COGNITIVE BIAS IN THE RUBRIC EVALUATION OF STUDENT PERFORMANCE IN STANDARDS-BASED GRADING MODELS

BY

ANTHONY R. REIBEL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Education in Educational Policy, Organization and Leadership
with a concentration in Educational Administration and Leadership
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Assistant Professor Jennifer Nelson, Chair and Director of Research
Associate Professor Rachel Roegman, Chair
Professor Yoon Pak
Teaching Associate Professor Mary Beth Herrmann

Abstract

Over the past decades, alternative grading models that rely on teachers' professional judgment of the quality of student work have been embraced as a more equitable method for grading since they claim to provide more specific, timely, and personalized feedback on students' knowledge and growth (Buckmiller et al., 2017; Knight & Cooper, 2019; Muñoz, & Guskey, 2015). These models include competency-based grading, portfolio-based grading, and standards-based grading, with standards-based grading being more widely practiced than the others (Buckmiller et al., 2020; Erickson, 2011; Iamarino, 2014). Proponents of standards-based grading argue that it is less subject to cognitive and racial biases of teachers than conventional grading systems (Feldman, 2019; Quinn, 2020). This study scrutinized this claim by investigating whether standards-based grading models resisted such teacher biases. Specifically, this study investigated racial discrimination, shifted standard bias, and expectancy bias using the quantitative approach of a factorial vignette experiment. Additionally, qualitative feedback to students about their performance was analyzed regarding teachers' attributions for student performance. The findings of this study suggest that using a standards-based rubric to grade and essay did not activate any racial bias; however, expectation bias, a shifted mastery standard, and attribution error were present.

Keywords: cognitive bias, standards-based grading, expectancy bias, attribution error, racial bias, heuristics, judgment, student evaluation, shifting standards

Acknowledgments

I want to acknowledge my advisors, Dr. Jennifer Nelson and Dr. Rachel Roegman, for their support and guidance. Thank you for strengthening my focus.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Some researchers and educators consider standards-based grading models to be a workable solution to the problem of inequitable grading practices (Feldman, 2018; Wormeli, 2018). Researcher David M. Quinn (2020) found in his study, “when teachers use a rubric that orients grading decisions to a limited number of specific, demonstrable criteria, they show no bias in their grading decisions. When teachers are asked to rate student work along a vaguer spectrum of performance, based on meeting “grade-level” standards, their grading favored the white student” (p. 72). However, grading models that use rubrics, which standards-based grading does, are not impervious to teachers' cognitive biases (Scarlett, 2018). As a result, there may be an increased risk of miscalculated grades, as students might "feel reluctant to invest themselves in a domain where they could be subjected to biased judgment or treatment" (Cohen & Steele, 2002, p. 304). Therefore, when considering a change to standards-based grading, it would be prudent to investigate the intersection of cognitive biases that can affect decisions, precisely racial prejudice, shifting standards, expectancy bias, and attribution error.

Problem

Standards-based grading models are often advertised as a more equitable evaluation of student work (Townsley & Wear, 2020; Marzano & Hardy, 2023; Muñoz & Guskey, 2015). However, these models are not impervious to non-academic factors such as teachers' values and cognitive biases (Guskey, 2014). Further, these factors not only have the potential to increase miscalculated grades, but "students will feel reluctant to invest themselves in a domain where they could be subjected to biased judgment or treatment” (Cohen & Steele, 2002, p. 304).

Further, the standards used for grading may not be agreed upon by the staff, district, state, or even the country in standards-based grading models. Therefore, a teacher's subjective values, beliefs, and assumptions can influence standard interpretation (Scarlett, 2018). Rubrics can often be applied differently (Kahneman et al., 2021); the interpretation of the term "excellent" on a standards-based grading rubric can vary from teacher to teacher (Gentrup et al., 2020). The potential issues with professional judgment models raise the following question: to what extent do teacher biases affect their decisions when evaluating student work? When studying bias and decision-making in a domain such as grading, a practical place to start is a broad study of human judgment.

Judgment and Decision-Making

Judgment

In *Noise*, Kahneman et al. (2021) define judgment as "a measurement in which the instrument is a human mind" (p. 39), and further, judgment is a "conclusion that can be summarized in a word or phrase" (p. 40). Moreover, sound judgments "depend on what you know, how well you think, and how you think" (p. 225). According to Kahneman and colleagues, judgment has two primary deficiencies: *bias* and *noise*. They define *noise* as "the unwanted variability in judgments that should ideally be identical [and] this noise can lead to rampant injustice… and errors of many kinds" (p. 21). They classify noise further into several types: level noise, pattern noise, and occasion noise. In an education context level noise occurs when different teachers grade student work harder or easier; pattern noise is when teachers might disagree on which criteria or components deserve more or less scrutiny when grading, and occasion noise might be when teachers disagree among themselves (teachers grade one test harder than a similar test on a different occasion). Through multiple experiments and consistent

findings, Kahneman et al. (2021) have shown strong evidence that “mood, fatigue, weather, and other sequence effects: many trigger unwanted variations” in our judgments (p. 91). In education, these noises create systemic distortions, where students might have completely different experiences due to variability in scoring, feedback, and instruction. Hereafter, sources of noise are referred to as level, pattern, occasion, and system variability.

