.49	.05
	Standards-Based System	25	2.33	.54	.11
	Hybrid: Combination of Both	21	2.37	.53	.12
Teachers' Grading Habits Domain	Traditional System	82	2.41	.37	.04
	Standards-Based System	25	2.61	.49	.10
	Hybrid: Combination of Both	21	2.60	.54	.12

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation; *SE* refers to standard error

Outcomes of Research Questions

Research Question 1 Analysis

RQ1: What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?

a. H_0 There is no relationship between teacher perceptions of grading practices and their grade distributions.
b. H_A There is a relationship between teacher perceptions of grading practices and their grade distributions.

This research question was examined using a Pearson *r*-correlation. A Pearson *r*-correlation requires multiple assumptions to be tested in the data analysis. Pearson's correlation is only appropriate when a linear relationship exists between the two variables being analyzed (Laerd Statistics, 2018). Linearity was confirmed through the examination of a scatterplot of the two variables. Figures 3-7 exhibit the linearity between variables for each TPGPS domain used to examine teacher perceptions of grading practices. The scatterplots also allowed for the identification of outliers within the data. No significant outliers were identified in any of the TPGPS domain analyses.

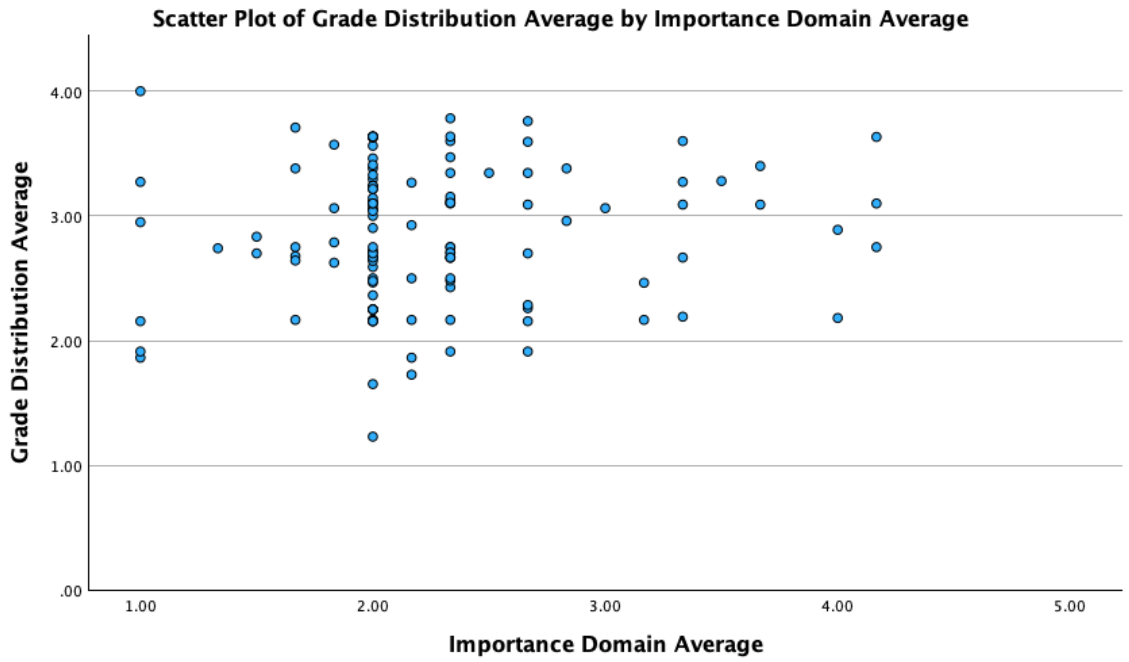


Figure 3

Scatter Plot of Grade Distribution Average by TPGPS Importance Domain Average

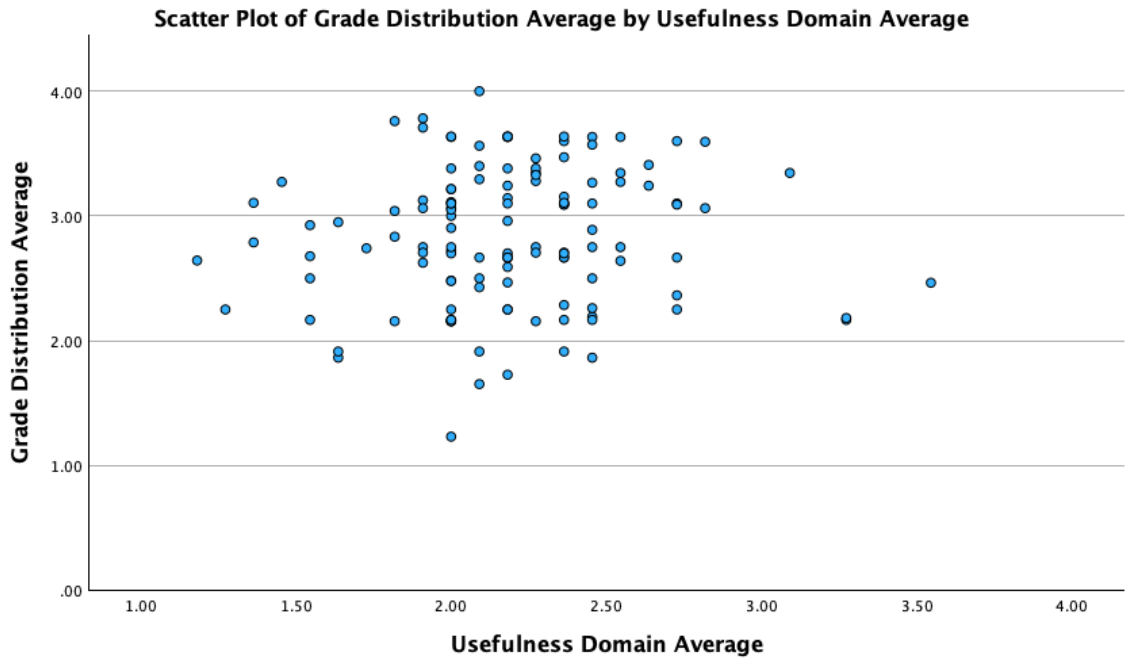


Figure 4

Scatter Plot of Grade Distribution Average by TPGPS Usefulness Domain Average

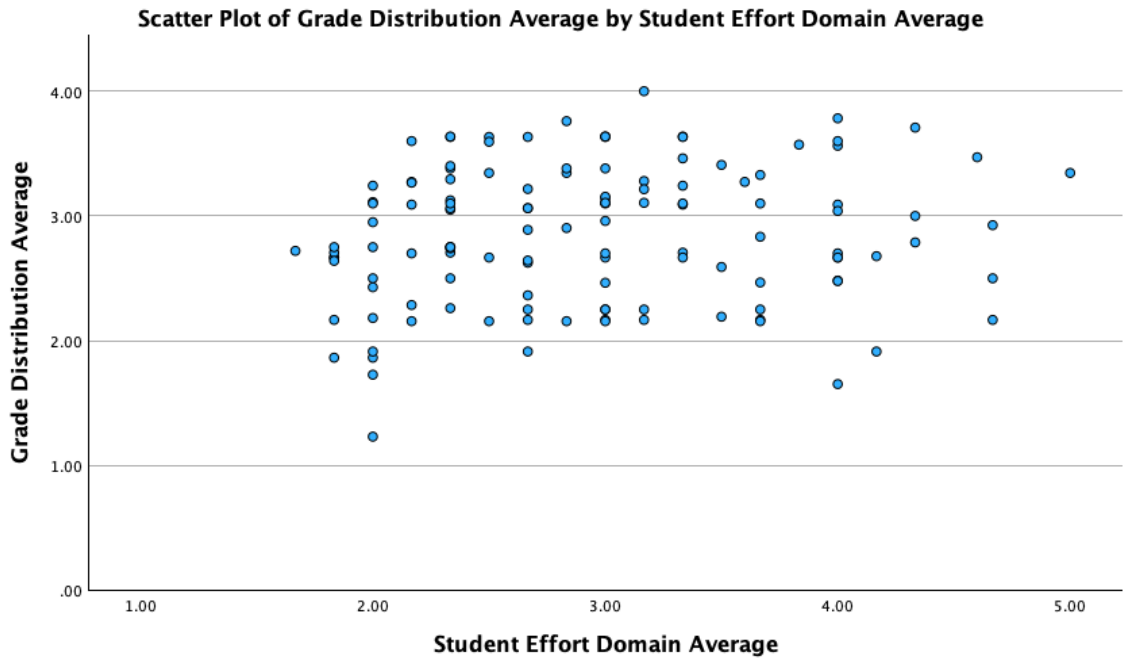


Figure 5

Scatter Plot of Grade Distribution Average by TPGPS Student Effort Domain Average

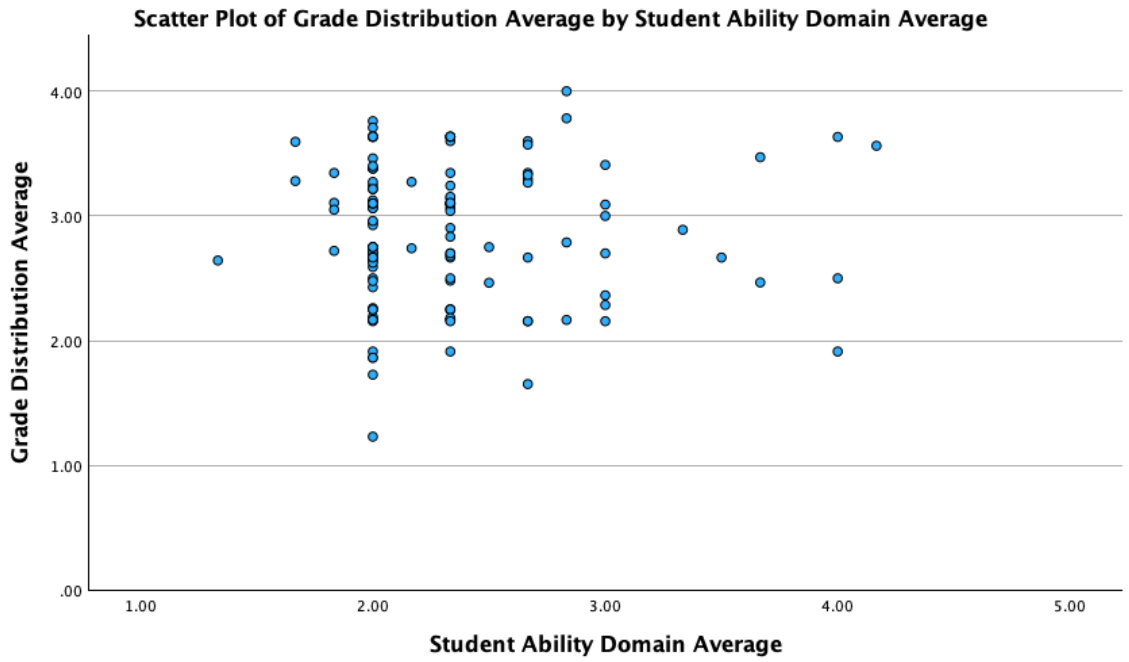


Figure 6

Scatter Plot of Grade Distribution Average by TPGPS Student Ability Domain Average

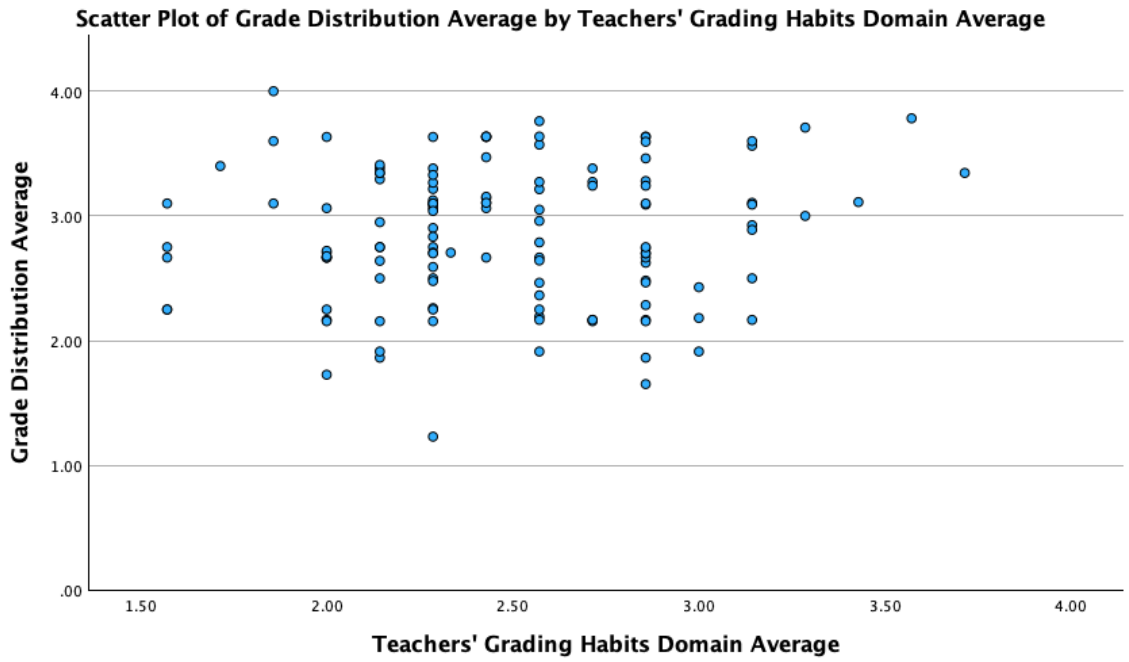


Figure 7

Scatter Plot of Grade Distribution Average by TPGPS Teachers' Grading Habits Domain Average

Domain average scores from the TPGPS and grade distribution scores were linear and normally distributed, as assessed by visual inspection of Normal Q-Q Plots. This meets the assumption of normality for the Pearson *r*-correlation to be valid. Figure 8 and Figure 9 provide examples of normal Q-Q plots, and the tests for normality for all variables followed similar patterns. It should be noted that Laerd Statistics (2018) discusses that larger sample sizes (above 50 cases) can lead to statistically significant results when using the Shapiro-Wilk test for normality. Data could be normal when identified as being statistically significant. Graphical interpretation is what this researcher used and is the preferred methodology (Laerd Statistics, 2018).

