2023

Impacts of Standards-based Grading and Alternate Grading Strategies on Student Success

Brian Remington
Bethel University

IMPACTS OF STANDARDS-BASED GRADING AND ALTERNATE GRADING STRATEGIES ON STUDENT SUCCESS

A MASTER'S THESIS SUBMITTED TO THE FACULTY OF BETHEL UNIVERSITY

BY

BRIAN REMINGTON

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE IN MASTERS OF ARTS

DECEMBER 2023

BETHEL UNIVERSITY

IMPACTS OF STANDARDS-BASED GRADING AND ALTERNATE GRADING STRATEGIES ON STUDENT SUCCESS

Brian Remington

December , 2023

APPROVED

Thesis Advisor: John Bergeland, Ph.D.

Program Director: Molly Wickam, Ph.D., MBA

Abstract

Standards-based grading has changed how schools have graded its students for the past few years. Under standards-based grading students receive several grades for each benchmark skill that the student displays their knowledge of. This is a huge departure from what I have always known through my own education. For my Master's thesis, I reviewed the recent literature around standards-based grading and other grading experiments. I wanted to see what the effect of the changes have been, and more importantly what advice can be drawn from the recent research literature on the subject. Overall, teachers' and students' experiences with alternative grading philosophies were found to be positive.

Table of Contents

List of Tables and Figures

CHAPTER I: INTRODUCTION

Grading is an important part of education. A good grading system should strive to fairly assess students as well as provide meaningful feedback to students, teachers, and others (Kunnath, 2017; Guskey, 2011; Marzano & Hefleblower, 2010). Despite its importance, grading practices have been static for much of our educational system's history. However, this has changed somewhat recently, as schools are starting to tinker with their grading systems. These changes have been implemented at different speeds and to varying extents in different school districts, subjects, and grade levels. In both of my field placements and in my student teaching, the school district changed from traditional grading practices to standards-based grading. I saw firsthand how these practices were different from the grading practices that I experienced as a student. I had my own thoughts about where the new practices succeeded and where they fell short. I also spoke with teachers and administrators about the two grading systems.

I have been in favor of standards-based grading since I heard about it. My own experience in the education system has not always been consistent. Early in my academic career I struggled under traditional grading practices. During periods of my youth, I didn't care about school or grades and missed my fair share of assignments. I also remember feeling dissatisfaction toward grading, because I felt it rewarded students for being diligent and not for being better at the subject area. I now realize these thoughts were immature and I wish that I had applied myself more. But, this experience has left me open to reforms in grading. For this reason and others, my thesis will serve as an opportunity to better understand standards-based grading and its impacts in school.

Problems of Traditional Grading Systems

Before addressing grading reforms, it is helpful to understand some of the criticisms of traditional grading. O'Connor & Wormeli (2011) stated the criticism of traditional grading and identified four areas where it falls short: accuracy, consistency, meaningfulness, and supportiveness of learning.

O'Connor & Wormeli (2011) argued that traditional grading was not accurate for a few reasons. First, grades included nonacademic factors (e.g., a geometry student being rewarded for having an organized notebook). Including these nonacademic factors alongside real competency-based standards make grades less of a clear indication of students' geometry competence. Second, grading included group work, which made it unclear what each individual student knew. Third, scores were averaged, which gave an unclear picture of student achievement. If a student receives an "F" at the beginning of the semester and an "A" at the end, this could mean that the student receives an "C" for the class. This is not accurate for the student, as the student does not show average mastery of the material. Instead, the student performed exceptionally well in some ways and very poorly in others. It could also mean that the student struggled early, but now knows the material and is able to show their level of understanding. If this is the case, the "A" grade is the correct grade to show the students skill level and understanding at the end of the course. For this reason, according to standards-based grading philosophy, early formative work should not be given the same weight as later summative work. O'Connor & Wormeli (2011) stated, "It's unethical and inaccurate to include in a grade digressions in performance that occur during the learning process, when a grade is supposed to report students' mastery at the end of that process" (p. 40). Fourth, receiving a zero on an assignment holds undue weight on a student's grade as a whole. Under standards-based grading's

four-point scale, each point has equally skewing influence. If a student receives a one or zero, the student can still turn their score around. When a student receives a zero out of a 100 point scale, the authors feel that student tends to feel despair. It will take many full point assignments to bring back a grade, and the student may feel that it is too much from which to recover.

Another problem with grading is consistency between classes. Ideally, a grade by one teacher should correlate to a grade of another teacher in the same class, but this is not always the case. The authors argue that this is made worse by lack of clarity and focus as to what is being graded and the 100-point scale, which allows for too much teacher discretion.

Standards-based grading is meant to address these failings. Each standard is clearly graded individually, which gives clarity to the process. If there is group work in the class, each student is graded on their own contribution. Students are not graded for the learning process. Formative work is not graded heavily and there are ample opportunities to retake assessments. There customarily is no time limit on late work, and each standard is graded individually according to a four-point scale.

Myths about Standards-Based Grading

There are a lot of misconceptions about standards-based grading. Wilcox & Townsley (2022) rebutted common myths about standards-based grading practices with the research done in the field. These common myths allege that standards-based grading reduces rigor and motivation of the students. There are also myths regarding the difficulty of implementing the reforms into the classroom.

One myth claims that standards-based grading will reduce rigor in the classroom. This is partly due to a problematic definition of rigor. Wilcox & Townsley (2022) argued that classrooms should move from one of rote memorization to one where a variety of skills are being

taught. Standards-based grading practice encourages a classroom where students learn effectively. In addition, standards-based curriculum removes the fluff from the classroom. There is no longer credit given for extra credit, participation, or completion which inflate students' grades.

There is a myth that students who learn under standards-based grading will be disadvantaged at the college level, because college courses instead use traditional letter grades. The studies show that students who learned under standards-based grading perform well at college. There is strong theory to suggest that standards-based grading's emphasis on understanding and skills translate to college success. The myth that parents and students will not understand the grade book is unfounded. Just because it is different from what they may be used to does not mean they will not understand it. An aim of standards-based grading provides a better picture of what the student does and does not know.

Recommendations under Standards-Based Grading

Marzano & Heflebower (2011) reviewed existing literature and provided recommendations to teachers in grading.

The first recommendation was to get rid of the omnibus grade, which refers to traditional grading whereby the student receives an aggregate grade. The main problem of the omnibus grade is that it is not immediately clear to students what they know and where they fell short. Although teachers rarely give a grade for student behavior and conscientiousness, these qualities affect final grades. Under standards-based grading, there is less of an emphasis on completion of homework and participation, which should reduce the importance of these non-standards-based factors. Standards-based grading clearly separates grades into their separate standards, and focuses on student performance regarding those standards. In a pure standards-based grading

system there is no final one letter grade. Instead each standard is clearly listed with the student's grade listed for each step. Marzano & Heflebower (2011) provided the example of a report card under standards-based grading, which is provided in Figure 1.

Figure 1

First Quarter Report for a Middle School Mathematics Student

Measurement Topics	Score	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Number systems	2.5								
Estimation and mental computation	1.5								
Ratio/Proportion/Percent	2.0								
Patterns	3.5								
Equations	2.5								
Data Analysis	1.0								

The darker part of the report card represents the students performance at the beginning of the grading period and the lighter parts show the student's performance at the end of the grading period. In practice very few schools can or should move completely to a standards-based grading environment. High schools and middle schools still need to assign letter grades for purposes of displaying each student's performance in relation to the other students, which is useful for both higher education and employers. Marzano and Heflebower (2011) state "if public pressure demands that students receive an overall grade or percentage score, a school or district can still employ the benefits of the approach shown in Figure 1 [the sample report card provided] by including the bar graphs of the report card, along with traditional omnibus grades" (p. 36). They then provide the recommended conversion found in Table 1.

Table 1

Suggested Conversion from Four-Point Scale to Letters

A	A-	B+	B	B-	C+
3.51-4.00	3.00-3.50	2.84-2.99	2.67-2.83	2.50-2.66	2.34-2.49
C	C-	D+	D	D-	F
2.17-2.33	2.00-2.16	1.84-1.99	1.67-1.83	1.50-1.66	0-1.49

Marzano and Heflebower (2011) recommended replacing the 100-point scale with a four-point scale. Under a four-point scale a score of two would be given when a student is able to show mastery of simpler content; a score of three shows that the student met the target standard and a score of four shows that the student was able to show an understanding of more complex content. This makes the score easier to understand at a glance.