Implications of Judgment Variability

Objectivity is of utmost importance in certain professions, such as education, medicine, or the law. However, as mentioned previously, judgment can be distorted by internal and external forces (Rom & Conway, 2018). For example, in the legal field, judges can be influenced by factors outside their consciousness, such as weather or the time of the day (Dobelli, 2013). Dror et al. (2021) call this "biasability," which “refers to the systematic impact that factors can have on professionals’ ability to think or reason about the information at hand objectively” (p. 1). Dror and Charlton’s (2006) forensic science study on fingerprinting found that criminal investigators can often draw different conclusions from the same fingerprint evidence presented to them on different occasions.

Heuristics

When making a judgment, what goes on in an individual's head is not deliberate processing but more of an automatic response to their environment (Bargh & Chartrand, 1999). Individuals make judgments and decisions often based on unconscious heuristics, which include their values, prejudices, and biases (Gilovich et al., 2002). Tversky and Kahneman (2002) described heuristics as a form of "'natural assessments' elicited by the task at hand that can influence judgment without being used deliberately or strategically" (p. 5). These assessments promote "judgment of likelihood in the absence of deliberative intent" (p. 5). Similarly, "implicit

associations have a stronger impact when people are unable or unwilling to devote cognitive resources to their behaviors and decisions, instead relying on gut reactions" (Warikoo et al., 2016, p. 510). If people do not have the time or capacity to reflect on their thoughts and impulses before they act on them, they often rely on heuristics and intuition, which can be vulnerable to internalized stereotypes or prejudices. Three common heuristics are “availability, representativeness, and anchoring and adjustment” (Gilovich et al., 2002, p.3). The availability heuristic is “when the individual assesses a specified target attribute of a judgment object by substituting a related heuristic attribute that comes more readily to mind” (Kahneman, 2002, p. 466). In other words, substitution with an example that easily comes to mind. For instance, if someone knows tall people who play basketball, they are likely to assume, using the availability heuristic, that a tall person they see somewhere naturally plays basketball. The representativeness heuristic tests how closely the subject of one's observation relates to an internalized prototype. For example, if someone assumes that all tall people play basketball, they might use this specific heuristic to generalize that every tall person they see plays basketball. The third type of heuristic is *anchor and adjust*. People start with an initial value and then judge relative to that anchor. For example, if someone says, "They are nice for a businessman," it implies that they used an anchor to show that they had a base-level assumption that businessmen are typically not nice.

Heuristics and Education

In education, teachers may rely on heuristics to make quick decisions because of the lack of time or their students’ continuous need for attention in the classroom (Gilovich et al., 2002). These quick decisions can influence their evaluations of student work (Zhan et al., 2021), as they must make cognitively demanding micro-decisions (heuristics) to support various student needs. Both internal factors (e.g., dispositions, values, expectations) and external factors (e.g.,

environment, events) can activate teacher biases. Factors such as morning or afternoon classes, hours of sleep, classroom temperature, or preceding learning events can impact a teacher's judgment (Kahneman et al., 2016).

Heuristics and Student Performance Evaluation

When a teacher evaluates their students, they may filter information through heuristics, a quick evaluative strategy that helps determine performance. However, any distorted filtering of contextual information may lead a teacher to shift their standard of quality or make unsubstantiated assumptions to judge a student's performance (expectation bias, shifting standards theory). Each of these theoretical frameworks has been discussed in more detail in Chapter 2, along with an exploration of how the interplay between these frameworks may help explain variability in the scoring of student work in a standards-based grading system.

Purpose

As schools contemplate shifting from conventional grading practices (points and averaging) to systems that utilize the professional judgment of evidence (standards-based, competency-based, portfolio-based), it is crucial first to examine how teachers judge student performance (Knight & Cooper, 2019). Many educators believe that standards-based grading models are just and equitable, owing to the absence of mechanical performance calculations such as averaging, weighting, and grading scales (Feldman, 2018; Townsley & Wear, 2020). This study investigates the validity of this common perception with the following research questions:

> To what extent are teachers influenced by their cognitive biases when using standards-based rubrics to evaluate student performance?

The specific sub-questions are as follows:

1) To examine potential racial bias, I ask: How does a student's implied race affect the grade teachers award on a standards-based rubric?

2) To examine potential shifted standards and expectation biases, I ask: How does a teacher's additional background information on the student, as presented in a student learning profile, affect the grade teachers award on a standards-based rubric?

3) To examine potential attribution error processes, I ask: How does a student's implied race or learning profile relate to the content conveyed in teacher feedback?

These questions align with the research investigating how relevant and non-relevant factors influence evaluators' performance assessment (Dror & Charlton, 2006; Kahneman et al., 2021).