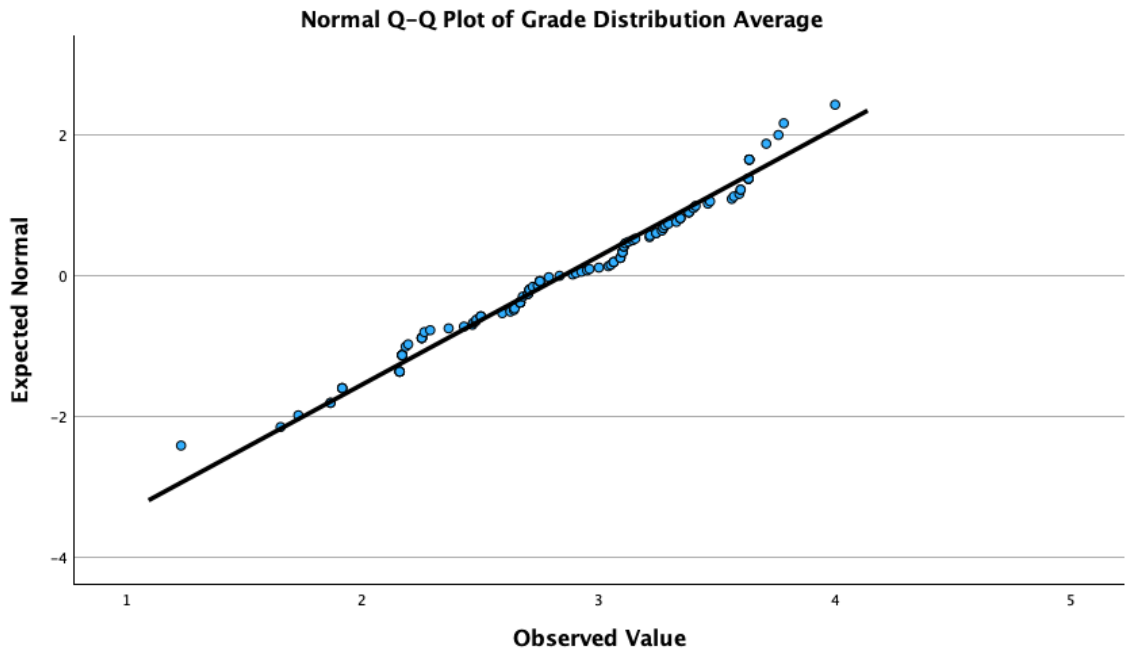


Figure 8

Normal Q-Q Plot of Grade Distribution Average

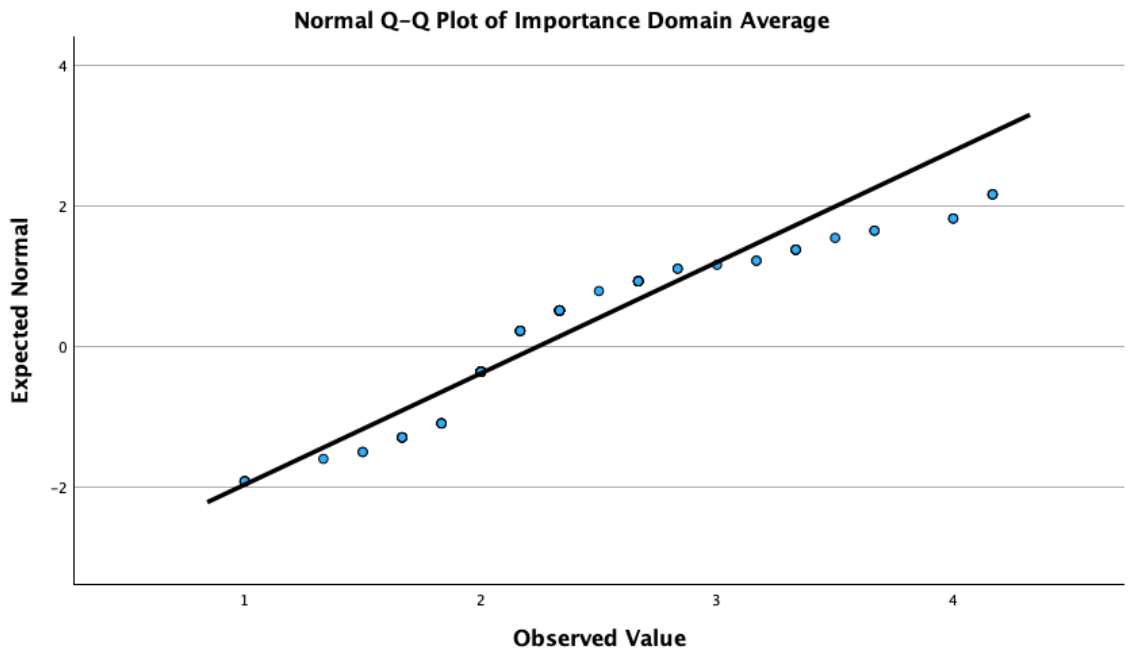


Figure 9

Normal Q-Q Plot of TPGPS Importance Domain Average

A Pearson's product-moment correlation was run to assess the relationship between the TPGPS domain averages of respondents and their grade distribution averages. One hundred and twenty-eight responses were analyzed. Preliminary analyses showed the relationships to be linear, with all variables normally distributed as assessed by the Normal Q-Q scatterplots, and there were no outliers. No statistically significant correlation existed between the teachers' perception of importance, usefulness, student effort, or grading habits and their grade distribution average, as exhibited in Table 9. While a slight positive correlation existed between student ability domain averages and grade distributions (p = .02), the correlation had no statistical significance.

Table 9

Pearson Correlations for Main Study Variables

TPGPS Domains	Correlations	Grade Distribution Average
Importance Domain Average	Pearson Correlation	.083*
	Sig. (2-tailed)	.352
Usefulness Domain Average	Pearson Correlation	.069*
	Sig. (2-tailed)	.442
Student Effort Domain Average	Pearson Correlation	.135*
	Sig. (2-tailed)	.130
Student Ability Domain Average	Pearson Correlation	.018*
	Sig. (2-tailed)	.837
Teachers' Grading Habits Domain Average	Pearson Correlation	.084*
	Sig. (2-tailed)	.344

Note. * = statistically significant at $p < .05$ level.

The relationships between a teacher's perceptions of importance, usefulness, student effort, student ability, or grading habits and their grade distribution average were not statistically significant. Therefore, we cannot reject the null hypothesis or accept the alternative hypothesis.

Research Question 2 Analysis

RQ2: What relationship exists, if any, between teacher characteristics and their grade distributions?

a. H_0 There is no relationship between teacher characteristics and their grade distributions.
b. H_A There is a relationship between teacher characteristics and their grade distributions.

Research question two was analyzed using the one-way ANOVA, which is used to determine whether there are any statistically significant differences between the means of two or more independent groups (Laerd Statistics, 2017b). In the analysis of this question, we are looking for differences between the grade level a teacher teaches in and their grade distribution, as well as differences among those teachers who completed traditional teacher education preparation programs and those who did not and their grade distributions.

The one-way ANOVA has six assumptions that must be met, three of which are based on the study design. The first assumption requires one continuous dependent variable; the grade distribution average meets this assumption. Additional study design assumptions include the inclusion of one independent variable that consists of two or more categorical, independent groups. Both grade level taught and completion of a teacher preparation program, which will be tested independently, meet this assumption. The third assumption requires the independence of observations, and this study meets this requirement as no one participant belongs to multiple tested groups. The additional assumptions required for the one-way ANOVA are as follows:

The one-way ANOVA requires no significant outliers in the independent variable groups (Laerd Statistics, 2017b). The data were analyzed for outliers using a boxplot, as shown in Figures 10 and 11.

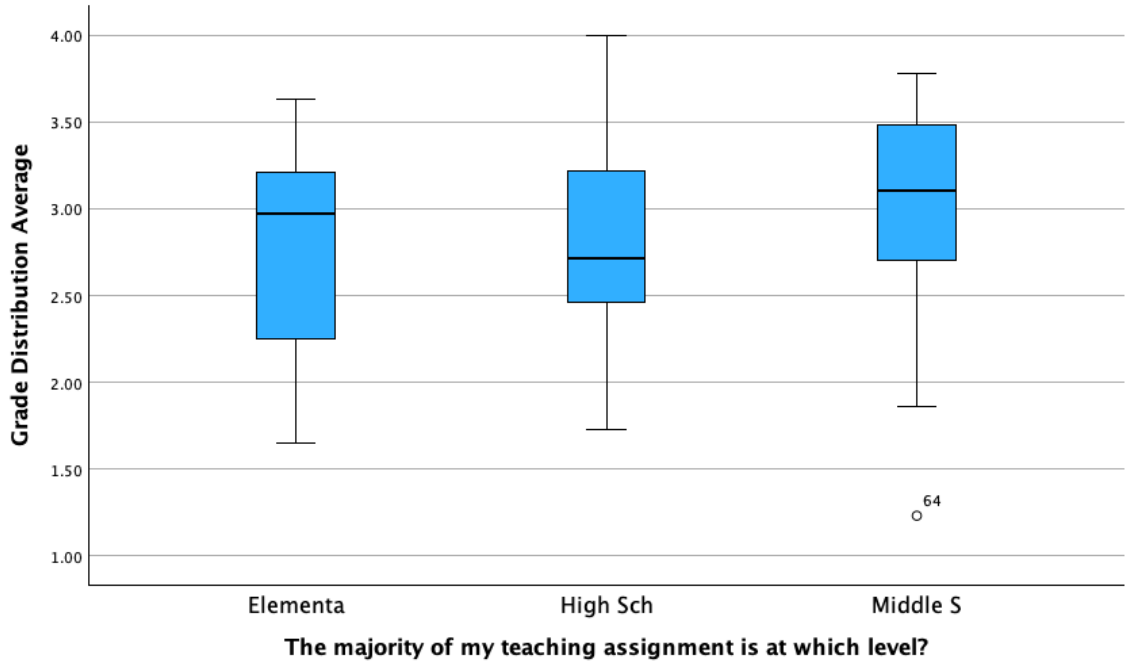


Figure 10

Boxplot Analysis of Grade Level Taught and Grade Distribution Average

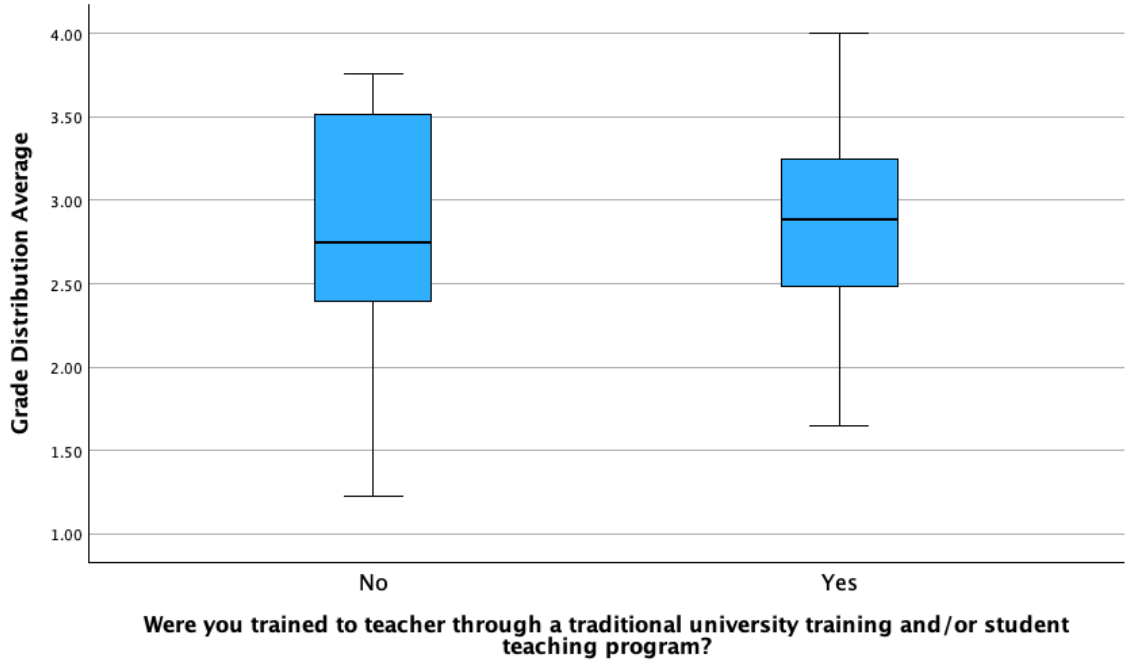


Figure 11

Boxplot Analysis of Teacher Preparation and Grade Distribution Average

Outliers in the data were addressed by running the one-way ANOVA with the outliers included and again without the outliers included to look for variances in the results. The outliers were included in the analysis reported here if no variance was detected. Altering the data by removing the outliers may introduce bias into the analysis (Ghosh & Vogt, 2012).

Testing of normality was also required of the data for the one-way ANOVA. Scores for grade distribution averages were normally distributed for all groups, as assessed by visual inspection of Normal Q-Q Plots, referenced in Figures 12 – 16.