Assigning full points only when students went above and beyond makes good theoretical sense. It is intended to foster a growth mindset which encourages them to reach above and beyond instead of just trying to do enough to receive an "A." It is important to note that this is a sizable departure from our traditional understanding of grading. Scoring less than full points for an assignment that met all of the criteria is something that is difficult for students and parents to accept. This is made better by the conversion which still awards a low "A" for a 3 score, but is still difficult. In my experience of applying standards-based grading during student-teaching, I heard many complaints about this feature from students.

Marzano and Heflebower (2011) also recommend expanding the assessment options available to students, so that they are able to show their mastery in different ways. It also advised not to grade practice and allow students to update their scores on previous measurement topics, as they reach higher levels of mastery in the course.

Importance to Minnesota and Beyond

This topic is especially important for Minnesota, which has seen many school districts transition to a standards-based grading curriculum. In both my observations and student teaching experience, I was in schools that employed standards-based grading. I am currently teaching adults, but if I end up teaching high school or middle school, I anticipate there being a good chance that I am either working with standards-based grading or being part of a discussion to transition to standards-based grading because it is becoming more common in Minnesota.

Guiding Question and Outline

The guiding research question for this thesis is: What is the impact of standards-based grading and other alternate grading strategies on student success? The research will also address the following three items: 1) What has been the feedback from both teachers and students? 2) What are the short-term and long-term effects on students who have learned under a standards-based grading system? 3) Has changing the grading system ameliorated some of the problems we see in grading, especially in grading non-STEM subjects?

In this paper, Chapter 1 introduces the paper and gives its rationale. Chapter 2 discusses the individual reviewed studies in my research. Chapter 3 provides an analysis and evaluation of the research and includes thoughts on future personal implementation and future research recommendations.

CHAPTER II: LITERATURE REVIEW

All of my articles were found through ERIC. I tried searching JSTOR, Google Scholar, and ProQuest, but I found few articles and started searching just ERIC and EBSCO Host. It is

only upon writing this section, reviewing the provided sample, and revisiting my process that I now realize that I was using EBSCO Host incorrectly, as EBSCO Host was only searching ERIC. Sometimes ERIC did not have the full text of an article, in which case I would attempt to locate the full text in Google Scholar or from a database that came up in a simple Google search.

My first searches included both "standards-based grading and high school." I checked the boxes for both scholarly articles, peer reviewed, and full-text. From the initial search results, I chose the articles that were about high school and were non-STEM subjects. This approach netted very few quality articles, but it served to get me started.

From there, I dropped the search term "high school" and eventually I just searched "grading." This netted quite a few results, but many were not included in my thesis as they were unrelated. So, I winnowed the articles down by how closely they fit the concepts of my initial guiding question, which included "standards-based grading," "high school," and "non-STEM subjects."

I was hesitant to expand my search to include math courses, for I felt my focus on non-STEM subjects gave my focus some uniqueness. In light of this, I ignored math courses, and instead broadened my search to include other grading reforms or alternate grading practices. I included articles about proficiency-based grading, equity-based grading, and ungrading. I included articles that focused away from grading in traditional ways including peer-based grading, portfolios and practices such as these. Similarly, to preserve the focus on non-STEM subjects I included a pair of studies both looking at inconsistencies of grading writing.

At the end of my researching and writing process, I included standards-based grading articles focusing on the college level, and standards-based grading articles focused on math and science. These were a departure from my initial plan, but they were necessary for me to get the

required number of sources and these still provide valuable information to my thesis as a whole.

After completing my literature review of the articles, I organized the articles into what I felt were dominant themes. These corresponded with the main points of the articles that I reviewed. Therefore, Chapter II is divided into five sections 1) Experiences with Standards-based grading and its impacts on mindsets, 2) Experiences with alternate graded, ungraded, and student-led graded course, 3) College readiness, 4) Impact on struggling students, and 5) Challenges with grading, communicating directions, and assigning grades.

Experiences with Standards-Based Grading and its Impacts on Mindsets

Knight & Cooper (2019) sought to understand teachers' thoughts and experience with standards-based grading. The study observed and received feedback from seven teachers. The teachers were drawn from five high schools that use standards-based grading. Of the teachers' specialties, there was a mix of high school grade levels 9-12 and subjects. In total, the teachers taught science, English/Language Arts (ELA), special education, math, and music. Each participant took part in three 60–90-minute interviews. The first interview asked the teacher about their experience as a student and as a teacher to provide context. The second interview asked how their planning, instructing, and assessing had changed under standards-based grading. The third interview asked teachers to make meaning out of their experience. The data collected from these interviews were triangulated with observation data and classroom materials as well as data regarding student achievement, behavior, and school policy. One classroom period was observed and the materials were collected from that period.

Teachers stated that abiding by the recommendations of standards-based grading made their planning, instructing, and assessing more purposeful. The seven participants also said that standards-based grading had a positive impact on communicating to students, because there were

clear goals for each lesson. On exams, students received clear feedback. Because of clear rubrics, students could articulate things like, “I am really good at introductions, but I struggle with conclusions” (Knight & Cooper, 2019, para. 27). The clear feedback also led to improvements in parental involvement because they were better able to understand their children’s performances. The teachers cited improvements to the learning environment and student behavior. Every teacher agreed that it created an environment that was more conducive to learning. Allowing students to retake exams led students to better accept mistakes, but to be unaccepting of failure. Being comfortable with mistakes and clear feedback led to increases in the growth mindset of students as well as students taking accountability for their own learning. Overall, students were able to take risks and grow. Teachers had a few complaints in the implementation of standards-based grading. The schools used it alongside traditional letter grades. Converting standards-based grades to traditional grades led to certain students receiving higher grades and other students receiving lower grades, which led to different feelings amongst parents and students. The teachers also spoke of an inconsistent application of standards-based grading principles between teachers in the same school. Some teachers especially felt frustrated at not being able to reward work ethic, professionalism, and skills such as these.

Peters et al., (2017) used a large-scale survey to collect at a high school of nearly 500 students. This high school was situated in a predominantly White, educated, upper socioeconomic community in Iowa of just over 2,000 people. It is located near a public university and only 10% of its population qualify for free or reduced lunches. This particular school district had already implemented standards-based grading in the primary grades, but was now in the process of implementing it at the secondary level as well. Across the community, many families expressed concern over the change. To better understand the community’s

concerns and the concerns of families in general, the researchers asked students to complete a survey. Peters et. al, (2017) received questionnaire responses from 376 students in the fall of 2012 and from 230 students in the spring of 2013. Data was received anonymously, so it is unknown how many students may have submitted their information twice. The responses were both quantitative in response to Likert style numbered responses, as well as open-ended. The questions eliciting open-ended responses were (1) What do you see as the biggest weakness/disadvantage of standards-based grading? (2) What do you see as the next biggest weakness/disadvantage of standards-based grading? (3) If I could change one thing about standards-based grading, it would be . . . . (Peters et. al, 2017 p. 14).

Criticism fell into five themes: (a) the SBG implementation process, (b) grading issues, (c) preparation for university and future employment, (d) social issues, and (e) issues related to current teaching, learning, and motivation.

Most students had some concern about the new grading system. For implementation, students expressed concern over inconsistencies between application by teachers and the lack of application of the concepts by some teachers. For difficulty, students were concerned that it was harder to achieve high grades under standards-based grading, because they could not use homework to cushion bad test scores. As for college preparation, students were worried that their transition to traditional grading at the college level would be impacted. As for motivation, students stated that it was too easy to get by without trying. For example, if a student did not do an assignment, there wasn't as much of an immediate consequence, and if the student did not study sufficiently, they could just redo the exam. This made it hard for the student to be motivated in the class. The study concluded that these were valid concerns, but that all of the concerns could have been/could be addressed by proper implementation of standards-based

grading.

In order to better understand standards-based grading from the teacher perspective, Williams (2023) asked three questions:

> What are middle school teachers' perspectives of differentiated instruction in concert with standards-based grading? How do teachers perceive the role of differentiated instruction in a standards-based grading evaluative process in middle schools? How do teachers perceive the ways in which standards-based grading influences differentiated instruction in middle schools? (p. 130)

In order to answer these questions, they performed semi-structured interviews with teachers in Westwood Middle School, a 12,000-student public school in the southern United States. It had employed a traditional grading system, but transitioned to standards-based grading in 2018. The participants of the study included middle-school teachers in grades 7 or 8. Three teachers taught ELA, one taught math, two taught social studies, and one taught science. Teacher participants ranged from 3-22 years of teaching experience and 2-6 years of implementation of standards-based grading. Each participant was interviewed twice. Their responses were put through a multiple-step coding process. First, their responses were analyzed to find common phrases and then those were clustered into themes. The second step comprehensively reviewed these themes as they related to each other and to assessment as a whole.