This dissertation explores how teachers use relevant information, such as proficiency indicators, success criteria, and other academic variables, versus non-relevant information, such as student race and learning profile information, when judging student performance. Results can therefore detect whether several potential biases, such as racial bias, shifting standards, expectancy bias, and attribution error, may influence how teachers evaluate student performance. This dissertation's findings can inform school leaders' decisions regarding a change to standards-based grading since leaders may consider this change to be isolated to the grade book format (Heflebower, 2019; Marzano, 2011) without exploring how biases can distort judgment. This narrow perspective of a change to a standards-based grading model may lead to the implementation of such a system falling short of its promise of equity by continuing to communicate unreliable grades, distort feedback, perpetuate social inequities in the learning process or even distort a student's self-concept (Bandura, 1997; Marsh et al., 2018).

Definition of Key Terms

The following concepts are used in this dissertation's theoretical and analytical framework.

Racial Bias – a tendency to prefer one person's race to another and to favor that race (Maryfield, 2018)

Rubric-based grading – the use of "an evaluation scale that is preferentially used by teachers (and by students in self-assessment and peer-assessment tasks) to assess competence descriptors" (Velasco-Martinez & Tojar-Hurtado, 2018, p. 119)

Mastery – "a construct that cannot be observed directly but can be inferred from an observable performance on a set of items or tasks related to a particular concept, skill, or subject" (Guskey & Anderson, 2013, p. 21)

Heuristics – "'natural assessments' elicited by the task at hand that can influence judgment without being used deliberately or strategically" (Gilovich et al. 2002, pp. 4-5)

Attribution – "judgment of why a particular incident occurred" (Weiner, 1972, p. 203)

Attribution theory – "how people make causal explanations, about how people answer questions beginning with 'why?'" (Kelley, 1973, p. 107)

Attribution error – "the tendency to overestimate individuals' influence [on outcomes] and underestimate external, situational factors" (Dobelli, 2013, p.107).

Implicit bias – "the automatic cognitive associations or affective predispositions individuals have with different social groups." (Starck et al. 2020, p. 274)

Cognitive bias – "when people are subconsciously influenced by prior beliefs and expectations formed on the basis of contextual information that is irrelevant to [what] they are evaluating" (Kukucka, et al., 2017, pp. 453-454)

Judgment – "measurement in which the instrument is the human mind" (Kahneman et al., 2021, p. 39)

Organization of the Manuscript

This dissertation is organized to communicate how cognitive bias influences teachers' decisions as they evaluate student performance. Chapter 2 reviews the existing literature and examines relevant studies concerning the potential bias in a teacher's evaluation of their student's performance. Chapter 3 outlines the methodology used to collect and examine the potential impact of bias on standards-based rubric grading among a random sample of teachers. Chapter 4 presents the results, and Chapter 5 presents conclusions drawn from the study's findings and considerations for future research.

CHAPTER 2

REVIEW OF LITERATURE

A standards-based grading system is thought to be more equitable for students since teachers who use this system often interpret learning evidence in place of awarding points and determine grades via mastery rubrics (Marzano, 2011; Townsley et al., 2019, 2020). However, since many standards-based grading systems rely on human judgment (teacher judgment) to evaluate performance, student grades may be prone to the teachers' internalized biases, personal values, or ethnic incongruencies (Legette, 2021). This chapter briefly discusses the biases inherent in the judgment of teachers regarding their students' performance, followed by a critical analysis of decision-making and evaluation processes, including where these processes might be more susceptible to bias.

This literature review mainly discusses studies, reviews, and theoretical works published between 2010 and now, although certain earlier publications are also included due to their relevance to this study. Key search terms used were standards-based grading, rubric-based grading, teacher bias, grading bias, decision-making bias, racial bias, attribution, expectations, shifting standards, cognitive distortions, racial bias in grading, and performance evaluation bias.

Standards-based Evaluation of Student Performance

The following sections discuss two grading scenarios central to standards-based grading models: rubric grading and essay grading.

Rubric Grading

Rubrics have been commonplace in standards-based grading (Gobble et al., 2017) and are considered an essential tool by many teachers in such a grading system (Velasco-Martínez et al., 2018). Although they are often considered tools that provide more accuracy, fairness, and stability to the process of grading student work, they may only appear to be standardized

(Jonsson & Svingby, 2007; Panadero & Jonsson, 2020). In other words, they may be vulnerable to teacher biases like any other grading device or practice.

Although there are various rubric structures and styles, this study uses standards-based rubrics (proficiency scales) because these types are the more frequently used in education. First, standards-based grading rubrics are commonly used to analyze a student's mastery (Reibel et al., 2020) and its associated criteria. In contrast, conventional (point-based) rubrics are often used to evaluate a task completed by the student (Schimmer et al., 2018). Second, standards-based grading rubrics are intended to be used throughout the learning process, while traditional rubrics often appear at the end of the learning process (Reibel & Thede, 2020). Furthermore, in standards-based grading, rubrics are intended to allow students to navigate and manage their learning and become aware of the realities of their learning (Velasco-Martínez & Tójar-Hurtado). Ideally, they enable students to have valuable conversations about learning outcomes with their teachers, peers, and parents. Standards-based rubrics often contain three components (Gobble et al., 2017): mastery scales that holistically rank student performance or knowledge level of a standard. In this study, the skill analyzed is writing, and the biases studied from this aspect include shifted standard, racial bias, and expectancy bias. Second, criteria for success, prerequisite skills, and knowledge are required to attain the standard's mastery level successfully. This study analyzes the criteria of thesis, sophistication, and evidence. Similar to the mastery scale the biases studied from this aspect include shifted standard, racial bias, and expectancy bias. And third, a feedback area for the teacher and student to discuss mastery, performance, and the interplay of performance with the criteria. Further, the bias studied from this aspect is attribution error.