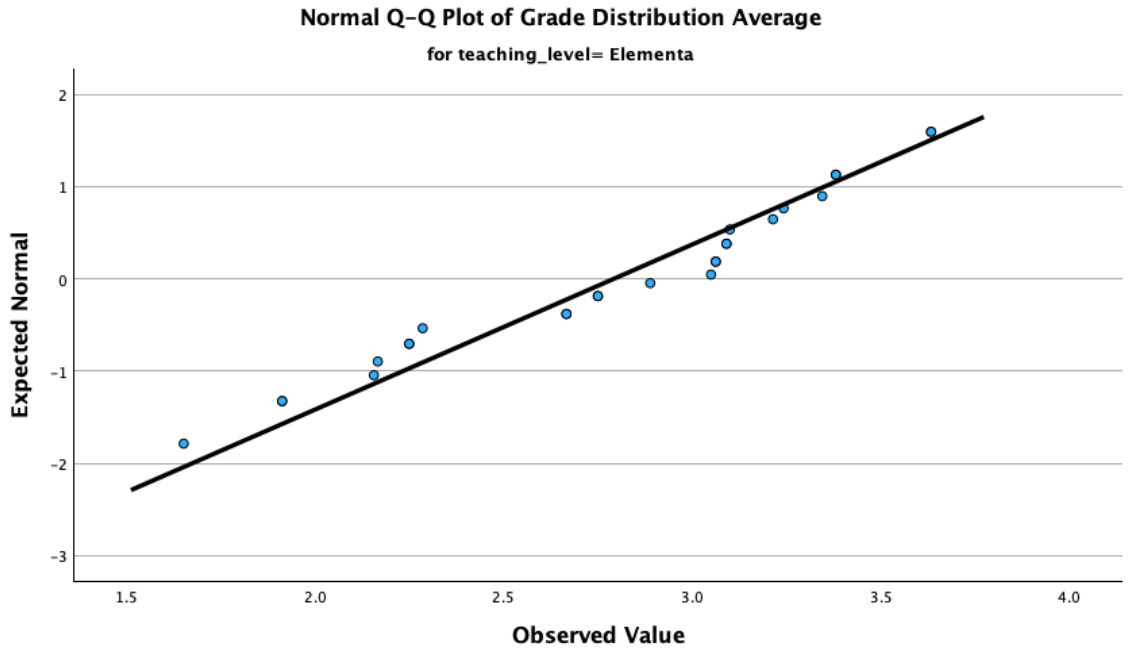


Figure 12

Plot of Grade Distribution Average for Elementary School

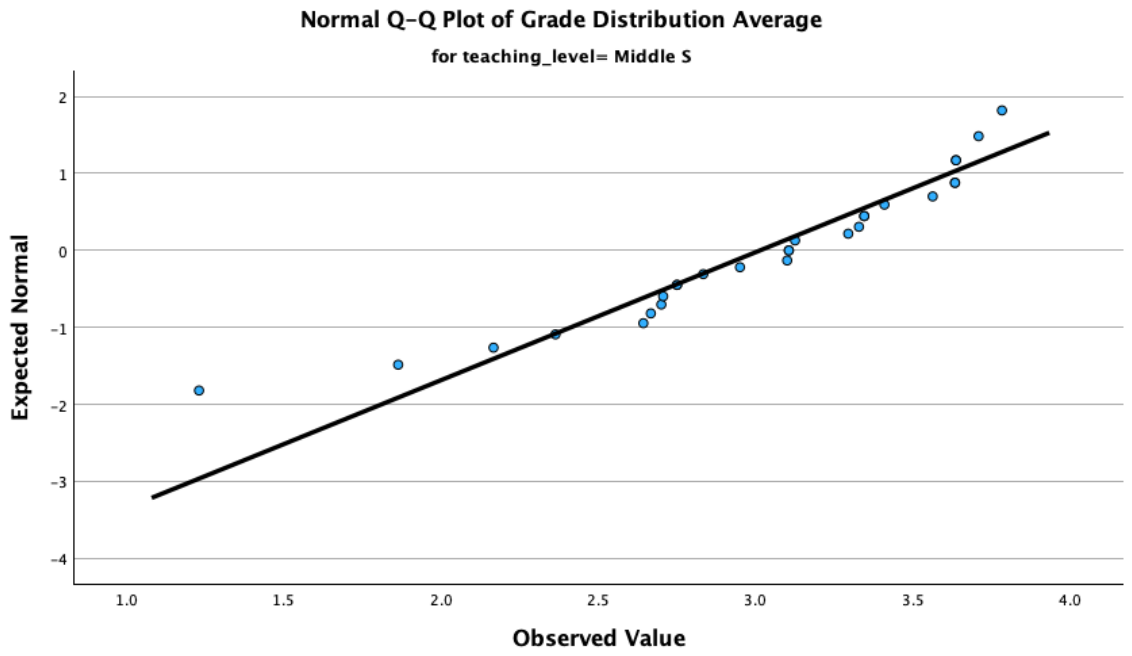


Figure 13

Plot of Grade Distribution Average for Middle School

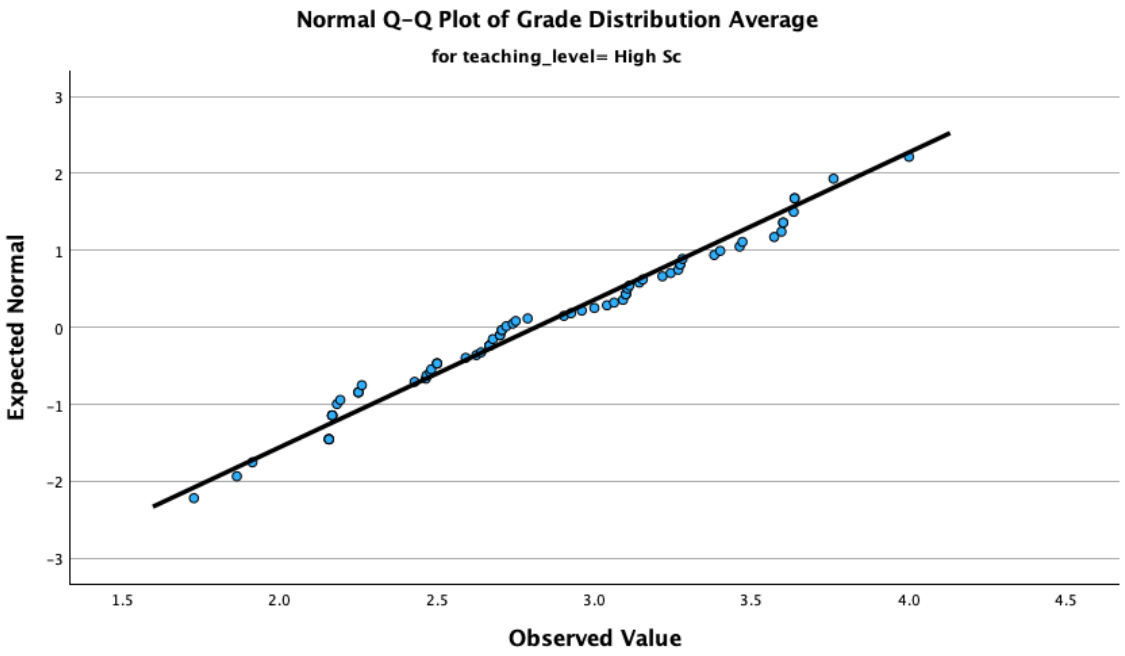


Figure 14

Plot of Grade Distribution Average for High School

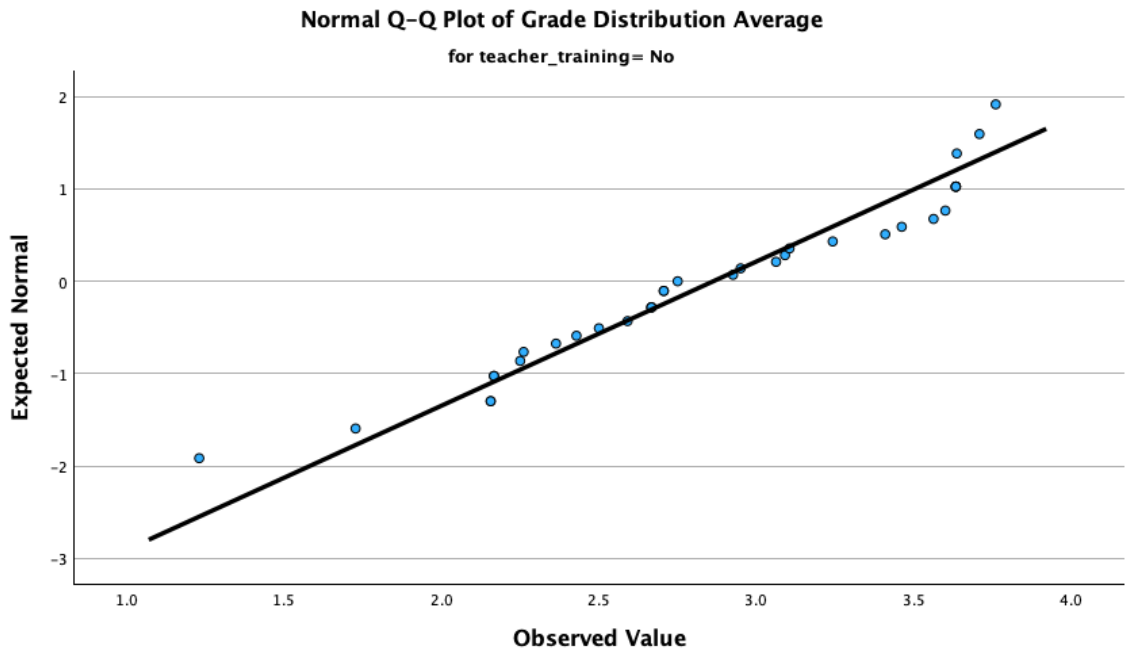


Figure 15

Plot of Grade Distribution Average for Teacher Training: No

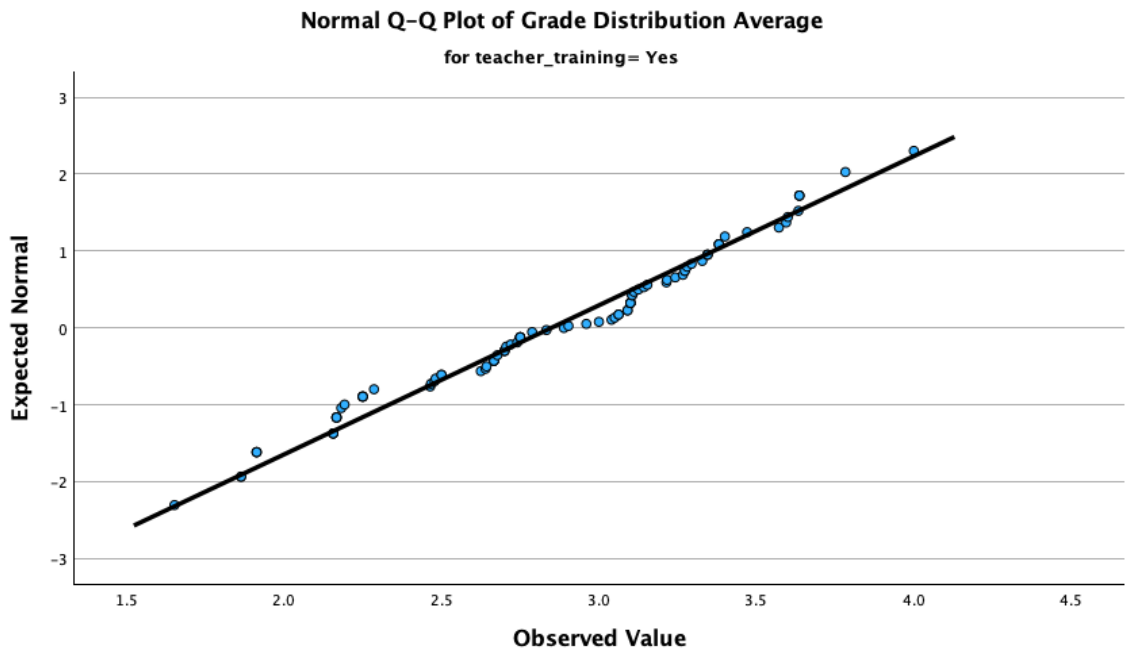


Figure 16

Plot of Grade Distribution Average for Teacher Training: Yes

Part A: Grade Level Taught and Grade Distribution Average

A one-way ANOVA was conducted to determine if the grade distribution average differed for teachers at varying grade levels. Participants were classified into three groups: elementary school (n = 26), middle school (n = 26), and high school (n = 74). Descriptive statistics are exhibited in Table 10. No outliers impacted the analysis, as assessed by boxplot; data was normally distributed for each group as determined by Normal Q-Q Plot; and there was homogeneity of variances, as assessed by Leven's test of homogeneity variances, p = .830 (see Table 11).

Table 10

One-Way ANOVA Descriptive Statistics: Grade Level Taught

Grade Level Taught	*N*	*M*	*SD*
Elementary School (PK-5)	26	2.79	.56
Middle School (6-8)	26	3.01	.62
High School (9-12)	74	2.81	.52

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation.

Table 11

Levene's Homogeneity of Variance Test: Grade Level Taught

Levene Statistic	df1	df2	Sig.
.187	2	123	.830

The grade distribution average increased from elementary school (n = 26, M = 2.79, SD = .56) to high school (n = 74, M = 2.81, SD = .52) to middle school (n = 26, M = 3.01, SD = .62), in that order, but the differences between grade level taught groupings were not statistically significant, $F(2, 123) = 1.452$, $p = .238$ (see Table 12). The group means were not statistically significantly different ($p > .05$), and, therefore, we can accept the null hypothesis and reject the alternative hypothesis, meaning there is no difference between a teacher's grade level taught and the grade distribution of that teacher (see Figure 17).

Table 12

One-Way ANOVA Source Table: Grade Level Taught

Grade Level Groupings	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	.880	2	.440	1.452	.238
Within Groups	37.283	123	.303		
Total	38.163	125			

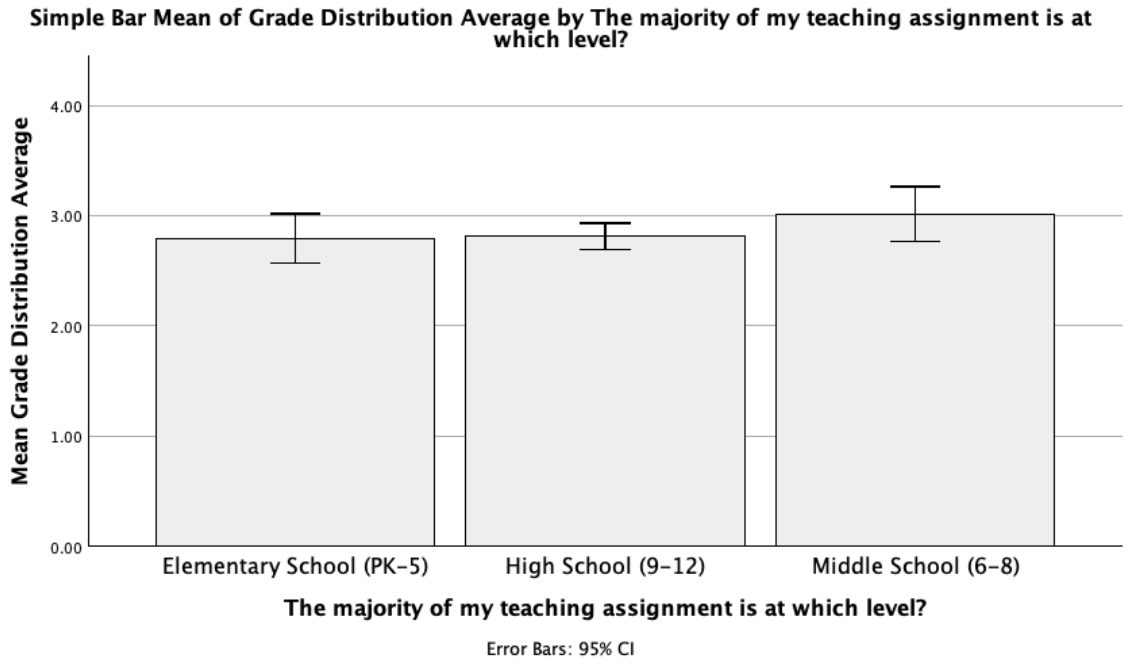


Figure 17

Bar Chart of Mean Grade Distribution Averages Based on Grade Level Taught

Part B: Teacher Preparation Program and Grade Distribution Average

A one-way ANOVA was conducted to determine if the grade distribution average differed for teachers at varying grade levels. Participants were classified into two groups: teachers who completed a traditional university-based teacher preparation program (n = 93) and those who did not (n = 35). Descriptive statistics are exhibited in Table 13. No outliers impacted the analysis, as assessed by boxplot; data was normally distributed for each group as determined by Normal Q-Q Plot; and there was homogeneity of variances, as assessed by Leven's test of homogeneity variances p = .078 (see Table 14).

Table 13

One-Way ANOVA Descriptive Statistics: Teacher Preparation Program Completion

Teacher Preparation Status	*N*	*M*	*SD*
Completed Teacher Prep Program	93	2.85	.52
Did Not Complete Teacher Prep Program	35	2.86	.64

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation.

Table 14

Levene's Homogeneity of Variance Test: Teacher Preparation Program Completion

Levene Statistic	df1	df2	Sig.
3.150	1	126	.078

The grade distribution average increased from teachers who completed a teacher preparation program (n = 93, M = 2.85, SD = .52) to teachers who did not complete a teacher preparation program (n = 35, M = 2.86, SD = .64), in that order. However, the differences between completion of a teacher preparation program or not were not statistically significant, $F(1, 126) = .019$, $p = .891$ (see Table 15). The group means were not statistically significantly different ($p > .05$), and, therefore, we can accept the null hypothesis and reject the alternative hypothesis, meaning there is no difference between a teacher's completion of a teacher preparation program or not, and the grade distribution of that teacher.

Table 15

One-Way ANOVA Source Table: Teacher Preparation Program Completion

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	.006	1	.006	.019	.891
Within Groups	38.424	126	.305		
Total	38.430	127			

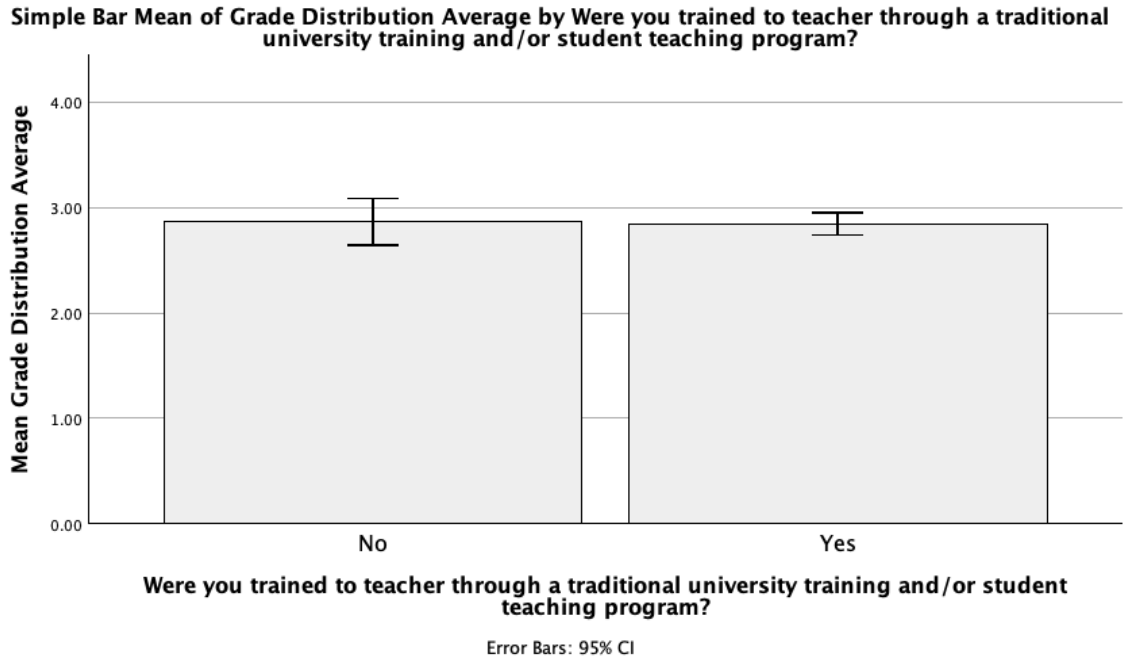


Figure 18.

Bar Chart of Mean Grade Distribution Averages Based on Completion of Teacher Preparation Program

Research Question 3 Analysis

RQ3: What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based), or hybrid (combination) grading practices and their grade distributions?

a. H_0 There is no relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.
b. H_A There is a relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.

Research question three was analyzed using the one-way ANOVA, which is used to determine whether there are any statistically significant differences between the means of two or more independent groups (Laerd Statistics, 2017b). In this analysis of this question, we are looking for differences between the grading methodology used in practice and their grade distribution.

As previously discussed, the one-way ANOVA has six assumptions that must be met, three of which are based on the study design. The first assumption requires one continuous dependent variable; the grade distribution average meets this assumption. Additional study design assumptions include the inclusion of one independent variable that consists of two or more categorical, independent groups. The grading methodology used in practice meets this assumption. The third assumption requires the independence of observations, and this study meets this requirement as no one participant belongs to multiple tested groups. The additional assumptions required for the one-way ANOVA are as follows:

The one-way ANOVA requires that there be no significant outliers in the groups of the independent variable (Laerd Statistics, 2017b). The data were analyzed for outliers using a boxplot in Figure 19.

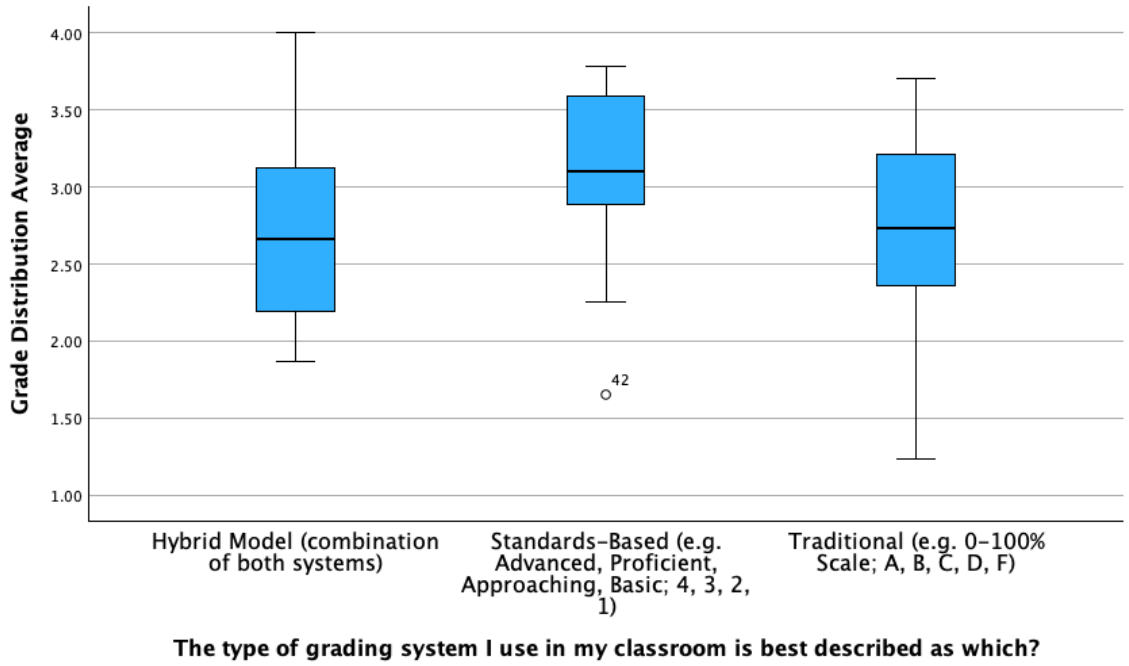


Figure 19

Boxplot Analysis of Grading Methodology Used in Practice and Grade Distribution Average

Outliers in the data were addressed by running the one-way ANOVA with the outliers included and again without the outliers included to look for variances in the results. The outliers were included in the analysis reported here if no variance was detected. Altering the data by removing the outliers may introduce bias into the analysis (Ghosh & Vogt, 2012).