Overall, there were mixed feelings about how differentiation interacted with standards-based curriculum. Differentiation benefits standards-based grading in a few ways. Many teachers felt that standards are not made to fit the needs of all children, and differentiation is needed to help all students to meet the standards. Also, they felt that standards give a clear target by which teachers can assess students' strengths and weaknesses. This then helps them to

differentiate the curriculum appropriately. So overall, both standards and differentiation help each other.

There were some common themes about how differentiation and standards-based education did not work well together. In regard to assessing students, differentiation understands that students should be able to demonstrate their understanding in different ways, but the standards are not so flexible. Standards alone do not give a clear roadmap as to how students should be taught or how they should be assessed. Also, differentiation teaching while also progressing towards standards are two different goals. Trying to reach both goals puts the teacher under a lot of pressure, which impacts the instruction.

In Scarlett (2018), a professor detailed his own implementation of standards-based grading at the post-secondary level. He started with a review of literature for his own purposes so that he was able to adhere to its core principles. He instituted standards-based grading in four steps: clarifying curricular aims, choosing standards-based assessment evidence, weighting evidence, and arriving at a final grade. He and his students reflected on the changes in grading, and he offered advice for other professors who are similarly inclined to change their grading to standards-based.

To change his grading practices, he first identified learning targets for the class. He also identified non-academic skills that were important to his course, such as turning assignments in on time. Lastly, he wrote down his criteria for grading each. From these learning targets, he wrote assignments. He found that his assignments did not need to change much. Despite not changing significantly, his assignments were improved by looking at them through a goal-based lens. For weighting assignments, he identified which assignments were formative and which were summative. He also sought not to give too much weight to grades that were early in the

semester, while the student was still learning on how to address the course. In arriving at the final grade, he chose to depart from the advice of standards-based advocates. He assigned points for the process score. He reasoned that he was teaching college students. He worried that students may do well in the course without conducting their academics as they are expected to at the college level. Nevertheless, process scores were reduced from his earlier grading practices. At the end of the course, product grades were multiplied by .80 and process scores by .20.

Overall he felt his changes positively affected the course. The feedback from his class was mixed but was more positive than in previous semesters. He and his students agreed that the course's aims were clearer and feedback was stronger. Re-assessment opportunities helped students by forcing them to re-look at old material. Changing to standards-based grading impacted his course but equally important the process of changing the grading made him a better teacher. The process gave him a new lens to look at his own policies. In the process the curriculum was also made stronger and more focused on the desired outcomes.

Meyer, et al. (2009) sought to understand the relationship between motivation and achievement. The study took place in New Zealand where there is a National Certificate of Educational Achievement (NCEA) which has a standards-based criterion-referenced assessment. This is an assessment that is taken by all students, but for those seeking higher education, the assessment is important for their further academic prospects. The study's population was drawn from 20 diverse schools in New Zealand. The study gathered data through survey responses (elicited 2-4 weeks prior to the start of student final examinations) from 3,790 students who were in their final three years of schooling. They were diverse in race, language, population size, and school size. The survey had many questions on student motivation. The student would rate each item according to a four-point scale (1 = this doesn't matter to me at all; 4 + this is a big factor in

making decisions). The relationships between the responses and the students' results were then analyzed.

Meyer (2009) sorted questions in the survey into different categories, which allowed the researchers to draw quantitative data from the survey. Scores from different categories were used to compare students' mindsets. Some questions were sorted into two categories: doing my best vs. doing just enough. Other responses were sorted into the following categories under the heading "influence on subject choice" including utility, expediency, and interest. Utility was designated for those that students felt the subject was important for their future aspirations. Expediency was high for students who were motivated by the easiest course of action. Lastly, interest was high for those students who were taking a subject because of their personal interest in it.

Overall, there was a strong correlation between high scores and those who were aiming to do their best as opposed to those who did just enough. Doing my best was the strongest predictor of variance in grades. It accounted for 23%. Doing just enough accounted for 8% variance. There was also a small increase in grades for those who were participating in the class for interest as opposed to expediency. This accounted for 0.3% variance.

Lewis (2022) examined the impacts of standards-based grading on students' mindset and test anxiety. Specifically, the study asked three research questions: "(1) Does students' test anxiety change over time in an SBG class and their other classes? (2) Do students' growth mindset and achievement goal orientation change over time in an SBG class and their other classes? (3) Is the magnitude of any differences constant across demographic groups, particularly those underrepresented in mathematics?" (p. 7).

The study surveyed students at the start and end of three different courses (i.e., Calculus

II, Linear Algebra, Differential Equations) taught by two different professors at one university in South Alabama. The survey asked the students to rate their agreement with the statements as they applied to this course and as they applied to their other courses. This would provide a difference as the classes taught in this study all employed standards-based grading but most of their other courses did not.

After receiving these survey results, the study's authors re-analyzed the data from previous similar studies. That study was from previous semesters at the same university and asked the same questions. That study included questions related to growth mindset.

Overall, 74 of 115 students (14 = female, 63 = male) completed both the surveys in Study 1 and 94 of 221 participants completed the re-analyzed surveys in Study 2. The study did show that the students had lower anxiety in their standards-based graded courses. This was especially true in the survey administered at the end of the course and showed specifically reduced post-semester test anxiety. The reduction in anxiety was largest among the female students. Female students have consistently shown higher levels of school related anxiety (von der Embse et al., 2018). Their anxiety in these courses decreased over time. In the post-semester survey, the anxiety between women and men was equal.

The study did not find any difference between the courses in regard to growth mindset. The author stated that expecting large changes in a mindset from a single class may have been too ambitious. Similarly, the study did not find much change in students' performance goal orientation. There was a modest difference in the students' mastery avoidance goal orientation (see Table 2). This may suggest that students were more motivated by other more productive goal orientations, but evidence of those increases were not found directly.

Table 2

Test Anxiety and Achievement Goals Means

	Pre-Semester	Post-Semester (Other Classes)	Post-Semester (SBG class)
Test Anxiety	11.93	12.52	11.46
Mastery Approach	13.60	13.34	13.38
Mastery Avoidance	11.84	12.66	12.00
Performance Approach	12.42	12.73	12.78
Performance Avoidance	12.03	12.78	12.84

Experiences with Alternate Graded, Ungraded, and Student-Led Graded Course

Percell (2019) collected data from five high school teachers, but the study focused on one teacher participant who stood out as especially encompassing ideals of democracy and social justice. This teacher served as a case study for common themes relating to democracy, citizenry, and justice. This commitment to these themes was impacted by his unorthodox grading system. This is discussed below.

This teacher, who the study referred to as Simon, taught social studies at Middleton High School. Middleton draws from a community that is 22% from lower socio-economic backgrounds and 10% of whom have individualized education plans (IEPs).

Simon did not employ a traditional grading system. Instead, the students were assessed in three areas: participation, progress, and performance. The students could earn a minus (unsatisfactory), a checkmark (sufficient), and a check plus (outstanding). He regularly conferenced with his students about where the students stood in his class, so they could receive more detailed feedback. His assessments of their work counted for 50% of the grade; these were not an authoritative statement by the teacher, but they were determined through a conference in

which students felt they were a part of the process.

After analyzing the data specific to Simon, the study identified four democratic considerations regarding his grading system: freedom, democracy, critical thinking, and civility. Freedom was achieved because students were not simply thinking about what answer would give them the most points, because they were a part of the assessment process and could make their case about their answer. This allowed the student to mull over answers before thinking how they felt it would be best to answer. Democracy was achieved because the classroom placed the teacher in a less-authoritative position. Critical thinking was fostered because the students were encouraged to answer questions with multiple solutions and were encouraged to question other parts of the class. Civility was achieved as the open dialog between student and teacher led to better relationships. Without the usual competitiveness of grading, Simon reported a cooperative atmosphere.