Figure 4 (Appendix F) depicts one type of a standards-based rubric. Teachers may use this standards-based rubric to evaluate a student's mastery level of the standard, in this case, writing, with the associated criteria. The participants in this study received an Advanced Placement (AP) rubric with a four-point mastery gradation using Common Core State Standards for Grade 11 Language Arts to guide their scoring. This rubric (Figure 4, Appendix F) was chosen because it contains many qualities of an effective standards-based rubric (Brookhart & Chen, 2015; Marzano, 2011; Velasco-Martinez & Tojar-Hurtado, 2018). Figure 3 (Appendix F) is a completed example of this standards-based rubric.

Essay Grading

To arrive at a letter grade, standards-based teachers often review students' essays for evidence that they have met the outlined essay criteria instead of point totals, averages, or other numerical grading policies (Jokić, 2017; Scarlett, 2018). In other words, essay grading using a standards-based model involves collecting and interpreting student-produced evidence to give students a reliable rating of their competence in course standards (Reibel & Thede, 2020).

Concerns with Standards-based Rubrics

The balance between classification (feedback about performance) and communication (guidance about how to get better) in standards-based education is delicate. With too much classification and too little communication, students might see the rubric as a static ranking instead of a formative opportunity (Henry, 2018). On the other hand, a heavily communication-based rubric might promote mimicry (Miller, 1956; Bacchus et al., 2020). The result in either scenario is that the student may not view the feedback or evidence from their work as transformational for learning.

Rubric Ambiguity

Since standards-based evaluation systems often rely on these performance scales as their judgment tool, the number of scale levels can also become problematic. Educators might assume a grading scale with more classification levels to be more precise; however, more gradations may only give a precision illusion (Guskey, 2013). The number of gradations on performance evaluation tools may lead to more evaluator bias, although this can be alleviated by narrowing down the rating scale (Rivera & Tilcsik, 2019). When more gradations are added to the scoring scale, a teacher is faced with several options to record the score, increasing pressure on their decision-making processes, and the likelihood of incorrect scoring, because of the need for teachers to justify the difference between each gradation added to a mastery scale (Guskey, 2014). Citing Dwyer (1996), Guskey states, "setting more cutoff boundaries (levels or categories) in a distribution of scores means that more cases will be vulnerable to fluctuations across those boundaries and, hence, to more statistical error" (p. 70). In other words, the more levels on a rubric, the more likely the teacher to mislabel the student's mastery. Along these lines, rubrics can often be applied differently (Kahneman et al., 2021); the interpretation of the term "excellent" on a standards-based grading rubric can vary from teacher to teacher (Gentrup et al., 2020).

Potential Biases in Evaluating Students' Performance

As mentioned earlier, judgment can be influenced by various factors, including personality traits, personal values, cognitive style, idiosyncrasies, and weighting of different considerations. "Performance ratings are highly variable and depend more on the person assessing than on the performance being assessed" (Kahneman et al., 2021, p. 7). This variance is not without consequence. According to Kahneman et al. (2021, p. 21), "Unwanted variability

in judgments that should be identical can create rampant injustice, high economic cost, and errors of many kinds." Moreover, "people want to feel that when judged, those evaluations represent the values of the system, not [an]individual['s] judgments" (p. 53).

Cognitive Bias #1: Racial Bias

Implicit racial biases are "associations made by individuals in the unconscious state of mind [that] cause individuals to unknowingly act in discriminatory ways" (Maryfield, 2018, p. 1), and activated racial bias often positions people of color as potentially vulnerable (Bodenhausen & Richeson, 2010; Fiske, 1998, 2015; Rogers et al., 2020). Moreover, stereotypes and biases can be activated when someone is in the *presence* of another with a different ethnicity (Casper et al., 2010; Sassenberg et al., 2005; Spencer et al., 2016).

Although there has been incremental progress, the national school student archetype remains a White middle-class student (Greenwald & Farnham, 2000; Lewis & Diamond, 2015; Preston, 2007). This privileged White bias functions in many school dimensions—practice, policy, and personnel. Moreover, racial bias may manifest as illusory race-neutral policies, exclusionary pedagogy, and discriminatory grading practices (Lewis & Diamond, 2015; Paslay, 2021). Notably, some studies (Babad et al., 1982; Starck et al., 2020) find that educators are no more or less racially biased than the average American.

A teacher's ability to examine implicit stereotypes and biases can be compromised by physical and psychological demands (De Houwer et al., 2009), such as time, resources, and stress. The lack of resources and time can lead teachers to create inaccurate representations of their students, which may affect their pedagogy (Emdin, 2021). For example, teachers may refer students belonging to the racial majority to specific programs more often than their minority counterparts (Tenenbaum & Ruck, 2007). Additionally, since "white educators teach the vast

majority of racial minorities in the U.S. (Leonardo, 2021), [this] over time could significantly contribute to overall racial disparities in academic achievement" (p. 3).