The data for the one-way ANOVA also required testing of normality. Scores for grade distribution averages were normally distributed for all groups, as assessed by visual inspection of Normal Q-Q Plots, referenced in Figures 20 – 22.

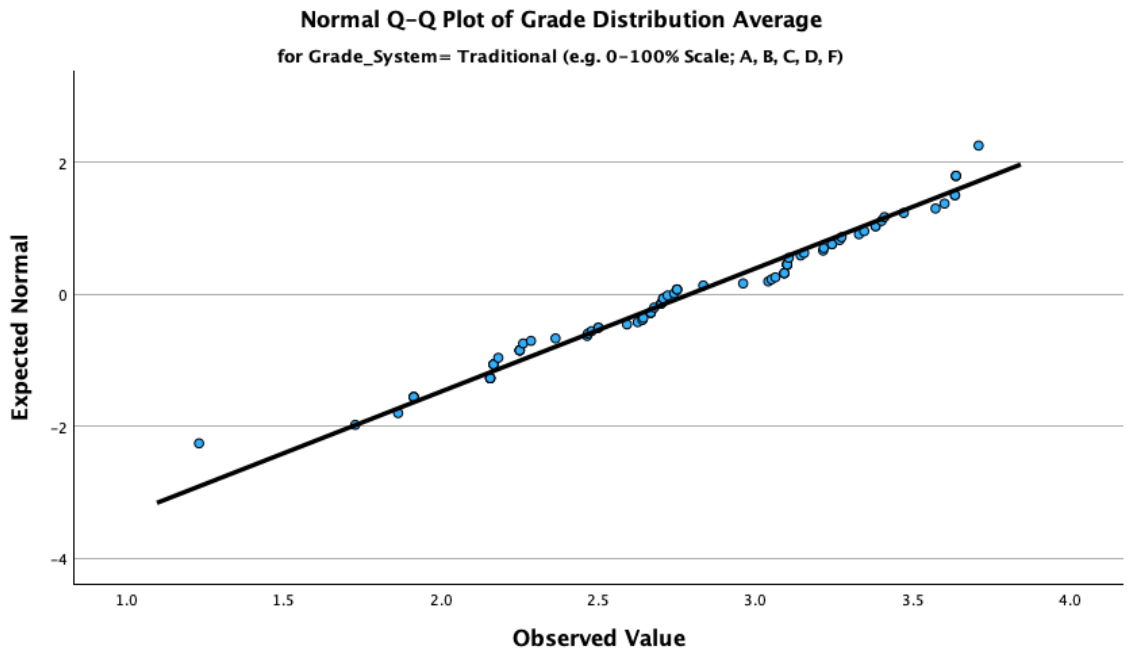


Figure 20

Plot of Grade Distribution Average for Traditional Grading System

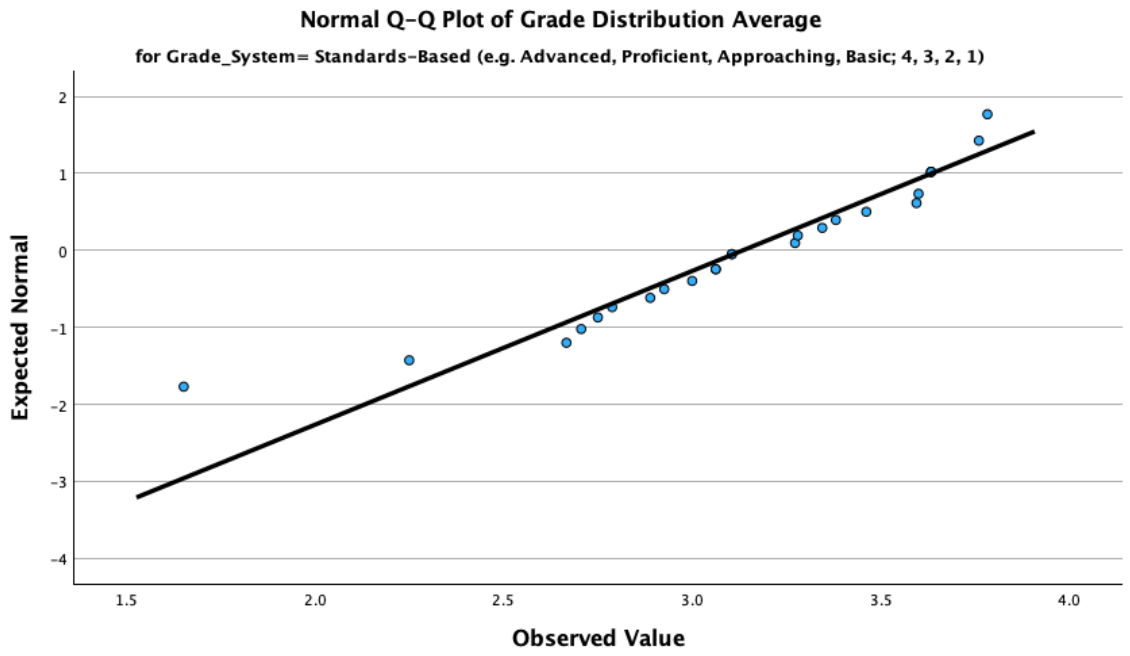


Figure 21

Plot of Grade Distribution Average for Standards-Based Grading System

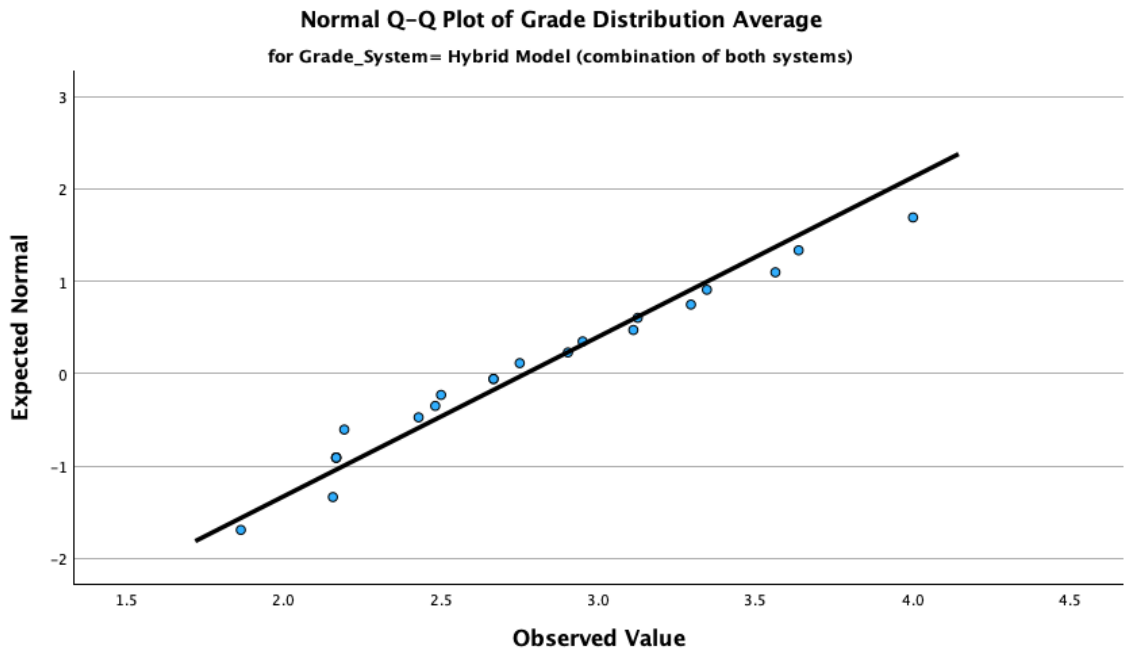


Figure 22

Plot of Grade Distribution Average for Hybrid Model Grading System

A one-way ANOVA was conducted to determine if the grade distribution average differed for teachers using different grading methodologies. Participants were classified into three groups: traditional grading system (e.g., 0-100% scale; A, B, C, D, F) (n = 82), standards-based grading system (e.g., advanced, proficient, approaching, basic; 4, 3, 2, 1) (*n* = 25), and hybrid model grading system (combination of both) (*n* = 21). Descriptive statistics are exhibited in Table 16. No outliers impacted the analysis, as assessed by boxplot; data was normally distributed for each group as determined by Normal Q-Q Plot; and there was homogeneity of variances, as assessed by Leven's test of homogeneity variances, $p = .536$ (see Table 17).

Table 16

One-Way ANOVA Descriptive Statistics: Grading Methodology Used in Practice

	N	*M*	*SD*
Traditional System	82	2.79	.52
Standards-Based System	25	3.13	.50
Hybrid System	21	2.77	.58

Note. N refers to sample size; *M* refers to mean; *SD* refers to standard deviation.

Table 17

Levene's Homogeneity of Variance Test: Grading Methodology Used in Practice

Levene Statistic	df1	df2	Sig.
.627	2	125	.536

The grade distribution average increased from the hybrid system (n = 21, M = 2.77, SD = .58) to the traditional system (n = 82, M = 2.79, SD = .52) to the standards-based system (n = 25, M = 3.13, SD = .50), in that order.

The differences between grading methodologies used in practice were statistically significant, $F(2, 125) = 1.234$, $p = .016$ (see Table 18). To examine the data further, a Tukey post hoc test is appropriate to test all possible group comparisons and identify which specific groups differ (Laerd Statistics, 2017b). Table 19 exhibits the results of the Tukey post hoc test.

Table 18

One-Way ANOVA Source Table: Grading Methodology Used in Practice

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	2.468	2	1.234	4.290	.016
Within Groups	35.962	125	.288		
Total	38.430	127			

Table 19

Tukey HSD Multiple Comparisons Table

(I) Group	(J) Group	Mean Diff (I-J)	Std. Error	Sig.
Hybrid Model	Standards-Based Model	-.36538	.15877	.059
	Traditional Model	-.01955	.13118	.988
Standards-Based Model	Hybrid Model	.36538	.15877	.059
	Traditional Model	.34583*	.12254	.015
Traditional Model	Hybrid Model	.01955	.13118	.988
	Standards-Based Model	-.34583*	.12254	.015

Note. * The mean difference is significant at the 0.05 level.

There was an increase in grade distribution average from the traditional model group (M = 2.79, SD = .52) to the standards-based group (M = 3.13, SD = .50), a mean increase of .34, 95% CI [.05, .63], which was statistically significant (p = .015). The group means were statistically significantly different ($p < .05$); therefore, we can reject the null hypothesis and accept the alternative hypothesis, meaning there is a difference between a teacher's grading methodology used in practice and the grade distribution of that teacher (see Figure 23).

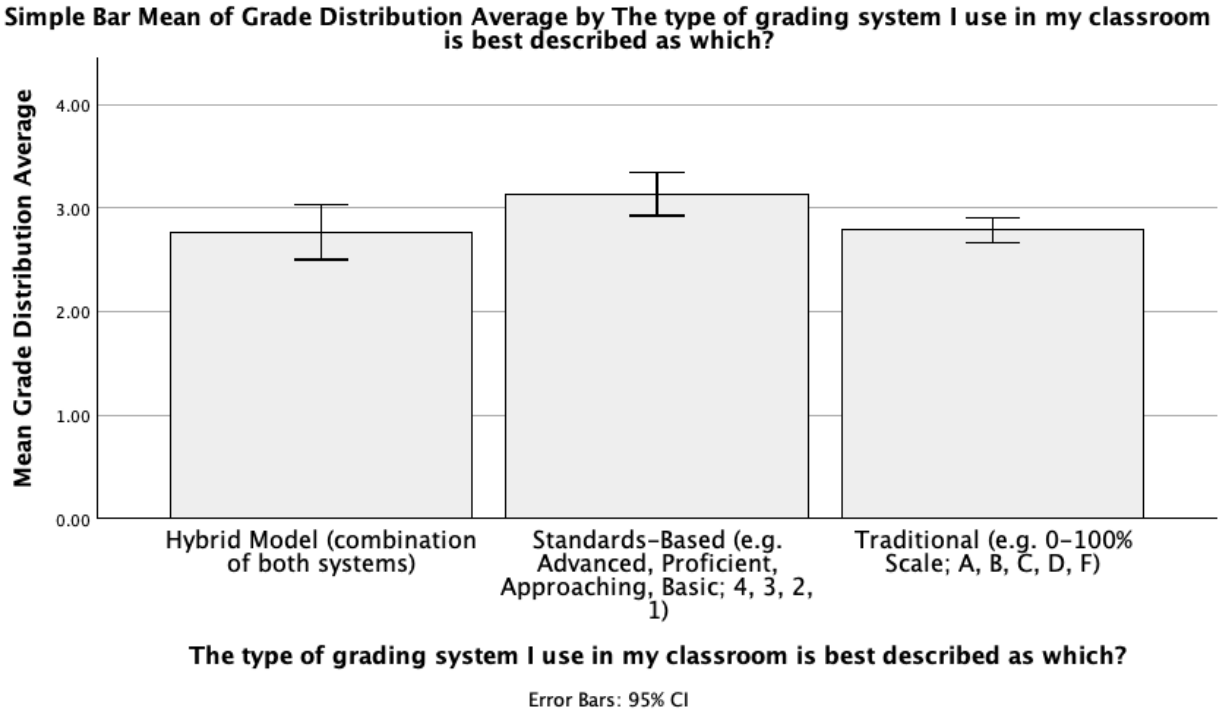


Figure 23

Bar Chart of Mean Grade Distribution Averages Based on Grading Methodology Used in Practice

Summary

Chapter Four provided a detailed analysis of demographic data, research questions, and methodology. The study analyzed the quantitative results of the Teacher Perceptions of Grading Practices survey instrument, along with relevant teacher characteristics about this study and self-reported grade distributions from respondents. Results indicated no significant relationships or differences between teacher perceptions of grading practices and the grade distributions of that teacher. Furthermore, the study results indicated no statistically significant relationships or differences in teacher characteristics of grade level taught or completion of teacher preparation program and their grade distributions.

Research question three detected a statistically significant difference. It examined the relationships between teacher groups using different grading methodologies in practice. Teachers using a standards-based model had a statistically significant grade distribution increase compared to traditional grading models.

Chapter Five will include a summary of the study and discussions of the findings for each research question. It will also include implications for practice and policy and recommendations for further research on this topic. The chapter will conclude with the impact that this study has had on the field of educational leadership.

CHAPTER 5
SUMMARY, DISCUSSIONS AND CONCLUSIONS

Introduction

In the previous chapter, quantitative data collected from respondents of the Teacher Perceptions of Grading Practices Survey (TPGPS) were disaggregated and analyzed. The TPGPS was comprised of three sections: (1) teacher demographic and characteristic data, (2) the TPGPS instrument, which includes five domains: Importance, Usefulness, Student Ability, Student Effort, and Teacher Grading Habits, and (3) grade distribution data reported as letter grades for one class period, for one academic term. Chapter 5 begins with a summary of the study, which includes the purpose of the study and methodology. A discussion of the findings for each research question is provided to contribute to the literature and body of knowledge on teacher characteristics and perceptions of grading practices and their potential effects on grade distributions of students. Finally, the researcher offers implications for practice for school-level administrators, school districts, and professional learning programs and closes with recommendations for future research.

Summary of the Study

Over a century of research has expounded on the inconsistencies in grading practices and outcomes. Starch (1913) highlighted the variability of scores on the same piece of student work. Historically, it has been accepted that teachers use various measures, from academic performance data to other factors not necessarily aligned with academic outcomes (Brookhart et al., 2020). Brookhart (1993) called out the central idea of why grades are so important, “Grading

is important to study because of the centrality of grades in the educational experience of all students" (p. 139).

This quantitative study investigates whether teacher perceptions of grading impact their students' grade distributions, whether teacher characteristics impact their students' grade distributions, and whether a teacher's grading system produces a different distribution outcome. The conceptual framework for this study was the Complex Systems Theory, which explains the challenge with grading reform efforts. Looking at one part of the system will not produce a result; instead, one must look at how the variables interact with one another.

The study utilized a voluntary response sampling and was open to all active K-12 teachers in the United States. Recruitment to research has traditionally been challenging and often results in a small sample for a particular study. This will slow the progress of valuable research and impact the outcomes (Wertheimer, 2013). To increase the sample size for this current study, social media platforms were leveraged to target survey respondents and cast a broader representation and sampling of data.

For research question one, a Pearson's correlation was run to assess the relationship between the TPGPS domain averages of respondents and their grade distribution averages. One hundred and twenty-eight responses were analyzed. No statistically significant correlation existed between the teachers' perception of importance, usefulness, student effort, or grading habits and their grade distribution average. While a slight positive correlation existed between student ability domain averages and grade distributions (r = .018), the correlation had no statistical significance (see Figure 24).