Muho & Taraj (2022) focused on students who were learning the English language in three high schools in Albania. It posited that a departure from traditional grading may lead to better results; traditional grading in this context refers to assessment based on whether students could demonstrate knowledge on tests. Muho & Taraj (2022) studied the effects on motivation to learn when students were assessed through a combination of portfolio use, self-assessment, and peer-assessment by extrapolating data from survey responses. The researchers distributed 300 surveys to students from three main high schools in Albania. No information was given as to how the students were selected. Of the 300 responses, only 278 of them were used in the study. The excluded responses had no relevant data to include in the study. The survey employed a Likert scale ranging from strongly disagree to strongly disagree. The survey included questions on portfolios directly, such as, “Students‘ portfolio as an assessment practice has changed my

way of thinking when I face problems;" questions about self-reflection, such as, "I feel involved in the learning process;" questions about peer review, such as, "I analyze myself and my peer at the same time;" and questions about teacher instruction, such as, "Teacher's questions stimulate debates and comments."

The study (both observation and from a review of the literature) found that the combination use of portfolio, self-assessment, and peer assessment increased motivation of the students. Students self-reported improvement in their own autonomy. They were able to take charge of their learning in ways that traditional grading practices did not allow. The students found that they learned more about themselves through the process as they were able to better assess their own strengths and weaknesses. In assessing themselves and others, the students were learning a skill that they would continue to apply to their English study as they moved further along it. This autonomy, the focus on themselves, and on developing skills of self-reflection led to increased motivation when compared to a system of top-down grading practices.

Sadler (2006) asked two questions regarding peer grading. Can student-grading be a valid substitute for teacher grades? And, is student grading used as a pedagogical tool for student learning? In order to examine these issues, the study conducted experiments in four seventh-grade, general-science classrooms. Each class experienced a different, randomly assigned grading method. "These interventions were: control (teacher-only grading), self-grading (plus teacher grading), and peer-grading (plus teacher grading)" (Sadler, 2006, p. 11). The classes involved in the study had been employing peer-grading similar to the grading set forth in this study since the beginning of that school year. As such, they were comfortable and experienced in assessing their classmates.

The test to be graded included nine fill-in-the-blank items, seven classification tasks,

thirteen matching questions, and five constructed-response items. This was intended to give both easy-to-grade questions as well as difficult-to-grade questions. Students spent a whole class period creating and applying a rubric to be used. Each student then graded two anonymous students' assessments. In the first assessment that they graded, they rated a peer. For the second assessment, they self-graded. In the second, students also completed a feedback form so that they could record their thoughts on the peer review process. After the grading, students took a second unannounced assessment to see if they had learned from the process; this was compared to a control group that did not participate in peer and self-grading activities.

Comparing both peer-grading and self-grading to teacher-grading led to a few observations. One, self-grading led to an average increase of 1.9 over the grade given by the teacher; whereas peer-grading led to an average decrease 3.3 points below teacher grades with two substantially low outliers. Two, these scores were not uniformly different. Three, poorly performing students (especially) tended to underrate themselves, and better performing students tended to be underrated by their peers. Four, despite these somewhat consistent differences and trends, the scores otherwise showed a high degree of correlation with the scores given by the teacher. The results of the second test suggested that self-grading was a useful learning tool. In addition to performing better on the test, students reflected that they felt they learned from the self-grading and peer-grading.

Guberman (2021) examined the feedback from his own class, where he experimented with ungraded assignments and self-graded final grades. The course itself was an upper-level elective, titled Society, Culture, and Rock and Roll. This was taught online to fifty students at a STEM-focused university over an eight-week summer session. This course had previously been taught with more traditional grading practices.

Instead of backward-designing the course around objectives, the course emphasized communication, research skills, and metacognitive learning skills. Instead of receiving a grade, the students would assess their own level of accomplishment for each outcome. The course was built around each student sharing five presentations around specific tasks, such as analyzing a song or presenting any topic connected to a specific decade of music. These presentations were shared in consistent groups of 12 students. “Students were expected to explain why they thought their participation and engagement demonstrated meaningful responses” (Guberman, 2021, p. 90). In addition to their peers, the teacher gave very detailed feedback. By allocating less time to rote grading the teacher was able to give better feedback than he would have been otherwise. When it came time to assign grades at the end of the class, the students would volunteer a grade that they thought they had earned through their work. If it was higher than the teacher thought was appropriate the teacher would follow up to discuss it with the student. Rather than inflating their grade, the students more often gave themselves lower grades than the teacher viewed as appropriate. “At the end of the semester, fewer than 10% of the 50 students suggested final grades that differed from my own expectations” (Guberman, 2021, p. 91).

At the conclusion of the course, 20 out of the 50 students provided optional feedback through a survey and written response. The students gave a response to each of the questions according to a seven-point scale, ranging from strongly disagree to strongly agree. The questions pertained to three focus areas: learning climate, motivation, and informed learning.

The students' responses were positive. They felt that the learning environment was made better by the class' grading scheme. They felt that the instructor was helpful and encouraging. The feedback was praised as being honest, in-depth, and spanning the whole class, instead of just the narrow assignment. The survey suggested that the students were more intrinsically motivated

which was made more possible by the grading system that allowed them to take control of their own learning by allowing them freedom to craft their own learning plans. Students were able to choose what topics they delved into, how they studied them, and then how they conveyed their knowledge of them. These responses were partly due to the class subject being “fun.” But even a course with a fun topic such as Rock and Roll could become unfun by being too strictly controlled. A student stated that it gave them “creative freedom” that they hadn't experienced in any other college class. In regard to content, the students gave positive responses on all of the measures of growth. Many students wrote about how they felt that their ability to use different sources of information and to learn continued to grow throughout the course.

Meinking and Hall (2022) drew upon research that suggests that grades do not enhance student motivation and create large amounts of anxiety in students. To test this, the study enlisted “23 high achieving second-year students at a mid-size liberal arts university in the Southeast region of the United States” (p. 3). All of the students took a course titled “Beauty and the Brain.” This class was interdisciplinary, including both elements of neuroscience and classical studies. Per the university requirement, the class assigned final grades for each student. Departing from other classes, this course would not give intermediate grades during the semester, but instead would provide only detailed feedback. The study reviewed both the students' work which was coded to sort the work into themes. In addition, the study’s authors met with each student individually to confirm predominant themes that were identified through the course work. Specific focus was given to feedback impacting motivation or stress.

During the course, student feedback was mixed. Many students struggled with the new model. They especially expressed difficulty in staying on track and consistently working on the course. Also, anxieties about grades did not go away and in some ways they were made worse.

Without being assigned a grade, students did not have validation about what would constitute an "A" performance. The different structure put students out of their comfort zones. One student was quoted,

> Because there is so little emphasis on grades and specific assignments and deadlines, I'm really finding myself floundering when it comes to small assignments that we have to turn in. I'm usually an incredibly organized, put-together person, but for some reason, I literally keep completely forgetting about assignments we're supposed to do in this class. (Meinking, p. 8)

This was reflected in another student's response, who found the grading scheme both challenging and rewarding. He stated that he never considered the point of assignments past just doing what was needed in order to receive a high score and was having difficulty adjusting.

The mid-course feedback was more positive. Some students reported feeling that the course gave them an opportunity to learn in a different way and it allowed them to take risks in approaching assignments in new ways. It gave them the ability to look into things that they were interested in without fear that it was outside the scope of the assignment.

The feedback at the end was overall positive. Students felt that they not only learned about the class but also learned about themselves in the process. They felt that they grew as students as a result. They most liked the autonomy that the system allowed them. Many students reported an increase in intrinsic motivation. One student especially stated that "I realized over the course of this semester that I do actually love learning and I love academia that I might actually end up going that path" (Meiking & Hall, 2022, p. 12).

Some students reported a decrease in extrinsic motivation, especially extrinsic motivation caused by class competition. In other courses, students would ask each other what they got and

try to outdo each other. This was absent in this course. Overall, extrinsic motivation still remained, because despite no intermediate grades, many students reported that they had difficulty not worrying about the final grade at the end. They felt the desire to compete for grades too ingrained to let go of.

College Readiness

Guskey & Buckmiller (2020) conducted a study to respond to criticism that standards-based grading does a poor job at preparing students for college. To understand whether this criticism had merit, the researchers distributed a survey to students who came from a standards-based grading high school and were now enrolled in a small, private, Midwest college. Thirteen students responded to the survey, comprising the study's sample. After returning the survey, the students participated in a semi-structured, 45-minute survey in which they discussed any problems they may be having with college and how standards-based grading impacted them.