Jacoby et al. (2016) hypothesized "that greater implicit bias increases whites' anxiety when teaching black students, and that the resultant distraction and depletion will diminish the quality of their instruction and, subsequently, student learning" (p. 51). In their study, they recruited 200 Princeton undergraduates for their study that “included cross-race and same-race dyads” (p.51). For their study, “The first participant, who was always white, was assigned to the instructor role. The second participant, who was black or white, was assigned to the learner role […] After the lesson, the participants were given a five-minute discussion period. Afterward, the participants were separated; the learner was given five minutes to complete a test, and instructors completed a measure of explicit bias” (p.51). Their study findings suggest that White teachers displayed more anxiety when teaching Black learners. The observable anxiety-related behaviors included speech incoherence when teaching, lack of eye contact with the students, and the teachers' choice of physical positioning in the classroom.

In an often-cited study by Van den Bergh et al., 2010, researchers used two tools to measure implicit teacher bias. One tool was an implicit association test to assess bias toward different ethnic groups. The second was a racism scale (McConaughey, 1986) that measured teachers' *explicit* prejudiced attitudes. The researchers compared racism scale scores to the IAT score. They state, “low- versus high-prejudiced people clearly have different representations of an ethnic minority group, which can result in different beliefs and different expectations with regard to the behavior of a member of an ethnic minority group” (p. 502). Additionally, those teachers with high implicit bias may have scored low on the explicit measure, as can happen with

explicit bias measures; participants give the most socially acceptable answers (Greenwald et al., 2016).

In a review of social psychological research on racial bias, Warikoo et al. (2016) examined the literature on racial bias and its effect on students. First, negative implicit associations toward low-achieving groups (stereotypes) often abound in classrooms "not only because they are automatic and difficult to control, but also because they are pervasive" (p. 508). Second, well-intentioned teachers may "sometimes act on unconscious biases towards students from stigmatized groups" (p. 510). Third, "implicit racial associations consistently correlate with problematic feelings and behaviors that emerge during interracial interactions" (p. 510). This may affect student performance and perpetuate problematic feelings between the teacher and the student.

Other studies have examined how racial bias is connected to verbal feedback given to students (Gentrup et al., 2020; Scott et al., 2019). Gentrup et al. (2020) videotaped teachers providing feedback to students about their performance. Independent observers then coded the video based on the types of feedback, language, and other cues the teacher gave the students while discussing their performance. Reviewing the patterns in these codes, the researchers found that the teachers directed more positive or neutral feedback (orally or via body language) when discussing performance with White children than with children of color. Many studies have found similar results: teachers tend to hold higher expectations for White students than African American students (de Boer et al., 2010; Honstra et al., 2018; Sbarra & Pianta, 2001).

Likewise, studies by Fox (2015), Oates (2003), and Joshi et al. (2018) showed how teacher-student racial congruence influenced teachers' scores of student performance. Oates (2003) found a disparity in teacher judgment of student performance, particularly pronounced in

the white teacher–Black student context. Another study (van Ewijk, 2011) found that racially incongruent teacher-student relationships may indirectly induce minority students to perform poorer because ethnic majority teachers appear to have lower expectations from ethnic minority students.

Although many studies demonstrate disparities in teacher expectations and evaluation of student academic achievement based on student race (e.g., Irizarry, 2015; Marcelo & Yates, 2019; Reardon et al., 2019), other studies that explored educator racial bias show inconclusive results, for example, a study by Pigott and Cowen (2000) found insignificant variance in student scores between White and Black students. In their study, Black students were presumed by both Black and White teachers "to have more severe school adjustment problems, fewer competencies, more stereotypically negative qualities, and poorer future educational prognoses than white students" (p. 193).

In line with this research, this dissertation also explores the extent to which implied student race factors into teachers' judgment of student work. For example, if a teacher is negatively biased toward students of color, they may see those students as not as talented as White students, which may cause them to judge the student's work based on that perception (Ferman & Fontes, 2021; Jussim, 1989). Indeed, "any internalized racial prejudice can activate biases and lead teachers to use discriminatory performance evaluations" (Wood & Graham, 2010, p. 177).

Self-reported Measures of Racial Bias. Historically, measurements of stereotyped thinking and prejudiced attitudes have only involved self-reporting procedures (self-administered surveys) that assess people's straightforward attitudes (Jussim et al., 2020). Although applicable, these tools may only reflect a person's socially desirable response and show limited reliability

due to the strong influence of self-presentation effects (Oysterman et al., 2001). Also, they are "not well-suited to capture thoughts and feelings outside of conscious awareness" (Greenwald & Banaji, 1995, p. 4). These measures are often criticized for their inability to capture emotions or thoughts “that people are unwilling to report due to social pressures or concerns” (Gawronski & Hahn, 2019, p. 770).