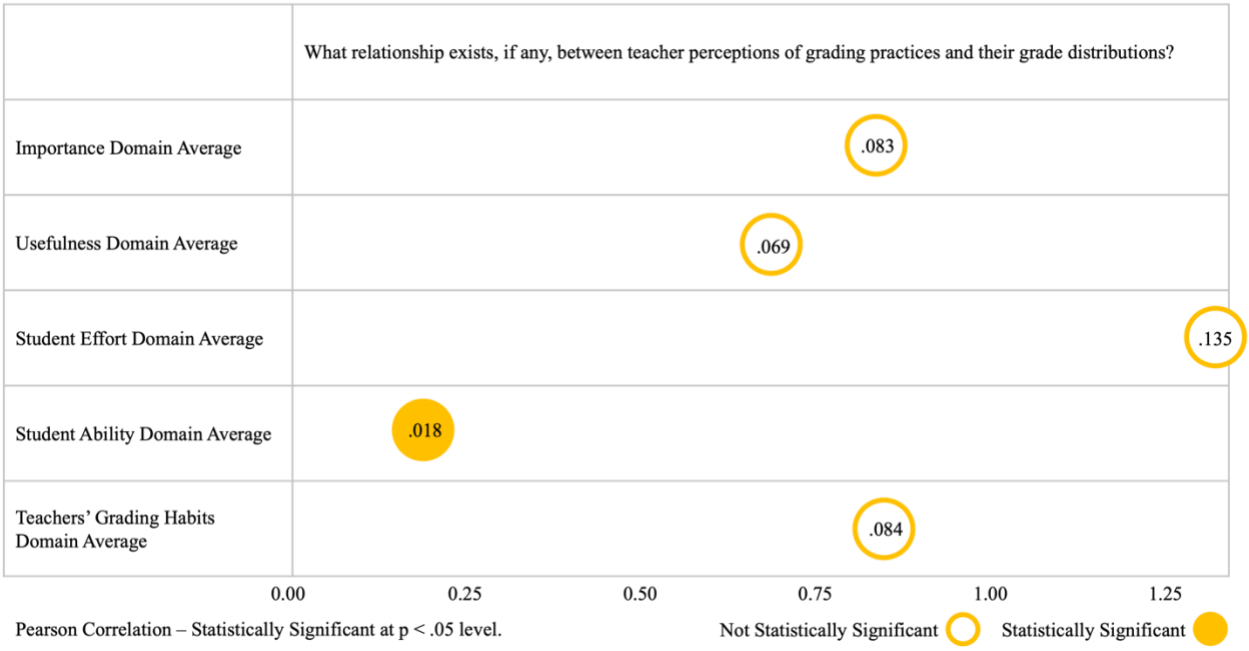


Figure 24

TPGPS Domains Pearson Correlation Statistically Significant Findings Data Visualization

For research question two, part A, the grade distribution average increased from elementary school (n = 26, M = 2.79, SD = .56) to high school (n = 74, M = 2.81, SD = .52) to middle school (n = 26, M = 3.01, SD = .62), in that order, but the differences between grade level taught groupings were not statistically significant, $F(2, 123) = 1.452$, $p = .238$, as outlined in detail in Chapter Four. The group means were not statistically significantly different ($p > .05$), and, therefore, we can reject the null hypothesis and we cannot accept the alternative hypothesis, meaning there is no difference between a teacher's grade level taught and the grade distribution of that teacher (see Figure 25).

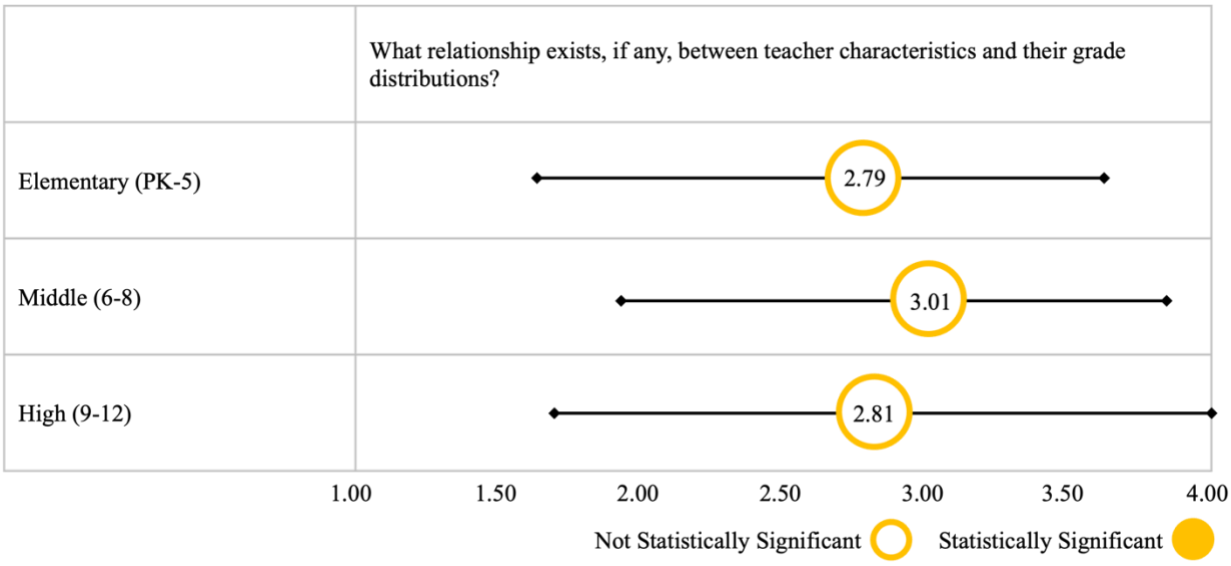


Figure 25

Grade Level Taught One-Way ANOVA Statistically Significant Findings Data Visualization

For research question two, part B, the grade distribution average increased from teachers who completed a teacher preparation program (n = 93, M = 2.85, SD = .52) to teachers who did not complete a teacher preparation program (n = 35, M = 2.86, SD = .64), in that order. However, the differences between completion of a teacher preparation program or not were not statistically significant, $F(1, 126) = .019$, $p = .891$, as outlined in Chapter Four in detail. The group means were not statistically significantly different ($p > .05$), and, therefore, we cannot reject the null hypothesis and we cannot accept the alternative hypothesis, meaning there is no difference between a teacher's completion of a teacher preparation program or not, and the grade distribution of that teacher (see Figure 26).

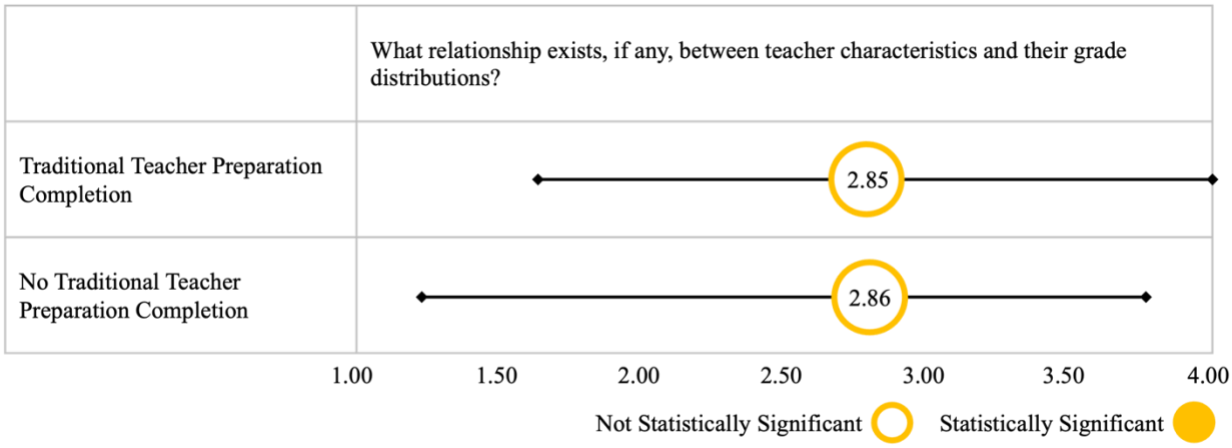


Figure 26

Teacher Preparation Program Completion One-Way ANOVA Statistically Significant Findings Data Visualization

There was an increase in grade distribution average from the traditional model group (M = 2.79, SD = .52) and the hybrid model group (M = 2.77, SD = .58) to the standards-based group (M = 3.13, SD = .50), a mean increase of .34, 95% CI [.05, .63], which was statistically significant (p = .015). The group means were statistically significantly different ($p < .05$); therefore, we can reject the null hypothesis and accept the alternative hypothesis, meaning there is a difference between a teacher's grading methodology used in practice and the grade distribution of that teacher (see Figure 27).

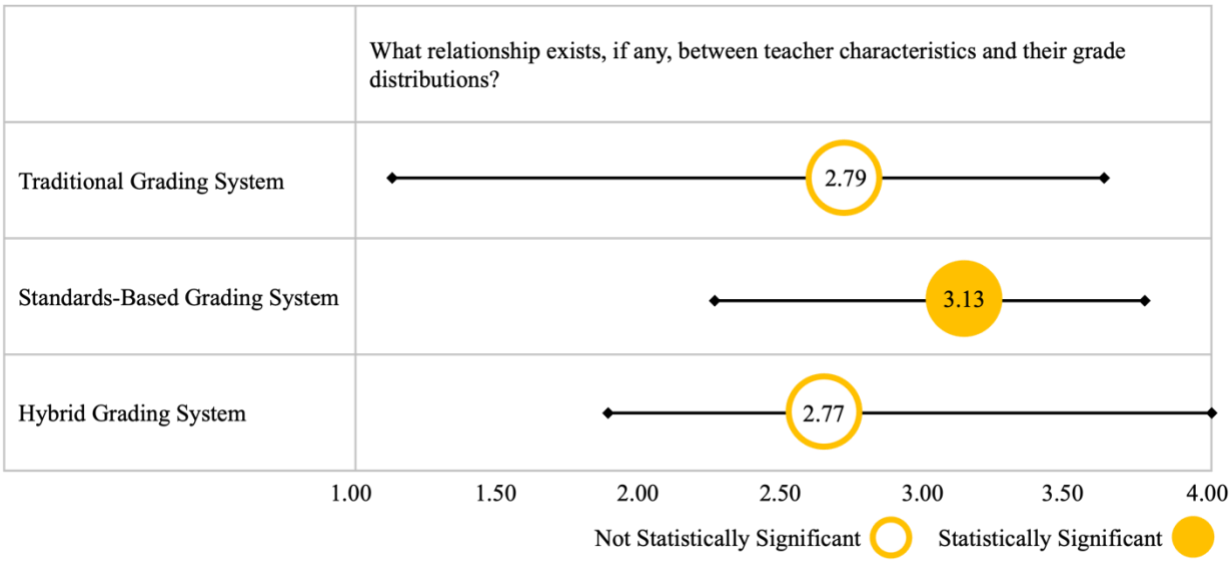


Figure 27

Grading System Used in Practice One-Way ANOVA Statistically Significant Findings Data Visualization

Discussion of Findings

Three research were created to guide this study and the relevant findings:

1. What relationship exists, if any, between teacher perceptions of grading practices and their grade distributions?
    a. H_0 There is no relationship between teacher perceptions of grading practices and their grade distributions.
    b. H_A There is a relationship between teacher perceptions of grading practices and their grade distributions.
2. What relationship exists, if any, between teacher characteristics and their grade distributions?

a. H_0 There is no relationship between teacher characteristics and their grade distributions.
b. H_A There is a relationship between teacher characteristics and their grade distributions.

3. What relationship exists, if any, between teachers using traditional grading practices (multiple choice, grading for completion), standards-based grading practices (performance-based, mastery-based), or hybrid (combination) grading practices and their grade distributions?
    a. H_0 There is no relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.
    b. H_A There is a relationship between teachers using traditional grading practices vs. standards-based grading practices vs. hybrid grading practices and their grade distributions.

To examine the relationship between the TPGPS domain score and their grade distribution, the researcher calculated a Pearson *r* correlation. The continuous variable for the Pearson r correlation was the grade distribution Average. To attain the grade distribution average, the researcher collected self-reported final letter grades for students in one course for one academic quarter as an addendum to the TPGPS instrument. These letter grades were converted to a traditional grade-point average score (e.g., A = 4, B = 3, C = 2, D = 1, and F = 0). A grade distribution average was then calculated by totaling the letter grade sum and dividing it by the total number of grades reported. The TPGPS is included in Appendix B. All other

relationships (e.g., grade level taught and grade distribution average, teacher preparation program enrollment and grade distribution average, and grading practices and grade distribution average) were examined using a one-way ANOVA.

Research Question 1

A Pearson r-correlation was used to examine research question one and to look for a relationship between the five domains for the Teacher Perceptions of Grading Practices Survey: (a) importance, (b) usefulness, (c) student effort, (d) student ability, and (e) teachers' grading habits. After ensuring that all assumptions were met, there were no statistically significant relationships between any of the five TPGPS domains and the grade distribution score of the respondents.

Therefore, this researcher can confidently state that there is no statistically significant relationship between teacher perceptions of grading practices as measured by the TPGPS and their grade distributions.

There may be several reasons why no statistical significance was found for research question one. Despite using a Pearson r-correlation to analyze the TPGPS domains and grade distribution scores, the inherent complexity of grading practices could be a contributing factor. Research suggests that teachers incorporate various aspects beyond just achievement into grading, such as student effort and ability. Stiggins et al. (1989) found that nearly half of teachers considered student effort, and a significant majority considered student ability when assigning grades. This complexity may have led to variations in how teachers distribute their grades, even if their general perceptions of grading practices align.

Limitations associated with sample size and external validity could also explain the need for more significance. For instance, Cross and Fray's (1999) large-scale study found that a quarter of teachers raised scores based on student effort. This study may not have captured the nuances within the teacher population, potentially leading to non-significant results. These factors and the inherent subjectivity of grading practices could explain why this study did not yield statistically significant relationships between teacher perceptions and their grade distributions.

Research Question 2

Research question two was analyzed using the one-way ANOVA, looking for differences between the teacher's grade level and their grade distribution, as well as differences among those teachers who completed traditional teacher education preparation programs and those who did not and their grade distributions.

The question was split into two parts: (a) grade level taught and grade distribution average and (b) teacher preparation program and grade distribution average. Each was explored separately. For the first part of the question, participants were classified into three groups: elementary school (n = 26), middle school (n = 26), and high school (n = 74). For the second part, participants were classified into two groups: teachers who completed a traditional university-based teacher preparation program (n = 93) and those who did not (n = 35).

For both parts of the research question, the group means were not statistically significantly different ($p > .05$), and therefore, we cannot reject the null hypothesis. We cannot accept the alternative hypothesis, meaning there is no difference between a teacher's completion of a teacher preparation program and the grade distribution of that teacher.

Multiple factors could be attributed to the lack of statistical significance in the findings relating to the current body of research. First, this study's traditional teacher preparation group may encompass a high degree of heterogeneity. Research by Hemmeter et al. (2008) suggests that traditional teacher preparation programs vary in their emphasis on effective grading practices. This variation within the *traditional* group could blur potential differences between teachers who completed these programs and those who still need to. A more nuanced categorization of teacher preparation programs, considering the specific content and focus of grading practice, might be necessary to detect a more transparent relationship.

The scope of grading practices examined in this study may also be a contributing factor. While the study focused on grade distributions, Brookhart (1991) highlights the multifaceted nature of grading practices, encompassing aspects like student effort and ability. A teacher's grading philosophy may have masked any relationship between these characteristics and grade distributions. Future research that explores a broader range of grading practices may yield more conclusive results.

Research Question 3

Research question three was analyzed using the one-way ANOVA, looking for differences between the grading methodology used in practice and their grade distribution. Participants were classified into three groups: traditional grading system (e.g., 0-100% scale; A, B, C, D, F) (n = 82), standards-based grading system (e.g., advanced, proficient, approaching, basic; 4, 3, 2, 1) (n = 25), and hybrid model grading system (combination of both) (n = 21).

The group means were statistically significantly different ($p < .05$); therefore, we can reject the null hypothesis and accept the alternative hypothesis, meaning there is a difference

between a teacher's grading methodology used in practice and the grade distribution of that teacher.

The statistically significant findings, indicating a strong connection between teacher grading methodology and grade distribution, align with the existing body of research on the impact of grading practices on student outcomes. Kelley's (2004) work on sample size and generalizability suggests that with a sufficient sample size (n=82) in the *traditional grading* group), statistically significant results can be interpreted with more confidence. This strengthens the argument that the observed relationship between methodology and distribution is not simply due to chance but reflects a genuine pattern in teacher grading practices.

The significance reinforces that different grading methodologies influence how teachers distribute grades. This finding complements Guskey's (2018) research on the complexities of grading practices. While Guskey emphasizes the multifaceted nature of grading decisions, the statistically significant results suggest that despite these complexities, the chosen methodology plays a clear role in shaping the overall distribution of grades.

Furthermore, the statistically significant results provide valuable evidence that teacher grading methodology has a measurable impact on grade distribution. This finding can inform future research that delves deeper into the specific mechanisms by which different methodologies influence this distribution while also considering the additional factors highlighted by Guskey (2018) that teachers incorporate into their grading decisions.