Overall, there was no detrimental impact from the grading system on the students' performance or problems; that is, most students did not struggle with a change to traditional letter-based grades. They instead struggled with ordinary college student issues such as homesickness and being in charge of their own schedule. The students spoke well of their high-school preparation, for it had given them a solid academic base. One student stated that he was struggling because he had become reliant on re-taking exams on which he had scored poorly; "Having the ability to retake major assessments and not having homework (count toward the academic grade) as a requirement (was a potential downfall of SBL in high school)" (Guskey & Buckmiller, 2020, p. #). This student was an exemplary student receiving a 4.0. The researchers concluded, "given that the typical GPA of first year students at four-year colleges in the United States is approximately 2.70, the participants in our study were highly successful in

their first semester of academic study at this private Midwestern university" (Guskey & Buckmiller, 2020, para. 20).

In order to compare assessment methods and their effects on both grades and ACT scores, Townsley & Varga (2018) identified and drew data from two comparable Midwestern high schools. The high schools were similar in enrollment, socioeconomic status, and ethnicity. One school employed standards-based grading practices, while the other used traditional grading practices. The study's total population was 327 students spread between the two high schools. The students involved were asked to provide their grade level, gender, cumulative GPA, grade of reach math course completed in each reporting period, grade for each English course completed in each reporting period, ACT composite score, ACT math subtest score, and ACT English subtest score. The study analyzed the data and compared grades and ACT scores to see whether one of the grading practices had an advantage for students taking the ACT.

First, despite the schools' different grading policies, each school gave grades in similar proportion to each other. This suggests that neither grading system led to inflated grades which facilitated school comparison.

In regard to test scores, the school that used traditional grades received higher ACT scores in both math and English. Students from the traditional-grading school scored a composite score 2.019 ACT points higher on average than the standards-based school. As for English, the traditional-grading school scored 2.654 higher than the standards-based grading school and for math, students from the traditional-grading school scored 3.277 points higher than the students from the standards-based grading school.

In 2010, Kentucky enacted Project Proficiency (PP) to address low reading and math scores. PP instituted standards and then held schools accountable by assessing students'

proficiencies. As a result, the teachers had common standards to teach to and subsequently taught to the same district-designed assessments. Additionally, the grading was on the basis of proficiency of the standards.

This background served as a good basis to test the effects of standards-based grading. In 2015, Pollio investigated whether standards-based or traditional grading had a stronger correlation with standardized test scores. In other words, did As and Bs correspond to proficiency scores in a standards assessment under either grading system? The study focused on minorities and disadvantaged students. As such, the study identified eleven high schools in the Louisville, Kentucky area that had implemented PP in the 2010 – 2011 school year. The demographic population of the school district was 51% White, 37% Black, and 12% other races. Nearly 62% of the district's population received free or reduced lunch. From this district, Pollio (2015) established two cohorts of students of eleventh graders who took an Algebra 2 course and assessment. The first cohort consisted of students who received Algebra 2 instruction in 2010 or 2011 according to the PP's standards-based instruction. The second cohort consisted of students from the same high schools who received Algebra 2 instruction in 2010 but not according to PP's standards-based instruction.

The study found of the students who did not take PP Math, 40% had received As or Bs. Of those 40%, only 26% of them scored "proficient" on the assessment. There was a larger correlation in grades and proficiency among the PP Math cohort. Of those students, 45% received an A or B. Of those students 55% were also proficient. According to Pollio (2015), it appeared that the standards-based reforms were successful at creating a stronger correlation between the students' grades and proficiencies in the standards and that more work needed to be done. Pollio (2015) commented, "educators should be concerned that 45% of the students in the

PP Math Cohort who achieved an A or a B in their Algebra 2 class still did not meet KCCT mathematics assessment proficiency" (p. 16).

Pollio (2015) did not formally ask for feedback, but a participant in the study collected data, which was included as "anecdotal observations." These were included to add "descriptive nuance" (p. 20). The participant's collected data refuted a possible criticism that teachers were simply teaching to the test. Teachers, parents, and students all had positive reactions to the changes. They felt that the change gave the lessons purpose. Pollio (2015) stated "Moreover, parent, teacher, and student discussions about grades appeared more meaningful and thoughtful. Instead of debating the number of points a student should have been awarded, more and more of these conversations focused on how the student demonstrated proficiency in a specific standard (p. 20)."

How Grading Practices Promote Fairness

Paff (2015) endeavored to fill a gap in the research about the impact of grading participation. The study sought to investigate the following three questions: 1) Does grading participation encourage participation? 2) How does it affect the quality of the classroom discussion? 3) Do students simply participate to receive credit and therefore participate in a more straightforward manner?

Students at a large public university in Pennsylvania were given a survey twice, once at the beginning and once at the end of one of two courses. The study's population included three sections of 50-student economic courses and a 60-student accounting course. The majority of the students enrolled in the courses were in their first or second year at University. The survey rated each according to a five-point Likert scale, ranging from strongly agree to strongly disagree.

The results of the pre-survey surprised the assumptions of the researchers. Only 43.2% of

students indicated that they participated more when participation was graded. Thirty-three percent of students agreed that grading participation disadvantaged some students, while 25% disagreed with that statement. Students reported hesitance to participate because of shyness (45%) and fear of being wrong (28%). When asked what the students recommended, they were split. Of these 43% of the students, 25 % felt that participation should be ungraded; 15% felt that participation should be mandatory, and 9% felt that participation should be awarded for attendance.

The post survey showed that students were unchanged in their attitudes on whether the policy encouraged them to participate more. At the end 42.4% said that they actually participated more because of the policy, which was just down from the 43.2% at the beginning of the courses. There was a large jump in the number of students who felt that the participation policy was fair. 80% of students reported that the grading in the course was fair to all students. At the beginning of the course, only 25% reported that grading would not unfairly disadvantage some students. The results of the post survey revealed the differing opinions on how participation should be handled. 33.7% preferred graded participation, 9.5% preferred ungraded, 11.8% preferred extra credit be given for participation, 13.6% preferred that only attendance should be graded, 25.9% preferred something else, and 6.5% of responses were left blank.

Gordon & Fay (2010) identified four practices that are common in education which are intended, in part, to promote grading fairness. These are curving grades, retaking examinations, discarding the lowest grade, and grading on the basis of improvement.

This study consisted of two samples of junior-level undergraduate students. The first consisted of all 193 students enrolled in Principles of Management, a core course in the business school. The second sample consisted of 463 students of the 1,283 students enrolled in multiple

sections of organic chemistry. Of the 193 of the first sample, there were 187 usable surveys. Of the 473 of the second sample, there were 426 usable surveys. The surveys asked students 14 questions on grading fairness. These responses were scored via a Likert scale, from Strongly Agree to Strongly Disagree. Next, they asked students to estimate how many of their courses used certain teaching and grading practices. Last, the survey also asked about background information from the students, including their current GPA and the average amount of time they spent studying.

Each question on the survey was correlated to grading fairness, interactional justice, distributive justice, and procedural justice. Interactional justice referred to teachers interpersonal styles, social sensitivity of the teacher, and how willingly teachers shared information. Distributive justice referred to whether the grades accurately represented the student's performance; procedural justice referred to consistency and fairness of the grading procedures.

The results were analyzed to examine which practices were associated with a related type of justice. Providing study guides, review sessions, and practice tests correlated with students' view of each of the three related to fairness. Grading practices such as dropping the lowest score, curving low grades, and allowing substitutes for low grades correlated only with interactional justice. Grading practices that modified grades were very slightly inversely related to distributive justice.

The results of this study suggest that students' perceptions of fairness are better served by teachers who provide students with tools to succeed as opposed to teachers who assess certain students by different criteria or by softening the effects of poor performances.

Arensmeier (2021) explored how Swedish grading reforms affected students with a specific focus on the lowest performing pupils in the Swedish system. The study's author

employed a discourse analysis, qualitatively assessing the policies publications and the school system's empirical data. It laid out a brief history of the Swedish reforms and criticized the reforms' lack of progress to their stated aims.

With roots in the 1930s, Swedish reforms in the 1960s and 1970s tried to encourage fairness and equality through the grading system. Traditional grading was replaced by norm-referenced grading, where students were graded against their peers. This was thought to help especially the gifted students from schools that have traditionally not sent many students into good universities. Under this grading system, students could not fail out of middle school, because even if they failed, students could still enroll in a secondary school.