Teacher Implicit Academic Achievement Association Task (TIAAAT). Another measure of implicit bias developed by Greenwald and Farnum (2000) is the Teacher Implicit Academic Achievement Association Task (TIAAAT). Like Implicit Association Tests (IATs), the central assumption is that the more likely the respondent has an implicit bias, the slower the reaction time to contradictory combinations, such as a flower–evil response latency paradigm. Statistically, bias is considered the "average response latency between conditions…[indicating] differential association strengths among concepts" (Nosek et al., 2014, p. 2). The main difference between TIAAATs and IATs is that TIAAATs measure the relative strength of the association with symbols of academic success and failure to uncover the teacher's implicit academic prejudice. TIAAATs study implicit bias, discriminatory attitudes, and academic discrimination. Greenwald and Farnum (2000) used symbols of academic success or failure (10/10, A+, Excellent; 1/10, F, Poor), following which they ascribed surnames to these signs of achievement or failure.

Peterson et al. (2016) collected student achievement data using a third-party exam from the beginning and end of the year and simultaneously collected data on teachers' explicit hopes for their students’ academic performances. Then at the end of the year, they reviewed implicit prejudiced attitudes toward academic achievement from the TIAAAT perspective, which, similar to the IAT, measured the teacher's level of implicit bias. They found that "teachers with generally

high expectations had higher-achieving students in reading but not in mathematics and that implicit prejudiced attitudes affected mathematics achievement" (p. 132). Further, the teachers who scored higher on the TIAAAT, associated with prejudice, seemed more predisposed to evaluate their racial minority students as "less intelligent and less promising" (Rubie-Davies et al., 2016, p. 136).

Audit Studies. Researchers have increasingly used audit studies to study discrimination in various domains (Butler & Broockman, 2011; Gaddis & Ghoshal, 2015; Hanson & Santas, 2014), including academia, the labor market, and social services. In many audit studies, researchers test for racial and ethnic discrimination by submitting artifacts with fake names or altered personal information that attempts to trigger racially biased thinking or action (Darolia et al., 2016).

One of the most significant concerns of audit studies is the names used to signal racial identities (Butler & Homola, 2017). In this study's context, the primary concern was that the results would be less valid if the participants did not assume "Jamal" to be a Black male. Nevertheless, an audit study is helpful despite this limitation because "they produce clear findings that address important questions but do not require highly complex or assumption-ridden statistical techniques to produce or understand" (Ghoshal, 2018, p. 313).

So far, no study has investigated discrimination in standards-based grading practices using an audit study. The present study uses an approach similar to an audit study, which can infer causality by isolating the main cause of interest and controlling for other factors—a survey vignette experiment (Mutz, 2011).

Cognitive Bias #2: Shifting Standards

Another bias that might affect teachers' judgment of student work is the shifting-standards bias. Teachers may hold an unreasonable or incorrect perspective of a student (Holder & Kessels, 2017) based on their implicit bias toward one's ethnic and cultural background; because of this, they may shift the quality standard.

In a standards-based grading system, teachers have the discretion to determine how a student's work meets the standard. This evaluative discretion may leave teachers' judgment susceptible to shifted standard bias (Schuster et al., 2021). More simply stated, the meaning of "good job" may depend on the perceiver's standard of quality, not an objectively defined one.

It is worth noting that some studies on teacher grading practices (Holder & Kessels, 2017) argue that the more discretion a teacher has in evaluating students, the more likely the disappearance of bias. Schuster et al.'s (2021) study challenged the assumption that immigrant students are not as good at speaking the native language as native Germans. The researchers asked participating teachers to evaluate the performance of a target student (either immigrant or non-immigrant) using either an objective (similar to point-based grading) or a subjective scale (similar to standards-based grading). They found that, on subjective scales, teachers evaluated immigrant students similarly to German students. In a similar study by Tobisch and Dresel (2017), they found that "teachers' achievement expectations are accurate for students with an immigrant background but that teachers overestimate students without an immigrant background and with high socioeconomic status" (p. 748).

Additionally, the same Schuster et al. (2021) study mentioned above also investigated shifting standards and gender bias in teachers' grading of STEM essays. They concluded that they found "direct experimental evidence of the shifting of reference standards to more leniency

towards the negatively stereotyped gender, which may be specific to a presumably student-focused primary school context" (p. 831). Also, the supposedly weaker gender received more elaborate feedback, suggesting a shift in the standard held by the teacher for that gender. In other words, bias did affect the grading of student work.

Likewise, in a study by Forgas (2011), the participants read an essay with a person's picture, either by a male professor or a young woman. Forgas wanted to know if the participants graded the essay differently based on the framing provided (e.g., the image); he found that they did. The essay with the picture of the woman was awarded lower grades than the male professor's. Other studies performed in the legal profession (Frankel, 1973; Bublitz, 2020) found similar patterns, concluding that the judge's sentences often vary based on predilections, views, or biases of the said judge, not the details of the case. In sum, slight amounts of non-essential information can produce substantial variations in the judgment of performance**.**

Shifted Standard of Mastery and Teacher Evaluation of Student Performance. Similar to the curiosities of Schuster et al. (2021) and Tobisch & Dresel (2017) about how *shifts* in the interpretation of standard can lead to unreliable evaluations of student work (e.g., STEM essays) or cause the learning process to appear unjust and inequitable to students, other researchers have made shifted standards a focus of their studies (Biernat et al., 1991; Biernat, 1995; Biernat, 2012; Bertrand et al., 2015). Overt or covert shifts in how a teacher applies a standard while grading student work can distort grades or confuse students about their knowledge or skill level (Kaiser et al., 2017).