Implications for Practice

Research reveals significant inconsistencies in grading practices throughout history, making it difficult to compare grades across time directly. Standards for what constitutes an "A" or a "B" have changed. As Pauline Cahen (2019) points out, "[historical studies of student performance on standardized tests show] a clear upward trend in scores over time, even when accounting for changes in the tests themselves" (p. 12). This suggests that grades may reflect not just a student's absolute knowledge or skills but their performance relative to the expectations of their specific era.

Decades of inconsistencies have created school grading systems that will be difficult to change. However, because there may be adversity, our students still deserve a more consistent system of grading that measures their learning more accurately.

The findings of this study provide the following considerations, recommendations, and implications for practice and policy:

1. *Rethinking grades for more transparent communication and student growth.* The findings of this study underscore the need for a multi-pronged approach to grading reform. Here are some actions steps informed by recent research:

   Focus on Standards-Based Grading: Implement standards-based grading systems that clearly define what students need to know and do to achieve specific grades. A study by Popham (2021) found that standards-based grading improved students' understanding of learning goals and fostered a growth mindset. This reduces subjectivity and ensures grades accurately reflect mastery of curriculum standards.

Embrace Feedback Loops: Move beyond static grades to incorporate feedback loops that promote ongoing learning. A study by Bangert et al. (2022) suggests that effective feedback includes specific, actionable steps for improvement. Encourage teachers to provide timely, descriptive feedback alongside grades to maximize student learning.

Diversify Assessment Strategies: Complement traditional exams with a broader range of assessments that cater to diverse learning styles. McMillan (2023) explored the benefits of formative assessments like self-evaluations and peer reviews. These provide a more comprehensive picture of student learning and promote student ownership of the learning process.

Implementing these steps can create a grading system that fosters clear communication and collaboration. Grades can then become a meaningful reflection of student progress, not just a subjective product of teacher perceptions.

2. *Tailored professional development for effective teaching.* The findings of this study highlight a need to personalize professional development opportunities. Here are some action steps informed by the research:

   Needs-Based Professional Development: This approach shifts from a one-size-fits-all approach to individual teacher and school needs. Yin et al. (2022) recommend ongoing formative assessments to identify areas where teachers require support. This allows professional development to target specific instructional strategies or curriculum knowledge gaps, potentially leading to improved student outcomes.

Content-Specific Training: Provide professional development opportunities that delve deeper into specific learning goals and curriculum content. Research by Guskey (2018) emphasizes the importance of pre-assessments to determine student understanding before instruction begins. Professional development focusing on formative assessment and differentiated instruction can equip teachers to address individual student needs more effectively.

Mentorship and Collaboration: Create teacher collaboration and mentorship opportunities within schools or districts. Goddard et al. (2021) found that collaborative professional development increased teacher knowledge and improved student learning gains. This could involve peer coaching, lesson study groups, or collaborative planning sessions.

By tailoring professional development to specific needs and focusing on effective teaching practices, we can create a system that empowers educators and enhances student learning outcomes regardless of their entry path into the profession.

3. *Rethink professional development for a more profound learning impact.* The finding that traditional teacher preparation programs do not necessarily lead to different achievement outcomes highlights a need to focus on the content and delivery of professional development. Here are some action steps:

   Target professional development learning goals: Avoid generic professional development and tailor it to specific learning goals and curriculum areas. Desimone et al. (2021) found that professional development programs focused on improving teacher content knowledge can produce substantial student learning gains. School

districts can collaborate with teachers to identify these areas and design targeted professional development programs accordingly.

Incorporate formative assessment strategies: Equip teachers with effective formative assessment strategies to identify student needs before, during, and after instruction. Research by Guskey (2018) emphasizes the importance of pre-assessments to determine student understanding before instruction begins. Including formative assessment training in professional development programs can better prepare teachers to address student learning gaps and personalize instruction.

Prioritize collaboration and coaching. Create opportunities for teacher collaboration and ongoing coaching to strengthen instructional practice. Goddard et al. (2021) found that collaborative professional development increased teacher knowledge and improved student learning gains. This could involve peer coaching models, lesson study groups, or collaborative planning sessions facilitated by experienced teachers. By implementing these steps, we can ensure that professional development is directly connected to improving student learning. Focusing on content-specific knowledge, formative assessment, and collaborative learning can empower teachers to significantly impact the classroom regardless of the preparation program.

4. *Empowering all students through effective assessment.* This study highlights the importance of moving away from grading practices that focus on a predetermined distribution of grades and instead focus on all students achieving high levels of learning. Here are some action steps that could be taken from the research:

Develop clear learning objectives. Research shows that clearly defined learning objectives are essential for student success (Guskey & Bailey, 2021). Students can focus their efforts and track their progress when they understand what they are expected to learn. Schools and districts can implement policies that require teachers to develop clear learning objectives for each lesson or unit.

Use formative assessments to inform instruction. Formative assessments check for understanding throughout the learning process rather than just at the end. This allows teachers to identify areas where students struggle and adjust their instruction accordingly (Stanford Center for Teaching and Learning, 2023). Schools and districts can provide professional development for teachers on developing and using formative assessments effectively.

Provide students with opportunities for feedback and revision. Feedback is essential for helping students learn from their mistakes and improve their work. Schools and districts can create policies that encourage teachers to provide students with regular feedback, both written and verbal. They can also create opportunities for students to revise their work based on feedback.

By implementing these steps, schools and districts can create a learning environment where all students have the opportunity to succeed.

In conclusion, the decades of inconsistencies in grading practices across time and between schools highlight the need to examine current grading systems critically. While the findings of this study suggest that teacher perceptions may not hold significant sway over grading outcomes, the current "hodgepodge" effect (Guskey, 2006) reduces complex student

learning into a single letter grade. This lack of consistency makes it difficult to measure student performance fairly and creates confusion among parents, students, and educators. Further research is needed to identify and implement more effective grading practices that accurately reflect student learning and achievement.

Implications for Policy

This study can also serve as a catalyst for policy change that can help support a more effective grading policy and practice. The following reforms can be implemented through local (school district), state, and national policy changes. Here are some recommendations for policymakers supported by the research and potential benefits:

Local (School District) Policies

Curriculum and Standards: Districts can adopt clear, well-defined learning standards that outline student expectations for each grade level and subject. Popham (2021) emphasizes the importance of clear learning goals for students to understand expectations and guide their efforts.

Grading Guidelines: Create consistent grading frameworks across the district. This ensures that teachers use similar criteria for assessments, reducing subjectivity. Marzano & Heflebower (2011) highlighted how inconsistent grading practices can lead to inaccurate representations of student learning.

Professional Development: Provide teachers with training on standards-based grading, feedback strategies, and diverse assessment methods. Bangert et al. (2022) emphasized equipping teachers with practical feedback skills to enhance student learning.

Pilot Programs: Encourage schools to pilot alternative grading models, like mastery-based learning, in which students progress by demonstrating understanding.

State Policies

Rethink Standardized Testing: States can evaluate the weight of standardized testing in grading. They could focus on formative assessments that provide continuous feedback for improvement. McMillan (2023) advocates for formative assessments as a crucial tool for ongoing learning and improvement.

Funding Incentives: Offer financial support to schools implementing successful grading reform programs. OECD (2016) suggests that financial incentives can have an impact when encouraging schools to adopt innovative practices.

Data Sharing Protocols: Encourage districts to share anonymized data on grading practices to identify and address inconsistencies. Data sharing can help identify best practices and inform policy decisions for continuous improvement within a state.

National Policies

Federal Funding with Conditions: The federal government could allocate funds to states that commit to developing and implementing school reform policies. Research suggests using federal funding with specific requirements can promote educational equity and desired outcomes (Perera et al., 2020).

Standardized Grading Frameworks: Develop national guidelines for precise and consistent grading practices while allowing local flexibility. Striking a balance between national standards and local needs can ensure a consistent approach while allowing for tailoring to specific contexts.

Research Collaboration: Fund national research initiatives to study the long-term impact of grading reforms on student learning and success. Continued research is crucial to evaluate the effectiveness of reforms and inform future policy decisions.

In conclusion, it should be noted that implementing these reforms may initially increase teacher workload. Policies should be paired with resources like professional development and reduced administrative tasks. Universities and colleges rely heavily on grade point averages. States and the federal government can work together to encourage holistic admissions practices considering multiple factors beyond grades. Overall, a multi-level approach with collaboration between local school districts, state departments of education, and the federal government is critical for enacting and sustaining meaningful grading reforms. By implementing these reforms and addressing potential concerns, we can create a grading system that fosters a love of learning, empowers students, and provides a more accurate picture of their academic progress.

Recommendations for Further Research

The body of research on grading is well-established and spans well over a century. However, the grading practices in today's schools have not changed much during that same period. Grading is one of the most sacred and personal practices that a teacher possesses. This is what makes reform so difficult.

As a follow-up to this study, several recommendations for further research will help inform practice in future years. Those recommendations are as follows:

1. *Population Sampling Size*

   Future research should increase the population sampling size of the study. There have

been many studies that have used the Teacher Perceptions of Grading Practices Survey; however, those studies were of a narrow scope and often included one or two districts. This researcher set out to capture a more extensive sampling representative of a more diverse makeup of states, districts, and schools. However, participation still leaves room for a much larger sampling size. Recommendations would include making the survey accessible for a much more extended period. A larger population sample size will allow for more varied experiences to be captured in any future studies. In educational research, utilizing a larger sample size strengthens the generalizability and credibility of your findings. Firstly, a larger sample provides a more accurate representation of the population you are studying. More data points reduce the risk of random sampling error, leading to a more reliable estimate of the population mean or proportion (Cohen et al., 2018). As Kline (2015) states, "Larger samples tend to produce more accurate estimates of population parameters" (p. 161). This precision gives greater confidence in applying your research conclusions to a broader educational context. Secondly, a larger sample size increases the statistical power of your study. This means you are better equipped to detect actual effects, reducing the chances of missing a significant relationship or mistaking random chance for a factual finding (Cohen et al., 2018). By employing a larger sample, you strengthen the ability of your research to identify impactful educational practices.

2. *Creation of Original Survey Instrument*

   Research question one included using domain scores from the Teacher Perceptions of Grading Practices Survey. Future research allows room for creating an original survey

instrument and tests other perceptions of grading practices not covered by the TPGPS. This will allow for further research to explore other factors that may impact student grade distributions that can inform the support of teachers' grading practices. Exploring other areas of teacher perceptions may allow for findings that help determine what, if anything, does impact student grade distribution outcomes. Developing an original survey instrument for your educational research can offer significant advantages. Firstly, it allows you to tailor the instrument to precisely address your specific research question and the unique context of your study. Existing instruments might need to capture the nuances of your area of interest perfectly. As DeVellis (2017) suggests, "A well-designed survey instrument can provide data that is directly relevant to the research questions" (p. 170). This targeted approach ensures that the data you collect directly informs your research goals. Secondly, developing your instrument allows for greater control over the validity and reliability of your measurement. By carefully crafting questions and conducting pilot testing, you can ensure the instrument accurately measures the intended constructs (DeVonellis, 2017). This level of control strengthens the foundation of your research and increases confidence in your findings.

3. *Qualitative Study*

The addition of qualitative methodology to this study will allow for additional context around the data points that may not be able to be captured by the narrow scope of the survey instrument used in this study. With grading and assessment practices, the study could benefit from teacher interview data regarding practice. While quantitative

research is vital in educational research, qualitative methods offer a unique and complementary perspective. Qualitative data allows researchers to delve deeper and understand the "why" behind the numbers. Through interviews, focus groups, and observations, researchers gain insights into the lived experiences and perceptions of students, teachers, and administrators (Creswell, 2014). This rich data provides context and meaning to quantitative findings, painting a more holistic picture of educational phenomena.

Furthermore, qualitative research excels at exploring new areas and generating new ideas. By allowing participants to share their perspectives freely, researchers can uncover unexpected themes and complexities that might be missed by predetermined survey questions (Merriam, 2009). This flexibility allows for adaptation and the discovery of nuanced aspects of educational practice, ultimately leading to a more comprehensive understanding of education.

4. *NAEP Regional Analysis*

The National Assessment of Educational Progress (NAEP) divides the United States into four geographic regions: Northeast, Midwest, South, and West. These regions provide a structured framework for analyzing educational trends and outcomes across different parts of the country. While NAEP data is commonly used to assess student achievement, its regional divisions can also be leveraged to examine grading practices more closely. Grading reform, focused on equity and consistency, often overlooks geographical dimensions that are critical for understanding regional disparities and tailoring interventions.

This recommendation proposes a comprehensive study to investigate grading practices and outcomes across NAEP regions. Key areas of focus include regional differences in grading policies, the impact of these practices on student performance, and equity issues within each region. Data collection would involve gathering information from a representative sample of schools, analyzing NAEP assessment data, state and district policies, and demographic factors. Quantitative analysis would identify significant trends, while qualitative case studies and interviews would provide contextual insights into local practices and influences.

Research on grading practices by NAEP regions would offer valuable contributions to policy and educational practice. It would provide evidence-based recommendations for regional policy adjustments to promote fair and equitable grading, offer practical insights for educators, and highlight successful regional strategies for broader adaptation. This approach not only enhances our understanding of educational equity but also establishes a foundation for future studies on other regional dimensions of educational practices, ultimately driving more informed and targeted grading reforms.

Despite a long history of research on grading practices, this study reveals a lack of correlation between a teacher's perceptions of grading and the grade distribution of their students. The author suggests this disconnect stems from the personal and deeply ingrained nature of grading for teachers, making reform efforts challenging. To move forward, the study recommends expanding research to include a broader range of teacher experiences, potentially leading to more robust and generalizable findings.

Conclusion

This study investigated teacher perceptions of grading practices and their effects on grade distributions. Through a survey design, the research examined how teachers' views on the importance, usefulness, student effort, and student ability influenced their grading practices. The findings indicated that teachers perceived these factors as necessary but did not significantly correlate with actual grade distributions. This suggests that other factors beyond teacher perceptions may hold more significant sway over grading patterns.

The results of this study have implications for educators, administrators, and policymakers. Educators should continue to reflect on their grading practices and strive for consistency. School administrators can play a role in establishing clear grading guidelines and providing professional development opportunities on effective grading practices. Policymakers may want to consider broader initiatives to create more standardized grading practices across educational institutions. Future research can explore the influence of factors beyond teacher perceptions, such as school culture, curriculum design, and standardized testing, on grading practices and student outcomes.

In conclusion, I share this quote from Nelson Mandela, “Education is the most powerful weapon which you can use to change the world.” This rings true, yet the endeavor of education reform can feel like yielding a blunt instrument against a fortress. The complexities of educational systems and the sheer scale of the challenge can leave even the most passionate reformer feeling discouraged.

However, despair is not an option. The potential rewards of a more equitable and effective education system for all students are too great. We must remember that even the most

formidable structures are built one brick at a time. Our ongoing efforts, combined with those of countless others who share this vision, can gradually reshape the educational landscape. Let us continue to chip away, to innovate, and to advocate for the education systems that our students deserve.

APPENDIX A
TEACHER PERCEPTIONS OF GRADING PRACTICES SURVEY

Please choose one of the following.

- ○ I have read the Informed Consent Statement and consent to participate.
- ○ I do not consent to participate. I understand my answers will NOT be included in this study.

Demographic Data

The majority of my teaching assignment is at which level?

- ○ Elementary School (PK-5)
- ○ Middle School (6-8)
- ○ High School (9-12)

I primarily teach which grade(s).

- ☐ PK
- ☐ K
- ☐ Grade 1
- ☐ Grade 2
- ☐ Grade 3
- ☐ Grade 4
- ☐ Grade 5
- ☐ Grade 6
- ☐ Grade 7
- ☐ Grade 8
- ☐ Grade 9
- ☐ Grade 10
- ☐ Grade 11
- ☐ Grade 12

I primarily teach which content area(s).

- ☐ English/Reading
- ☐ Mathematics
- ☐ Science

- [ ] Social Sciences
- [ ] Physical Education and/or Health
- [ ] Other Elective/Special

How many years have you been teaching?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10
- ○ 11
- ○ 12
- ○ 13
- ○ 14
- ○ 15
- ○ 16
- ○ 17
- ○ 18
- ○ 19
- ○ 20+

To which gender identity do you most identify?

- ○ Male
- ○ Female
- ○ Transgender Male
- ○ Transgender Female
- ○ Non-Binary/Non-Conforming
- ○ Prefer Not to Say

School & Training Data

Is your school a Title I school?

- ○ Yes
- ○ No
- ○ I am not sure.

Were you trained to teacher through a traditional university training and/or student teaching program?

- ○ Yes
- ○ No

Grading System Implementation

The type of grading system I use in my classroom is best described as which?