Eventually this grading system was the target of grading reform when a center-right political majority came into power. This party hated that the system lacked accountability. The politicians replaced the system with a criterion-referenced grading in the 1990s. Under the new system, students could fail, but it was assumed that most students would receive passing grades. This did not occur. Many lower performing students failed to meet the minimum expected grades. In the 2010s, this system was reformed to encourage even more accountability into the system. Students needed to meet criteria to reach a passing grade. Students who received failing grades would not be eligible for a regular secondary school setting.

Upon reviewing the school system’s empirical data, Arensmeier (2021) stated that the reforms failed to help struggling students. Students still received the lowest grades at the same proportions as before. There has been no evidence that the possibility for failing and losing eligibility for a regular high school setting affects student performance. Students are not motivated by trying to avoid failing. The author cited Lindblad et al. (2018) which showed the problems facing the lowest achieving students is not accountability but rather cognitive abilities

and lack of funding.

Traditional Challenges with Grading

Kunnath (2017) stated that grades are important but historically have done a poor job at communicating student performance. The study focused on three questions: 1) What influences affect teacher grading practices? 2) What grading rationale do teachers use to produce student report card grades? 3) What grading practices do teachers use to create student report grades?

Kunath distributed a survey to teachers that consisted of two parts: a quantitative and qualitative section. The quantitative section provided an overview of grading influences and teacher practices, and the qualitative part asked teachers' detailed understanding of their grading rationales. All teachers sampled were from a large urban school district in California. The school district had over 37,000 students and was 63% Latino. Sixty-four percent of the school districts' population qualified for free or reduced lunch. The school district employed 1,500 teachers with an average of fifteen years teaching experience. A total of 251 teachers returned the quantitative survey. Of those that returned the survey, 86% had taught for more than seven years. There was a relatively even response from all of the core subject areas. For the qualitative part, fifteen teachers took part in small group semi-structured interviews. The quantitative survey was based on a five-point Likert scale (i.e., 1 = not at all, 5 = completely). The qualitative small group interviews focused on "grading challenges; grading procedures; and the influence of their own education philosophy and beliefs, external factors, and classroom realities on their grading" (Kunnath, 2017, p. 73).

As for grading influences, the top five revealed by the quantitative study (in order of importance) were a desire for student success, philosophy of teaching and learning, desire to promote student understanding, desire to accommodate student individual differences, and needs

and student motivation and engagement. Pressures by parents and administrators, standardized testing, and managing behavioral issues all fell much below these factors. The specific values are provided in Table 3 below.

Table 3

Survey Results of Teacher Use of 17 Grading Practices

	Percentages					
Grading Practice	Not at all	Slightly	Somewhat	Largely	Completely	M
Student Ability Level	4	7	25	57	8	3.58
Student Academic Achievement	1	3	14	63	19	3.97
Student Disruptive Behavior / Conduct	75	19	4	2	0	1.33
Student Effort	6	25	31	32	6	3.06
Student Participation and/or Paying Attention	23	31	27	17	2	2.46
Student Improvement of Performance	13	25	41	20	1	2.71
Grade Distributions of Other Teachers	91	5	2	2	1	1.16
Student Performance of Other Students in Class	66	22	7	4	1	1.51
Student Performance Compared to Students from Previous Year	83	9	5	3	1	1.28
Specific Learning Objectives Mastered by Students	2	4	17	62	15	3.85
Formal or Informal School or District Policy	77	7	8	5	4	1.51

Student Effort, Improvement, Behavior	18	37	32	12	2	2.44
Student Completion of Homework	37	36	23	4	1	1.95
Quality of Student Completed Homework	18	25	35	20	2	2.64
Inclusion of Zeros from Incomplete Assignments or Assessments	11	16	36	18	18	3.17
Student Extra Credit of Academic Performance	46	47	6	1	0	1.61
Student Extra Credit for Non-academic Performance	88	10	1	1	0	1.15

As for the qualitative survey section on grading practices, teachers rated four categories as being especially important to their grading rationale: student academic achievement, specific learning targets mastered, student ability level, and student effort. Of note, only the first two of these four are recommended by grading experts. Conversely, rewarding student effort was one of the things that teachers spoke about in their interviews. “An English teacher explained her inclusion of effort in a grade to reward student work completion: ‘Everybody’s working, nobody just slacks off and does nothing . . . I personally want to make sure I honor that and respect that as part of my grade’” (Kunnath, 2017, p. 80).

In order to better understand the effect of rubrics, Marzano (2020) ran two experiments, which compared grading variance between teachers of the same six-item short answer responses. Specifically, both experiments used the same 10 student responses of real eighth graders writing about precipitation. The 10 responses represented different levels of knowledge, but all were perfect in regard to spelling and grammar. Students were identified as Student A, B and so on to control for bias regarding student names. Both experiments gave these ten responses to 10

different raters. All raters were experienced eighth-grade science teachers, who were highly knowledgeable about the subject. These raters averaged over 10 years of teaching experience.

In the first experiment, the 10 raters were randomly assigned to two different methods of scoring the students' constructed responses. Five raters were tasked with grading the assignment with an "unconstrained" grading method. These teachers first decided how much weight to give each part of the answer, and they then scored it according to how far the answer fell short of their standards. This allowed for the most flexibility in grading. The other five teachers were given a general rubric. For example, a perfect four-point response would be given when "the student has a complete and detailed understanding of the information important to the topic" (Marzano 2020, p. 252).

After data was collected from the first experiment, a second experiment was run. The same responses were given to a new set of experienced science teachers. The teachers were again randomly split and assigned different grading practices. This time one-half of the teachers used a constrained grading approach; this approach was similar to the unconstrained grading method in experiment 1, except for two changes. One, the six parts had a predetermined point value assigned; and two, the teachers had discussed together how they would grade the assignment using an assignment submission that was not part of the study. The second group of graders used a rubric that was topic and grade-level specific.

It is important to note that the study's population was quite low. Only 10 responses were used and each method was employed by only five graders. This means that the results may not be as generalizable as they otherwise would be. Deviations were extracted from the small sample size and stated as a population. This led to considerable sampling errors, which were noted by the study. That being said, there was a clear difference in the deviation of scoring between the

different graders. Variability for unconstrained point condition was 69.4; for the generic rubric condition, it was 48.21; for constrained point condition, it was 36.9; and for topic-specific rubric condition, it was 30.16. Marzano (2020) stated,

> Taken at face value, the results for the four conditions across the two experiments indicate that the topic-specific rubric condition produces more generalizable and dependable scores than does the constrained point condition that produces more generalizable and dependable scores than does the generic rubric condition that produces more generalizable and dependable scores than does the unconstrained point condition. (p. 264)

This study suggested that the same assignment would be scored in a more predictable way when scored by different teachers using a topic-specific rubric. The constrained-point model, which allowed teachers to meet with each other prior to grading, also showed high levels of consistency.

Hobson (2003) explained why poor directions and rubrics can lead to inconsistent grades and how to write better directions and rubrics. To preface this, the researcher noted that teachers should be aware of the challenges students face in completing assignments. Before beginning the assignment, students have to understand the instructions, identify an effective process to tackle the assignment, and ascertain how the assignment is to be graded. Teachers sometimes take this part of the assignment for granted. It is not always easy for the teacher to put themselves in the shoes of the student. Unclear directions can make these three parts difficult. As a result, students will often be forced to guess as to what the teacher is looking for.

One common mistake that makes assessing writing less consistent is poor or incomplete written instructions and rubrics. Hobson (2003) used an example of a common writing prompt

that a teacher used; “Write a 4–5 page review of a frequently used over-the-counter (OTC) or health-care product. You will be graded on accuracy of information and adherence to proper spelling and grammar usage (p. 53).”