Cognitive Bias #3: Expectation Bias

Expectancy theorists suggest that erroneous expectations for students, or anyone, can perpetuate social injustices, stereotypes, prejudice, and a lack of occupational opportunities (Fisk

& Taylor, 1984; Brault et al., 2014; Van den Broeck et al., 2020). Studies suggest that race-based differences in achievement may “reflect widespread and pervasive low expectations on the part of teachers for low SES and Black pupils.” (Strand, 2014, p. 241), which in turn *may* affect how students perform (van Ewijk, 2011). If a teacher has internalized low expectations for racial minority students because of an unconscious bias, these marginalized students may interpret teacher signals as indications that they are not as capable as other publicly affirmed students. This can result in the racial minority group performing lower than the racial majority group (Hornstra et al., 2018). For example, in an oft-cited study, Brophy and Good (1970) observed four first-grade classrooms and studied the interactions between teachers and students, trying to understand how socially significant relationships, such as that between a teacher and a student, teachers' behaviors differ between high-expectation and low-expectation students. They found similar results to Rosenthal and Jacobsen (1968) in that teachers pushed high-expectancy students and appeared to be complacent with low-expectation students.

Correspondingly, teachers' expectations often set students on a learning trajectory. More often than not, this positively affects a student's learning. However, "once a student is set on a particular learning trajectory, it is often challenging to interrupt the learning journey they are on" (Rubie-Davies et al., 2018, p. 12). According to Rubie-Davies et al. (2018), teachers often form expectations for their students based on the information given about their previous performances in different curriculum areas. This idea is similar to the cognitive bias of *excessive coherence* (Kahneman et al., 2021), where the person anchors on a detail or a narrative and makes subsequent judgments from this anchor to create a feeling of stability. "Regardless of the valid reasons for your belief, you will be inclined to accept any argument that appears to support your pre-judgment, even when the reasoning is wrong" (Kahneman et al., 2021, p. 170).

Educational studies suggest that "differences in academic achievement may be due to misguided expectations" (Strand, 2014, p. 151), potentially biasing teachers' evaluations of student performance. One way that expectations become misguided is through *self-fulfilling prophecies*. These occur when a teacher's initial expectation is inaccurate, yet the student acts and performs accordingly (Jussim et al., 1996). In some situations, the student may become whomever the teacher thinks the student is (Rosenthal & Jacobson, 1968).

Jussim (1989) surveyed a public school in Michigan, surveying sixth-grade students and teachers. When the school year started, teachers answered a questionnaire about each student's talent, effort, and performance. For example, how well they answered questions in class, did their homework, and tried hard on projects. At the end of the year, the questionnaires and the standardized test scores were compared to determine if the teachers' thoughts about the students in the questionnaire correlated with the third-party exam scores. He found that teachers "develop expectations for students early in the year, and students do indeed often confirm these expectations" (p. 469). The results showed that teachers' performance predictions (i.e., "I believe this student will do well on tests") correlated with the test scores. In contrast, talent ("I see this student as naturally talented") and effort ("this student is a hard worker") did not. These findings, reminiscent of Rosenthal and Jacobson's (1968) classic *Pygmalion in the Classroom* study, suggest that teachers who thought the students should do well on the state tests worked consciously or subconsciously to make that happen, giving more feedback, praise, or in-depth responses to these students.

Another concept central to expectancy theory is the *halo effect.* Haloing is when there is a pre-judgment and subsequent disregard of conflicting evidence because the first impression (the halo) is the anchor. In other words, the initial positive or negative reaction significantly affects

one's choices, actions, or reactions (Kahneman et al., 2021). An example of the halo effect in education can be seen in Hornstra et al.'s meta-study of expectation literature (2018). The researchers cite that "when high-expectation students gave incorrect answers or did not know the answer to a question, teachers were more likely to rephrase the question and offer another opportunity to respond. In contrast, low-expectation students were more often given the correct answer rather than the teachers rephrasing the question" (p. 325). The consequences of the halo effect can be that teachers use information about their students before a performance, set an expectation, and judge against that anchor. In other words, they may shift the lens through which they view the student's work because of their first impression (Rosenthal & Jacobson, 1968; de Boer et al., 2018).

However, not all research points to teacher expectancy bias. Some studies argue that teachers' expectations are accurate and minimally affect students' learning; therefore, they are not worth much consideration (Brophy, 1983; Jussim & Harber, 2005; Jussim et al., 2009). For example, "teachers' expectations for their students reveal that most teacher perceptions of students are accurate and based on the best available information, and most of those inaccurate are corrected when more dependable information becomes available" (Brophy, 1985, p. 304). Other studies that examined whether teachers' expectations change over time found similar results (Kuklinski et al., 2000; Marshall et al., 1986; Martinek, 1980; McKown & Weinstein, 2008).