- ○ Traditional (e.g. 0-100% Scale; A, B, C, D, F)
- ○ Standards-Based (e.g. Advanced, Proficient, Approaching, Basic; 4, 3, 2, 1)
- ○ Hybrid Model (combination of both systems)

Importance Domain

Please score each of the following statements in the Importance Domain.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Grading is an important criterion for judging students' progress.	○	○	○	○	○
Grading has an important role in classroom assessment.	○	○	○	○	○

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Grading has a positive effect on students' academic achievement.	○	○	○	○	○
Grading practices are important measures of student learning.	○	○	○	○	○
Grading practices are important measures of student achievement.	○	○	○	○	○
Grading has a strong impact on students' learning.	○	○	○	○	○

Usefulness Domain

Please score each of the following statements in the Usefulness Domain.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Grading helps me categorize students' knowledge/performance as above average, average, and below average.	○	○	○	○	○
Grading can help me improve instruction.	○	○	○	○	○
Grading can encourage good work by students.	○	○	○	○	○
Grading helps me in deciding what curriculum to cover.	○	○	○	○	○
Grading is a good method for helping students identify their weaknesses in a content area.	○	○	○	○	○
Grading can keep students informed about their progress.	○	○	○	○	○
Grading provides information about student achievement.	○	○	○	○	○

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Grading documents my instructional effectiveness.	○	○	○	○	○
Grading provides feedback to my students.	○	○	○	○	○
High grades can motivate students to learn.	○	○	○	○	○
Grades of zero can demotivate students to learn.	○	○	○	○	○

Student Effort Domain

Please score each of the following statements in the Student Effort Domain.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I consider student effort when I grade.	○	○	○	○	○
I give higher report card grades for students who show greater effort.	○	○	○	○	○
I will pass a failing student if he or she puts forth effort.	○	○	○	○	○
Grades are based on students' completion of homework.	○	○	○	○	○
Grades are based on the degree to which students participate in class.	○	○	○	○	○
Grades are based on students' improvement.	○	○	○	○	○

Student Ability Domain

Please score each of the following statements in the Student Ability Domain.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I consider student ability in grading.	○	○	○	○	○
Grades are based on students' problem solving ability.	○	○	○	○	○
Grades are based on students' critical thinking ability.	○	○	○	○	○
Grades are based on students' independent thinking ability.	○	○	○	○	○
Grades are based on students collaborative learning ability.	○	○	○	○	○
Grades are based on students' writing ability (quality of writing, not handwriting skills).	○	○	○	○	○

Teachers' Grading Habits Domain

Please score each of the following statements in the Teachers' Grading Habits Domain.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
If a student fails a test, I will offer him or her a second chance to take the test.	○	○	○	○	○
If a student fails to complete an assignment, I will assign him or her a grade of zero.	○	○	○	○	○
If a student fails to complete an assignment, I will subtract grade points progressively until the assignment is turned in.	○	○	○	○	○
I often give students opportunities to earn extra credit.	○	○	○	○	○

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I often look at the distribution of grades for the whole class after I finish grading.	○	○	○	○	○
I have my own grading procedure.	○	○	○	○	○
I often confer with my colleagues on grading criteria.	○	○	○	○	○

Grade Distributions

Choose one class or period and enter the current grade distribution below. You will enter the raw number (for example, if you have 7 As in the class, enter 7 in the A field below).

Total # of As (Not Percentage)

Total # of Bs (Not Percentage)

Total # of Cs (Not Percentage)

Total # of Ds (Not Percentage)

Total # of Fs (Not Percentage)

Powered by Qualtrics

APPENDIX B
WRITTEN PERMISSION FOR SURVEY USE

Dr. Xing Liu
Professor and Associate Chair
Eastern Connecticut State University
Webb Hall

College of Community Innovation and Education

June 25, 2023

Dr. Liu,

I am a doctoral candidate at the University of Central Florida, completing a dissertation in the Executive Doctor of Education program at the College of Community Innovation and Education. I am requesting permission to use the Teachers' Perception of Grading Practices Survey in my research study. I am studying the correlation between teachers' perceptions of grading practices and their grade distributions. This tool will be essential in producing the right data for this study.

I plan to use the instrument in its entirety; eliminating questions that may not be relevant to the study is not likely. No questions will be adapted. This will be provided via closed social media groups of teachers - as I want to get the largest sample size possible and a large demographic representation.

In addition to using the instrument, I also ask your permission to reproduce it in my dissertation appendix. The dissertation will be published in the University of Central Florida online repository and available through additional academic research databases.

I would like to use [and reproduce] your Teachers' Perception of Grading Practices (TPGP) instrument under the following conditions:

- I will use the TPGP instrument only for my research study and not sell or use it for other purposes.
- I will include a statement of attribution and copyright on all instrument copies. If you have a specific attribution statement you would like me to include, please provide it in your response.
- At your request, I will send a copy of my completed research study to you upon completion of the study and/or provide a hyperlink to the final manuscript.

If you do not control the copyright for these materials, I would appreciate any information you can provide concerning the proper person or organization I should contact.

If these are acceptable terms and conditions, please indicate so by replying to me at michael.meechin@ucf.edu.

Educationally,

Michael Meechin
UCF Doctoral Candidate

From: **Liu,Xing (Education)** liux@easternct.edu
Subject: Re: TPGP Instrument Permission
Date: June 26, 2023 at 8:37 AM
To: Michael Meechin Michael.Meechin@ucf.edu

XL

Dear Michael,

You have my permission to use and revise the TPGP instrument.

Good luck with your research.

Xing

Xing Liu, Ph.D.
Professor, Research and Assessment
Associate Chair, Education Department
Eastern Connecticut State University
Webb Hall 153
83 Windham Street
Willimantic, CT 06226
(860) 465-5167
liux@easternct.edu

From: Michael Meechin <Michael.Meechin@ucf.edu>
Sent: Sunday, June 25, 2023 11:25 AM
To: Liu,Xing (Education) <liux@easternct.edu>
Cc: Michael Meechin <Michael.Meechin@ucf.edu>
Subject: TPGP Instrument Permission

Dr. Liu,

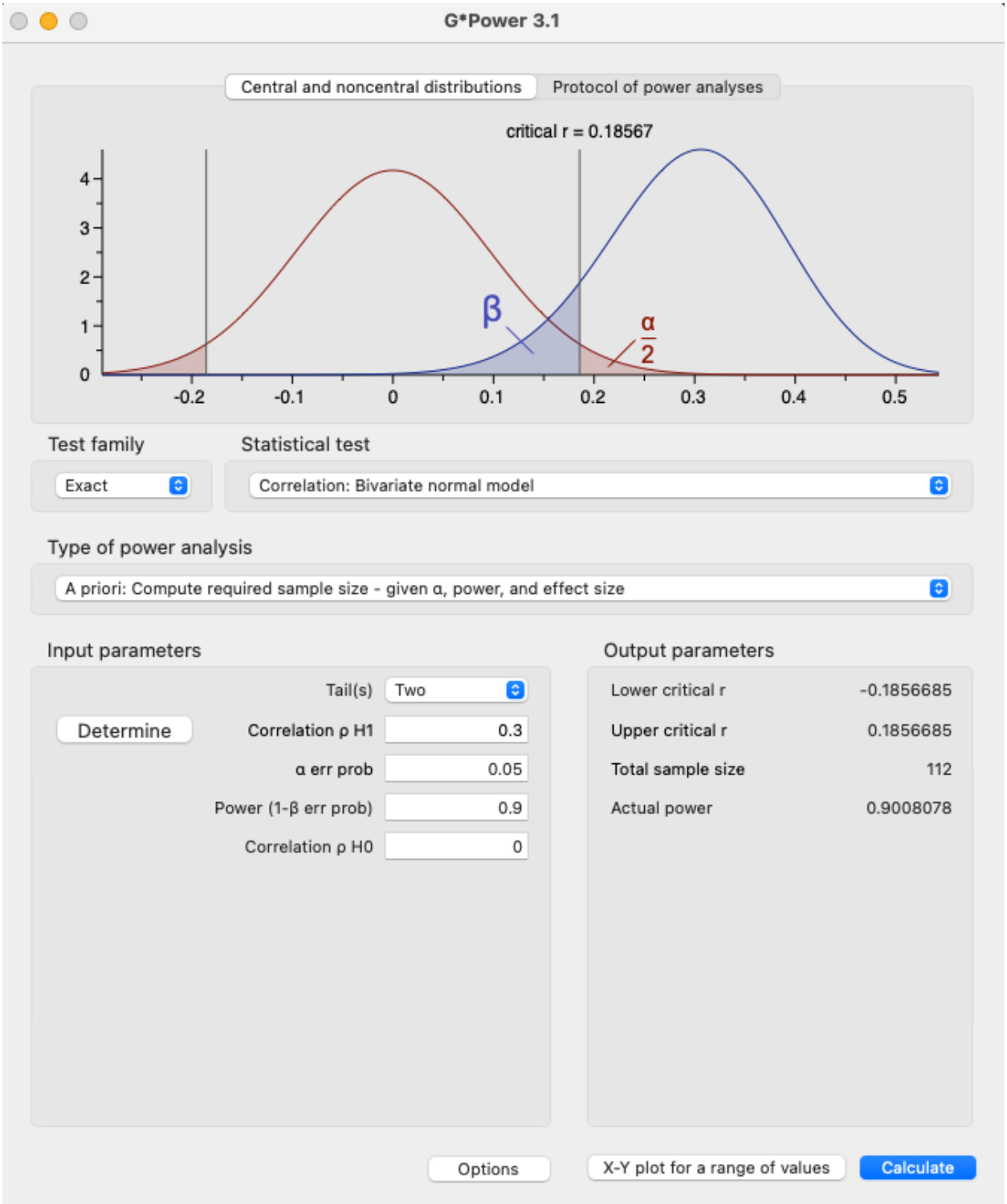
Please see the attached communication regarding permission to use the Teachers' Perceptions of Grading Practices instrument.

I look forward to your response. Thank you.

Michael Meechin
Ed.D. Educational Leadership Student
University of Central Florida

michael.meechin@ucf.edu
781-454-5700

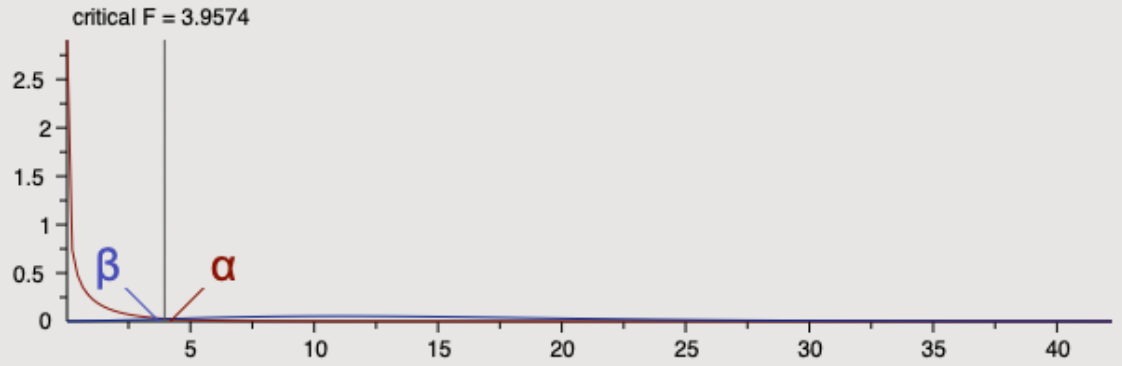
APPENDIX C
G*POWER CONFIRMATORY ANALYSIS FOR PEARSON


G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical r = 0.18567
4
3
2
1
0
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
β
α
2
Test family
Exact
Statistical test
Correlation: Bivariate normal model
Type of power analysis
A priori: Compute required sample size - given α, power, and effect size
Input parameters
Tail(s)
Two
Determine
Correlation ρ H1
0.3
α err prob
0.05
Power (1-β err prob)
0.9
Correlation ρ H0
0
Output parameters
Lower critical r
-0.1856685
Upper critical r
0.1856685
Total sample size
112
Actual power
0.9008078
Options
X-Y plot for a range of values
Calculate

APPENDIX D
G*POWER CONFIRMATORY ANALYSIS ANOVA 2 GROUPS

G*Power 3.1

Central and noncentral distributions | Protocol of power analyses



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input parameters

Determine

Effect size f	0.4
α err prob	0.05
Power (1-β err prob)	0.95
Number of groups	2

Output parameters

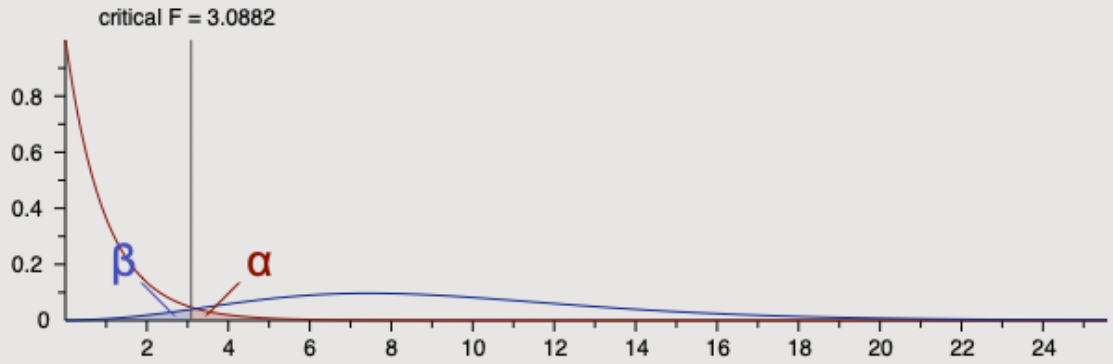
Noncentrality parameter λ	13.4400000
Critical F	3.9573883
Numerator df	1
Denominator df	82
Total sample size	84
Actual power	0.9518269

X-Y plot for a range of values | Calculate

APPENDIX E
G*POWER CONFIRMATORY ANALYSIS ANOVA 3 GROUPS

G*Power 3.1

Central and noncentral distributions | Protocol of power analyses



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Determine

Effect size f	0.4
α err prob	0.05
Power (1-β err prob)	0.95
Number of groups	3

Output parameters

Noncentrality parameter λ	16.3200000
Critical F	3.0882396
Numerator df	2
Denominator df	99
Total sample size	102
Actual power	0.9547295

X-Y plot for a range of values | Calculate

APPENDIX F
UCF IRB STUDY APPROVAL

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

November 28, 2023

Dear Michael Meechin:

On 11/28/2023, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study
Title:	A Correlational Study of Teacher Perceptions of Grading Methodologies and Student Grade Outcomes
Investigator:	Michael Meechin
IRB ID:	STUDY00005788
Funding:	None
Documents Reviewed:	• Meechin HRP-251-FORM, Category: Faculty Research Approval; • UPDATE 2023-11-15 Study 5788 Meechin-HRP-254-Explanation of Research.pdf, Category: Consent Form; • UPDATE 2023-11-15 Study 5788 Meechin-TPGPS Instrument.docx, Category: Survey / Questionnaire; • UPDATE 2023-11-17 Study 5788 Update-Meechin-HRP-255-Request for Exemption.docx, Category: IRB Protocol; • UPDATE 2023-11-26 Study 5788 Meechin-HRP-315-Survey Ad.docx, Category: Recruitment Materials;

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please submit a modification request to the IRB. Guidance on submitting Modifications and Administrative Check-in is detailed in the Investigator Manual (HRP-103), which can be found by navigating to the IRB Library within the IRB system. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Gillian Bernal
Designated Reviewer

APPENDIX G
UCF IRB STUDY CLOSURE

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

CLOSURE

May 3, 2024

Dear Michael Meechin:

On 5/3/2024, the IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	A Correlational Study of Teacher Perceptions of Grading Methodologies and Student Grade Outcomes
Investigator:	Michael Meechin
IRB ID:	CR00002927
Funding:	None, None
IND, IDE, or HDE:	None

The IRB acknowledges your request for closure of the protocol effective as of 5/3/2024. As part of this action:

- The protocol is permanently closed to enrollment.
- All subjects have completed all protocol-related interventions.
- Collection of private identifiable information is completed.
- Analysis of private identifiable information is completed.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Harry Wingfield
UCF IRB

REFERENCES

American Psychological Association. (2019). Addressing implicit bias in schools. Retrieved from https://www.apa.org/education/k12/implicit-bias

Arter, J., & McGaw, J. (2013). *Connecting assessment with instruction: Linking learning goals to meaningful learning experiences*. ASCD.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Asamoah, M. (2014). Re-examination of the limitations associated with correlational research. *Journal of Educational Research and Reviews*, *2*(4), 45-52. http://sciencewebpublishing.net/jerr/archive/2014/July/pdf/Asamoah.pdf

ASCD. (2012, February). Eight essential principles for improving grading. https://www.ascd.org/el/articles/eight-essential-principles-for-improving-grading

ASCD. (2012, February). Making the grade: What benefits students? https://www.ascd.org/el/articles/making-the-grade-what-benefits-students

Bangert, D., Vaughan, S., & Whitaker, C. (2022). Putting feedback into action. *Educational Leadership*, *79*(7), 16.

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

Barnes, M. (2018, January 10). No students don’t need grades. *Education Week*.

Bloom, B. (1968). Learning for mastery. *Evaluation Comment (UCLA-CSIEP), 1*(2), 1-12.

Bloom, B. (1981). *All Our Children Learning*. McGraw-Hill.

Bowie, L. (2021, April 8). As they investigate grading problems Baltimore school leaders defend system at city council hearing. *The Baltimore Sun*. https://www.baltimoresun.com/education/bs-md-augusta-fells-problems-20210408-npgkgnofl5g7vfpmwvx3tqk3wy-story.html

Briggs, J., & Peat, F. (1999). *Seven life lessons of chaos: Spiritual wisdom from the science of change*. Harper Collins Press.