The article identified areas in which this assignment was vague. In particular, it did not state the audience, appropriate format, research process, evaluation areas to be emphasized, expected use of sources, and efficient process for completing the project. The suggested rewrite of the assignment was almost a full page and even this could have been improved upon further. In addition to well-written thorough instructions, the teacher should provide a criteria tool for the assignment. In the example stated, this included a checklist whereby the grading criteria was broken down into smaller parts. The project was graded on focus, audience, and professional expectations. Focus was broken down into the likes of “the report reviews a single, specific OTC or health-care product, thesis is clear and appropriately supported, author provides detailed, documented information about benefits and detriments associated with product” (Hobson, 2003, p. 55). Lastly, a comprehensive rubric should be provided and adhered to. Similarly, a strong rubric should be thorough. For example, the study wrote a B grade equated to:

> (B) Strong
>
> A strong project presents a well-developed review of a specific product, focusing on important features of the product and its use, discussing them in a generally thoughtful way. Its intended readers have little reason to doubt the accuracy of the information presented or the assessments made, although they may desire more detailed information and more thorough analysis. The document demonstrates good control of the elements of effective writing. (Hobson, 2003, p. 56)

Brimi (2011) looked at the reliability of grading high school essays. In referencing a

century old study on the same topic, he noted that past results have shown quite a range of grades between different teachers and he wanted to know if the variance in the range of grades had narrowed. “In the 2007-2008 academic year, 90 ninth and tenth-grade teachers representing 12 schools were trained to use 6+1 as a teaching and assessment tool” (Brimi, 2011, p. 5). This was a new and mandatory assessment tool for the teachers. The assessment considered the following traits: “’Ideas,’ ’Organization,’ ‘Voice Word Choice,’ ‘Sentence Fluency,’ ‘Conventions,’ and ‘Presentation'”. In 2008, after training, teachers attended a follow-up session in which each teacher was asked to grade a single essay using the 6+1 approach. The essay, an argumentative piece entitled “Why Abortion Should Be Illegal” was chosen by the author of the study; Brimi (2011) picked it specifically because it contained many strong points as well as several flaws. The researcher acknowledged that the controversial nature of the topic may have affected the results, but as a general matter, the author’s viewpoint should not be a part of the grade.

Of the 90 participants, only 73 could be used for the study. One teacher did not return the paper, and 16 did not properly adhere to the numerical grade based on a 100-point scale. The teachers’ scores ranged from a high of 96 to a low of 50; the mean score was 81.1599. When the numerical scale was converted to letter grades, they fell as follows: 10 As, 18 Bs, 30 Cs, 9 Ds, and 6 Fs. Despite the broad range of scores, this result was consistent with past similar attempts. Elliot (1912) had similar ranges of 37 and 44. Brimi (2011) drew on past research to offer possible reasons to explain the range of scores. For example, the teachers may either have been ignorant of current grading practices or unwilling to change their views on a writing assignment; this was evidenced in part by the 19 teachers who failed to follow the 100-point score requirement. Hillock (2006) also believed that many teachers lacked preparation to teach composition past the basic requirements of state assessments. For example, teachers marked

errors for parts of the paper that were not wrong such as a delayed thesis in an argumentative essay and the use of "you" in an opening hypothetical situation . Another observation is that teachers' marks, grades, and comments were almost all concerned with the essay's errors. Very few teachers praised parts of the paper that were done especially well.

Brimi (2011) called for finding a balance between standardization and flexibility. Teachers and students can both find writing to rubrics as less fun. The student feels that they are simply following directions, and are less free to answer in a way that feels right to them. Similarly, the teacher feels that they are assessing writing by checking boxes and are not able to properly reward all the ways that make a piece of writing good. Assessing writing involves subjectivity. Good writing cannot be distilled into a simple checklist.

In a seminal work on this topic, Starch & Elliott (1912) investigated the consistency of grading assessments in English. Two papers were collected from actual ninth-grade students in Wisconsin. These papers were copied exactly as they appeared on the two exams. The handwriting, errors, changes, and all were reproduced. The assessments had six parts, all open-ended by design, which were a mix of grammatical rules, technical writing, and literature analysis. The final section requested a written three-paragraph essay, whether narrative, descriptive, or both. The first paper chosen was the better of the two. The topic for its three-paragraph essay was "Irish Hall of Fame" (Starch & Elliott, 1912, pp. 446-447). It described the building's purpose as a memorial as well as its architecture. The second paper's three-paragraph essay focused on the students' dog's relationship with the neighbor's dog.

The questions and the two responses were sent to 200 high schools in the North Central Association requesting they be scored by the top ninth-grade English teacher. The answers were returned and split into two categories based on whether 70 or 75 was the passing score.

The first observation was the “tremendously wide range of variation.” Starch and Elliott (1912) stated, “It is almost shocking to a mind of more than ordinary exactness to find that the range of marks given by different teachers to the same paper may be as large as 35 or 40 points” (p. 454). This wide range has a few ramifications. For the poorer-performing student who wrote paper B, the decision as to whether he received a passing grade depended on the teacher grading his paper. Also, the differences were not uniform. For example, a school that gave paper A a poor grade did not necessarily give a similarly poor grade to paper B. In general, “paper B is marked on average 8 points lower than paper A, yet nineteen of the one hundred and forty-two teachers marked it higher than paper A, and twenty-three marked it 15 or more points lower than paper A” (Starch & Elliott, 1912, p. 455). There were some consistent trends between schools. Schools with less than 150 pupils graded students more leniently. For paper A, this resulted in a median of 89.5 for small schools and a median of 86.8. Despite this leniency, the range of grades remained consistent.

CHAPTER III: DISCUSSION AND CONCLUSION

In writing this paper, I reviewed many articles addressing many different questions around grading. These employed different research methods, but most commonly these articles are qualitative, relying on questionnaire responses and semi-structured interviews. This gave me a better understanding of the experiences with different grading protocols, but they were not as definite or profound as I had hoped.

In regard to teachers' and students' experiences with standards-based grading, the results were somewhat positive. Both teachers and students remarked that they benefited from clearly defined purposes and goals. Both teachers and students found experiencing grading from a different angle allowed them to understand their own learning more deeply. There were also bad experiences with standards-based grading changes; many of these stemmed from inconsistent application and the system just being new to them.

In experiments with other alternative grading systems that did not strictly adhere to standards-based grading principles, the results were also positive. Both ungraded and student-led grading showed an increase in student intrinsic motivation. Students felt that they could take control of their own learning by implementing their own plans and strategies.

The studies are mixed in regard to the effect of standards-based education on college readiness. They suggest that students are not harmed by standards-based grading in that students still are able to succeed in college. There is some evidence that students who are graded in a standards-based setting do not perform as well on entrance exams. So it appears that they may be disadvantaged during the entrance process, but succeed once they arrive in college classes. It also just shows that more research needs to be done to determine the effect of standards-based grading on college bound students.

The studies about fairness show that there is justification for standards-based grading reforms. According to survey responses in Paff (2015) participation does not promote students' perceptions of fairness, while retaking exams and flexible due dates did in Gordon (2010). Similarly, struggling students are helped under a grading system that does not pile unnecessary pressures on them. Some pressures are part of the learning process, but under standards-based grading there is less pressure regarding deadlines and final results.

The studies regarding challenges in grading help to demonstrate a need for a grading system that is less discretionary. It also showed the need for clear goals, clear directions, and clearly stated grades, which is a central feature of standards-based grading. Similarly, discretion and consistency is a reason for the four-point grading system is intended to make grading.

Personal Application

Given the goals of standards-based grading and the strong theoretical foundation of it, I believe that school districts should continue to move toward it. Schools should also have a singular grading philosophy that all teachers follow. Teachers must follow the administration's decisions on what grading philosophy to use. As for me, I will follow the school guidance on grading, whether that be traditional or standards-based grading. Within the recommendations of the school district, I will apply some of the prescriptions of standards-based grading. I will strive for clarity of purpose and clarity in grading for competence.

My review has changed the way I view grading. Some of the ways are stated well in Guskey (2011), which advocated for a rethinking of grading. First, grading isn't simply about ranking students. Instead, grading should be part of facilitating growth in students, and this is why we grade practice lightly and allow them to retest. Second, grade distributions should not simply follow a bell curve where students are clustered in the middle. Instead we should expect

all students to clear certain objectives. Third, grading should not simply be a comparison to other classmates, but it instead should be in relation to what the student can do.

One of my key takeaways from reading the observations of those employing standards-based grading is that just thinking about grading from different perspectives is beneficial. Standards-based grading has given new guidance in how to think about grading. So, even if I don't abide by all of the precepts of it. I still hope to think thoughts such as 1) “What is my purpose in grading this way?” and 2) “How does my grading this assignment further academic goals? In other words, I need to be more intentional about grading.

I will apply Munoz & Guskey’s (2015) advice on communicating grades to students. I will try to clearly articulate the purpose and criteria of grading. I will separate product, process, and progress criteria. Product criteria presents what students know and are able to do at a given time. Product criteria is the most important grade. The main purpose of grading is to assess students’ product score. When it is necessary to convert standards-based grades to letter grades, product criteria is the only factor that should be considered. It is the criteria that should determine rank, admissions to college, and such things. The other two criteria are useful for the student and parents, but should not muddle the product score. Process criteria shows things like effort, work habits, participation, attendance, punctuality of assignments and so on. Progress criteria displays the students progress over time. These three criteria are all helpful to students and parents. Separating these three criteria helps to give meaning to grades. Since they are not combined, teachers don't have to consider how much weight each part should receive.

Limitations

I did not find as many sources as I had hoped. This may be for a few reasons including my research method and my expectations. Another reason that I didn’t find the articles or clear

outcomes that I expected to find is that testing standards-based grading is not easily done. It is hard to find an objective way to test this. Most of the studies that I found were testing different students, because a researcher cannot simply put the student through the same course twice. This leads to some issues, because comparing students from one year is not exactly the same. I believe this is even truer now after COVID, because students have had different academic experience with remote learning. It is also made more difficult to test because the aims of standards-based grading are not simply performance goals. Much of the goal is equity and fairness. These are difficult to address with a data driven test.

Implications for Future Research

More research should be done on the topic, because the effects of standards-based grading and other grading systems are still unclear. I would be interested in a study that looked at a classroom or school district that had been employing standards-based grading practices for a while. Most of the studies focused on classrooms or school districts that had recently switched. There is a good reason for this, because the new system can be directly compared with the old system. The problem with doing this though, is that much of the feedback concerns just how new the system is. Teachers complained about having to apply a new system and other growing pains (Knight & Cooper, 2019). Students similarly complained about having to learn a new system (Peters et al., 2017). On the other hand, much of the positive feedback is because it forces both teachers and students to look at a different way of grading. It would be interesting to see feedback from students who have been brought up with standards-based grading being exposed to traditional grading. Would they respond positively to it simply because it was new and interesting? This could be the case.

Equity

Some articles that I found gave compelling arguments that there has been insufficient focus on equity concerns. These are more difficult to test but they are a very important part of standards-based grading reforms.

Feldman (2019) argued that standards-based grading is a step in the right direction but more needs to be done. He stated that participation and effort, although innocuous sounding, inject unfair bias into grading. When this is done teachers often unfairly punish culturally accepted behaviors of other cultures. Similarly, homework again can be discriminatory, because disadvantaged students do not have tutors or quiet home environments of wealthier students. So, Feldman (2019) approves of standards-based gradings reducing the importance of homework.

Similarly Wehmeyer et al. (2004) argued that standards-based grading was not adequately tailored to address the needs of students with special needs. The authors believed that more needed to be done to improve students' self-determination. The standards-based grading system can be too rigid for these students in its singular focus on grading for understanding and performance. The researchers argue that teachers, administrators, and parents should understand the important role of grading in fostering skills that will aid students in their academic careers. They argue self-determination, which includes taking ownership of their learning and believing in themselves, is very important for all students, but especially those with special needs.

Both of these articles made compelling arguments and I am inclined to agree with them. It would be nice if there could be some objective data driven studies to properly test how equity is handled in a standards-based grading system. Again this may not be easily done, but it seems that there could be some way to accomplish it.

Conclusion

Overall, this paper has strengthened my resolve that grading is important to education. Its importance is more than just sorting students. Grading practices affect students' motivations, feelings of fairness, and behaviors. A student who receives one grade they feel is deeply unfair is enough to affect their whole attitude towards a class or to learning more broadly. Similarly, grading can be structured to motivate a student to reach new levels of understanding. On the other hand it can also be messy and incomprehensible, and leave the student bewildered.

Given its importance, we should continue to tinker with grading. A perfect grading system likely does not exist, but a better system does. Therefore, teachers and school districts need to experiment, collect data, and test ideas.

References

Arensmeier, C. (2022). Institutionalizing school failure: From abandoning to reintroducing a failing grade--the rationales behind Swedish grading reforms. *Journal of Educational Change, 23*(2), 221–252. https://doi-org.ezproxy.bethel.edu/10.1007/s10833-021-09421-7

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

Feldman, J. (2019). Beyond standards-based grading: Why equity must be part of grading reform. *Phi Delta Kappan*, *100*(8), 52–55. https://doi-org.ezproxy.bethel.edu/10.1177/0031721719846890

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

Guberman, D. (2021). Student perceptions of an online ungraded course. *Teaching & Learning Inquiry, 9*(1), 86–98.

Guskey, T. R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16–21.

Guskey, T. R., & Buckmiller, T. M. (2020). The impact of standards-based learning: Tracking high school students' transition to the university. *NASSP Bulletin, 104*(4), 257-269. https://doi-org./10.1177/0192636520975862

Guskey, T. R., & Jung, L. A. (2009). Grading and reporting in a standards-based environment: Implications for students with special needs. *Theory Into Practice*, *48*(1), 53–62. https://doi-org.ezproxy.bethel.edu/10.1080/00405840802577619

Hobson, E. H. (1998). Designing and grading written assignments. *New Directions for Teaching and Learning,* 74, 51–57.

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

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

Lewis, D. (2022). Impacts of standards-based grading on students' mindset and test anxiety. *Journal of the Scholarship of Teaching and Learning, 22*(2), 67–77.

Lindblad, I., Westerlund, J., Gillberg, C., & Fernell, E. (2018). Har alla barn i grundskolan förutsättningar att klara nya läroplanens krav? [Do all children in compulsory school have the ability to meet the requirements of the new curriculum?]. *Läkartidningen, 115*(6), 1–4.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*(3), 249–267.

Marzano, R. J., & Heflebower, T. (2011). Grades that show what students know. *Educational Leadership, 69*(3), 34–39.

Meinking, K., & Hall, E. E. (2022). Letting go of grades: Creating an environment of autonomy and a focus on learning for high achieving students. *Teaching & Learning Inquiry, 10*. https://doi.org/10.20343/teachlearninqu.10.21

Meyer, L. H., McClure, J., Walkey, F., Weir, K. F., & McKenzie, L. (2009). Secondary student motivation orientations and standards-based achievement outcomes. *British Journal of Educational Psychology, 79*(2), 273–293.

Muho, A., & Taraj, G. (2022). Impact of formative assessment practices on student motivation

for learning the English language. *International Journal of Education and Practice, 10*(1), 25–41.

Muñoz, M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. *Phi Delta Kappan, 96*(7), 64–68. https://doi-org.ezproxy.bethel.edu/10.1177/0031721715579043

O'Connor, K., & Wormeli, R. (2011). Reporting student learning. *Educational Leadership, 69*(3), 40–44.

Paff, L. A. (2015). Does grading encourage participation? Evidence & implications. *College Teaching*, *63*(4), 135–145.

Percell, J. C. (2019). Democracy in grading: Practicing what we preach. *Critical Questions in Education, 10*(3), 180–190.

Peters, R., Kruse, J., Buckmiller, T., & Townsley, M. (2017). "It's just not fair!" Making sense of secondary students' resistance to a standards-based grading. *American Secondary Education, 45*(3), 9–28.

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

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31.

Scarlett, M. H. (2018). "Why did I get a c?": Communicating student performance using standards-based grading. *InSight: A Journal of Scholarly Teaching, 13*, 59–75.

Selby, D., & Murphy, S. (1992). Graded or degraded: Perceptions of letter-grading for mainstreamed learning disabled students. *British Columbia Journal of Special Education,*

16(1), 92-104.

Starch, D., & Elliott, E. C. (1912). Reliability of the grading of high-school work in English. *The School Review, 20*(7), 442-457.

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

von der Embse, N., Jester, D., Roy, D., & Post, J. 2018. "Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review." *Journal of Affective Disorders 227: 483-493*. https://doi.org/10.1016/j.jad.2017.11.048

Wehmeyer, M. L., Field, S., Doren, B., Jones, B., & Mason, C. (2004). Self-determination and student involvement in standards-based reform. *Council for Exceptional Children, 70*(4). 413-425. https://doi.org/10.1177/001440290407000402

Williams, M. J. (2023). Teacher perceptions of differentiated instruction in a standards-based grading middle school. *Educational Research: Theory and Practice, 34*(1), 129–150.

Wilcox, J., & Townsley, M. (2022). Debunking myths of standards-based grading: Addressing the concerns and providing some strategies for implementing alternative grading practices. *Science Teacher, 90*(1), 29–33.