Brimi, H. (2011). Reliability of grading high school work in English. *Practical Assessment, Research, and Evaluation, 16*(17). https://doi.org/10.7275/j531-fz38

Brookings Institution. (2023, March 14). The pandemic has had devastating impacts on learning. What will it take to help students catch up? https://www.brookings.edu/articles/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/

Brookhart, S. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*, 35-36.

Brookhart, S. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123-142. http://www.jstor.org/stable/1435460

Brookhart, S. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education*, *7*(4), 279-301. https://doi.org/10.1207/s15324818ame0704_2

Brookhart, S. (2013). Grading as a motivating practice: Strategies and techniques. *Educational Leadership, 71*(1), 42-47.

Brookhart, S. (2017). *How teachers grade*. Corwin.

Brookhart, S., Guskey, T., McTighe, J., & Wiliam, D. (2020). Eight essential principles for improving grading. *ASCD, 78*(1). https://www.ascd.org/el/articles/eight-essential-principles-for-improving-grading

Brookhart, S., Guskey, T., McTighe, J., & Wiliam, D. (2020). Grading in a comprehensive and balanced assessment system. National Panel on the Future of Assessment Practices. https://www.dylanwiliamcenter.com/wp-content/uploads/sites/3/2020/04/DW02-08-Grading-Policy-Paper-2020.pdf

Buckley, J., Letukas, L., & Wildavsky, B. (2018). *Measuring success: Testing, grades, and the future of college admissions*. Johns Hopkins University Press.

Cahen, P. (2019). Assessment and learning. *Educational Assessment, 24*(2), 112-129. https://pubmed.ncbi.nlm.nih.gov/23140801/

Carothers, C. (2017). Gender differences in teacher perceptions and grading practices: A meta-analysis. *Review of Educational Research, 87*(1), 137-170. https://www.tandfonline.com/doi/full/10.1080/09645292.2023.2252620

Center for Applied Special Technology (CAST). (2018). Universal design for assessment (UDL) guidelines. https://udlguidelines.cast.org

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Creswell, J. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.

Cross, L., & Frary, R. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72. https://doi.org/10.1207/s15324818ame1201_4

DeVellis, R. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Durm, M. (1993). An A is not an A is not an A: A history of grading. *The Educational Forum, 57*(3). https://doi.org/10.1080/00131729309335429

Edgeworth, F. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*(3), 599-635. https://www.jstor.org/stable/2339898

Education Trust. (2021). The case for competency-based education. Retrieved from https://edtrust.org/resource/competency-based-education

Education Week Research Center. (2023, July 11). Students aren't rebounding from the academic effects of the pandemic. https://www.edweek.org/leadership/student-achievement

Edutopia. (2023, January 23). 3 grading practices that should change. https://www.edutopia.org/article/transitioning-to-evidence-based-grading/

EdWeek. (2020a). The coronavirus spring: The historic closing of U.S. schools (a timeline). https://www.edweek.org/leadership/the-coronavirus-spring-the-historic-closing-of-u-s-schools-a-timeline/2020/07

EdWeek. (2020b). Grading students during the coronavirus crisis: What's the right call? https://www.edweek.org/teaching-learning/grading-students-during-the-coronavirus-crisis-whats-the-rightcall/2020/04

Elkins, D. (2016). Grading to learn: An analysis of the importance and application of specifications grading in a communications course. *Kentucky Journal of Communication, 35*(2), 26-48.

Erickson, J. (2010). Grading practices: The third rail. *Principal Leadership, 10*(7), 22-24, 26.

Farr, B. (2000). Grading practices: An overview of the issues. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age for standards* (pp. 1-22). Norwood, MA: Christopher-Gordon.

Feldman, J. (2018). *Grading for equity: Achieving fairness in the assessment of student learning*. ASCD.

Finkelstein, I. (1913). The marking system in theory and practice. *Educational Psychology Monographs, 10*. https://hdl.handle.net/2027/loc.ark:/13960/t9x06xh0t

Gelinas, L., Pierce, R., Winkler, S., Cohen, I., Lynch, H., & Bierer, B. (2017). Using social media as a research recruitment tool: Ethical issues and recommendations. *The American Journal of Bioethics, 17*(3), 3–14. https://doi.org/10.1080/15265161.2016.1276644

Gershenson, S. (2015, August 18). The alarming effect of racial mismatch on teacher expectations. Brookings. Retrieved from https://www.brookings.edu/articles/the-alarming-effect-of-racial-mismatch-on-teacher-expectations/

Gershenson, S. (2020). *Great expectations: The impact of rigorous grading practices on student achievement*. Thomas B. Fordham Institute. https://files.eric.ed.gov/fulltext/ED603248.pdf

Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. *Proceedings of the Joint Statistical Meetings*, 3455-3460.

Gleeson, P. (2019, June 11). Why minority students get bad grades: The Pygmalion effect. *MultiBriefs*. Retrieved from https://exclusive.multibriefs.com/content/why-minority-students-get-bad-grades-the-pygmalion-effect/education

Grindberg, E. (2014, April 7). Ditching letter grades for a "window" into the classroom. *Cable News Network.* Retrieved from http://www.cnn.com/2014/04/07/living/report-card-changes-standards-based-grading-schools/

Grissom, J., Kalogrides, D., & Loeb, S. (2015). The micropolitics of educational inequality: The case of teacher–student assignments. *Peabody Journal of Education, 90*(5), 601–614.

Guskey, T. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for mastery." *Journal of Advanced Academics, 19*(1), 8-31.

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

Guskey, T. (2017). *Fair assessment practices in education: Applying the concepts*. Corwin.

Guskey, T. (2018). *Formative assessment for differentiated instruction: Connecting assessment to student learning*. Corwin.

Guskey, T. (2018). *Grading for equity: Achieving fairness in the assessment of student learning*. ASCD.

Guskey, T. (2019, October). Grades versus comments: Research on student feedback. *Kappan*. https://kappanonline.org/grades-versus-comments-research-student-feedback-guskey/

Guskey, T. (2021, May). Undoing the traditions of grading and reporting. *American Association of School Administrators*. http://my.aasa.org/AASA/Resources/SAMag/2021/May21/Guskey.aspx

Guskey, T. (2022, October). Can grades be an effective form of feedback? *Kappan*. https://kappanonline.org/grades-feedback-guskey/

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

Guskey, T., & Bailey, J. (2010). *Developing standards-based report cards*. Thousand Oaks, CA: Corwin.

Guskey, T., & Link, L. (2018). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 25*(3), 303-320. https://doi.org/10.1080/0969594X.2018.1555515

Haladyna, T. (2019). Assigning a valid and reliable grade in a course. *IDEA.* https://files.eric.ed.gov/fulltext/ED598950.pdf

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. United Kingdom: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hemmeter, M., Santos, R., & Osgood, D. (2008). The role of teacher preparation in promoting evidence-based practice in schools. *Educational Researcher, 37*(9), 141-157.

Ho, A., Imberman, S., Lachance, J., & List, J. (2018). The persistent influence of racial bias on high-school grading practices: Evidence from a social experiment. *American Educational Research Journal, 55*(2), 170-191. https://scholar.harvard.edu/files/dmq/files/quinn-racial-bias-grading-eepa-2020.pdf

Hodara, M., & Cox, M. (2016). Developmental education and college readiness at the University of Alaska (REL 2016–123). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

Heubeck, E. (2023, March 10). Should students get a do over? The debate on grading and re-doing assignments. *Education Week*. https://www.edweek.org/teaching-learning/should-students-get-a-do-over-the-debate-on-grading-and-re-doing-assignments-deepens/2023/03

Impedovo, J., Kuh, G., & Tinto, V. (2001). Validity of high-school grades in predicting student success beyond the freshman year. *The Journal of Higher Education, 72*(4), 326-340.

Importance of validity and reliability in classroom assessments. (2023). *Marco Learning*. https://marcolearning.com/the-two-keys-to-quality-testing-reliability-and-validity/

Jimenez, K. (2023, March 2). SAT, ACT testing requirements paused during pandemic are now permanently optional at some colleges. *USA Today*. https://www.usatoday.com/story/news/education/2023/03/02/sat-act-test-optional-policies-covid-become-permanent/11385454002/

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education, 86*(2), 103–123.

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

Kline, R. (2015). *Thinking statistically: What research data means, and why it matters* (2nd ed.). Pearson Education.

Kohn, A. (1994). Grading: The issue is not how but why. *Educational Leadership, 52*(2), 38-41.

Kohn, A. (1999). From degrading to de-grading. *High School Magazine, 6*(5), 38-43. https://eric.ed.gov/?id=EJ585547

Laerd Statistics. (n.d.). Pearson product-moment correlation. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Laerd Statistics. (2017a). Two-way ANOVA using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

Laerd Statistics. (2017b). One-way ANOVA using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

Laerd Statistics. (2018). Pearson's product-moment correlation using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

Link, L. (2018). Teachers' perceptions of grading practices: How pre-service training makes a difference. *Journal of Research in Education, 28*(1), 62-91. https://files.eric.ed.gov/fulltext/EJ1168160.pdf

Link, L. (2023). Every teacher grades differently, which isn't fair. *The Conversation*. https://theconversation.com/every-teacher-grades-differently-which-isnt-fair-201276

Link, L., & Guskey, T. (2022). Is standards-based grading effective? *Theory Into Practice, 61*(4), 406-417. https://doi.org/10.1080/00405841.2022.2107338

Liu, X. (2004, October). The initial validation of teacher's perception of grading practices. Paper presented at the 2004 Northeastern Educational Research Association annual meeting, Kerhonkson, NY.

Liu, X., O'Connell, A., & McCoach, D. (2006, April). The initial validation of teachers' perceptions of grading practices. Paper presented at the 2006 Annual Meeting of American Educational Research Association (AERA), San Francisco, CA.

Liu, X., O'Connell, A., & McCoach, D. (2008, April). Teachers' perceptions of grading practices: A cross-cultural study in the U.S. and China. Paper accepted for presentation at the 2007 Annual Meeting of American Educational Research Association (AERA), New York, NY.

Lobo, M. (2005). Descriptive research is the bench science of nursing. *Western Journal of Nursing Research, 27*(1), 5-6.

Luschei, T., & Jeong, D. (2018). Is teacher sorting a global phenomenon? Cross-national evidence on the nature and correlates of teacher quality opportunity gaps. *Educational Researcher, 47*(9), 556-576. https://doi.org/10.3102/0013189X18794401

Marco Learning. (n.d.). Grades vs learning – Shifting attention to what's important. [Blog post]. Retrieved from https://marcolearning.com/ultimate-guide-to-grading-and-feedback/

Marzano, R. (2007). *What works in schools: Translating research into action.* ASCD.

Marzano, R., & Heflebower, T. (2011). *Setting and using learning goals to promote student achievement.* Alexandria, VA: ASCD.

Marzano, R., & Kendall, J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction.* ASCD.

McMillan, J. (2018). *Classroom Assessment: Principles and Practice for Effective Teaching (7th ed.).* Pearson Education.

McMillan, J. (2023). *Classroom assessment: Principles and practice for effective teaching.* Boston, MA: Pearson.

McMillan, J., & Nash, S. (2002, March 4). Teacher classroom assessment and grading practices decision making. [Report]. Retrieved from https://files.eric.ed.gov/fulltext/ED447195.pdf

McMillan, J., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 202-213. https://doi.org/10.1080/00220670209596593

McNeil, L. (2000). *Contradictions of School Reform: Educational Testing, Course Taking, and the Curriculum.* Routledge.

Merriam, S. (2009). *Qualitative research: A practical guide (2nd ed.).* Jossey-Bass.

Michaelsen, A. (2020, July 16). What's the point of a grade? The digital classroom, transforming the way we learn. [Blog post].

National Center for Education Statistics. (2024). Digest of Education Statistics. Retrieved from https://nces.ed.gov/programs/digest/d22/tables/dt22_208.20.asp

National Council on Teacher Quality. (2020). Teaching now and tomorrow: The preparation of teachers in 2020. Retrieved from https://www.nctq.org/research/teaching-now-and-tomorrow-preparation-teachers-2020

National School Boards Association. (2020, February). Accurate and equitable grading. [Policy brief]. Retrieved from https://publications.csba.org/issue/spring-2022/whats-in-a-grade/

Northern, A., & Petrilli, M. (2018). Foreword and executive summary. In *Grade inflation in high schools (2005-2016)* (pp. 4-7).

O'Connor, K. (2007). Formative assessment and self-regulated learning: Developing student self-assessment competence. SAGE Publications, Ltd.

O'Connor, K., Jung, L., & Reeves, D. (2018). Gearing up for FAST grading and reporting. *Phi Delta Kappan, 99*(8), 67-71. https://kappanonline.org/oconnor-jung-reeves-gearing-up-for-fast-grading-reporting/

Organisation for Economic Co-operation and Development [OECD]. (2016). What are the incentives for people to invest in education? Retrieved from https://www.oecd-ilibrary.org/sites/dcabb78b-en/index.html?itemId=/content/component/dcabb78b-en

Perera, R., Carlson, D., Domina, T., Carter, III, J., McEachin, A., & Radsky, V. (2020, February 24). Can School Choice Support District-Led Efforts to Foster Diverse Schools? Retrieved from https://www.brookings.edu/articles/can-school-choice-support-district-led-efforts-to-foster-diverse-schools/

Popham, W. (2021). *Grading for effective learning: Why it matters more than ever.* Alexandria, VA: ASCD.

Popham, W. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: ASCD.

Reeves, D. (2004). The case against zero. *Phi Delta Kappan*, *86*(4), 324-325. Retrieved from https://www.researchgate.net/publication/285846142_The_Case_against_the_Zero

Reeves, D. (2011). *Elements of grading: A guide to effective practice.* Solution Tree Press.

Reeves, D., & Feldman, J. (2020). Grading during the pandemic: A conversation. *ASCD, 78*(1). Retrieved from https://www.ascd.org/el/articles/grading-during-the-pandemic-a-conversation

San Miguel, M. (2023). Frontiers in complex systems. *Frontiers in Complex Systems*. https://doi.org/10.3389/fcpxs.2023.1080801

Schillingburg, W. (2016). Understanding validity and reliability in classroom, school-wide, or district-wide assessments to be used in teacher/principal evaluations. Retrieved from https://www.azed.gov/sites/default/files/2016/10/Presentation_Resource_Understanding%20Reliability%20and%20Validity%202.pdf?id=57f6d9b3aadebf0a04b2691a

Siegenfeld, A., & Bar-Yam, Y. (2020). An introduction to complex systems science and its applications. *Hindawi Complexity, 2020,* 1-16.

Six, T. (2022, August 25). Parent sues Fayette County Public Schools, takes issue with grading in virtual classes. *Lexington Herald Leader*. https://www.kentucky.com/news/local/education/article264905424.html

Spencer, K. (2017, August 11). A new kind of classroom: No grades, no failing, no hurry. *New York Times*.

Stanley, G., & Baines, L. (2001). No more shopping for grades at B-mart: Re-establishing grades as indicators of academic performance. *The Clearing House, 74*(4), 227–230. http://www.jstor.org/stable/30189667

Starch, D. (1913). Reliability and distribution of grades. *Science, 38*(983), 630–636. http://www.jstor.org/stable/1640875

Starch, D., & Elliott, E. (1912). Reliability of the Grading of High School Work in English. *School Review, 20,* 442-457. https://doi.org/10.1086/435971

Steward, R., Hill, M., Neil, D., Pritchett, T., & Wabaunsee, A. (2008). What does GPA in an urban high school actually mean? *Educational Considerations, 36*(1), 11-16. https://coe.ksu.edu/EdConsiderations/issues/EdConsF08toc.pdf

Stiggins, R., & Chappuis, J. (2005). *Chasing learning: How the third revolution in education will change everything for students*. San Francisco, CA: Jossey-Bass.

Stiggins, R., Frisbie, D., & Griswold, P. (1989). Inside High School Grading Practices: Building a Research Agenda. *Educational Measurement: Issues and Practice, 8*(2), 5-14. https://doi.org/10.1111/j.1745-3992.1989.tb00315.x

Strobel Education Publishing. (2023, January 11). Leading change with effective grading practices. Retrieved from https://strobeleducation.com/onsite-pd-training/standard-based-grading-assessing/

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://doi.org/10.1177/1356336X13517437

Tomlinson, C., & McTighe, J. (2006). *Integrating differentiated instruction & understanding by design: Connecting content and kids*. Association for Supervision and Curriculum Development.

Wertheimer, A. (2014). Non-completion and informed consent. *Journal of Medical Ethics, 40*(2), 127-130.

Wiliam, D., & Black, P. (1988). Assessment and classroom learning. *Educational Leadership, 46*(5), 7-11.

Will, M. (2021, June 22). The complicated, divisive work of grading teacher-preparation programs. *Education Week*. https://www.edweek.org/teaching-learning/the-complicated-divisive-work-of-grading-teacher-preparation-programs/2021/06

Zhang, X., Song, L., & Wang, M. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications, 10*(1), 74. https://www.nature.com/articles/s41599-023-02385-4