Still, some studies argue that teacher expectations are inaccurate, more often than not. In one study with over 11,000 students, the degree of teacher expectations was accurate for only about 33% of the students (de Boer et al., 2010). In a similar study in New Zealand with two separate samples of 1,500 students, teacher expectations were accurate for only about one-third

of the students (Wang et al., 2018). In yet another New Zealand study with over 17,000 students, teacher expectations were unacceptably inaccurate for particular groups of students, that is, ethnic minority students, those from low socioeconomic groups, boys, and those with special needs (Meissel et al., 2017).

The Timing and Longitudinal Effect of Teacher Expectation on Student Outcomes. The timing of expectation formation is critical to the study of expectation bias. In one study, when the teachers received student information two weeks into the school year, no effects were found on subsequent student outcomes because the teacher had already developed a static expectation at that point (Raudenbush, 1984; Rubie-Davies, 2018); that is to say that teachers' early expectations of students might be resilient to change. However, studies similar to Sorhagen's (2013) recorded teacher expectations early on in students' academic careers. When they collected learning evidence several years later, student academic performance reflected the teacher expectations collected early in the student's academic career.

One of the more detailed studies on the longitudinal effect of teacher expectations is Rist's (1970). Like Rosenthal and Jacobson's (1968) study, Rist (1970) used a longitudinal study design to show how teachers' early judgments of student ability impacted students' later achievement. Rist conducted a three-year observational study starting when 30 students began kindergarten. The study was conducted in a poor urban setting; all the students and teachers were African American. Before meeting the students, the kindergarten teachers received information about their former position at the preschool, a list of families' welfare support, medical information, and parental concerns regarding the teacher and their colleagues' previous experiences with the student's siblings. Students were seated in a permanent seating arrangement within the first days of the school year. The researchers observed that the students in one group

seated at Table 1 had a clear physical appearance, were encouraged to speak often, and appeared more relaxed with the teacher. Moreover, the students seated at Table 1 came from homes where parents' incomes were higher, their education was more significant, and the family size was smaller. Tables 2 and 3 were more often occupied by students from divorced or lower-income homes, who did not engage with the class as often and appeared to have messier learning areas than Table 1.

By the time the study reached its second stage, the students had moved to the first grade. The carryover effect of this preferential treatment appeared again when the kindergarten students from Tables 2 and 3 moved to the same tables in the first grade. They interacted with the teachers less frequently and were excluded from classic activities, and their instruction became more infrequent. There appeared to be a compounding effect from kindergarten to the second grade on the teacher's expectations set in kindergarten, such as the over-under estimated achievement of the same students over time.

Regarding the socioeconomic profiles of students, Rist's study suggested that teachers tended to overestimate the ability of those with higher socioeconomic status, especially those perceived as more confident. On the other hand, teachers underestimated students from low socioeconomic backgrounds. In sum, it appears that teacher expectations are often initially inaccurate, but once formed, they tend to be stable (Rubie-Davies, 2017).

Cognitive Bias #4: Attribution Error

Attribution theory provides people with a framework for describing their beliefs about behavior and outcomes and explaining the causality of events (Harvey et al., 2014; Heider, 1958; Kelley, 1967; Weiner, 1974, 1980). Harvey et al. (2014) state that Fritz Heider, one of the first to research attribution theory (1958), "characterized people as naive psychologies with an innate

interest in understanding the cause of success and failure" (p. 128). In other words, people inherently want to know why things happen.

Attribution theory explains how people think of success or failure, even if people themselves may not easily understand this explanation (Hollyforde & Whiddett, 2002). Rooted in Heider's (1958) research, attribution theory seeks to understand how the layperson determines the cause of specific events and how and why individuals assign causality to events in their environment. Variations of the theory of attribution can be found in the works of Dweck (2018); Graham (2016); and Martinko & Mackey (2019).

Within the theory of attribution is a bias known as attribution error. Attribution error is "the tendency for attributors to underestimate the impact of situational factors and to overestimate the role of dispositional factors in controlling behavior" (Ross, 1977; p. 183). Moreover, "the actors [in an event] having a greater tendency to attribute behaviors to situational forces or constraints, and observers [of that same event] instead being more inclined to make attributions implicating the actor's abilities" (Wang & Hall, 2018, p.15). This error can create a distorted picture, thus creating biases and variability (Beckman & Rodriguez, 2021). Although, as found, behavior is ambiguous and may have many possible causes (Westra, 2019).

Weiner's Model of Attribution (1974). One commonly cited attribution framework is Weiner's model, which he often applied to the classroom context (1972, 1974, 2012, 2018). Weiner posits "that it is not just the success or failure of activities that engender pride or shame, but also the explanations that the person attributes to the causes of success or failure" (Weiner 1980, as cited on p.31 of Hollyforde & Whiddett, 2002). His model has two general dimensions: the locus of causality and stability. The locus of causality examines if success or failure results from an internal or external cause. In contrast, stability analyzes whether success or failure

results from stable (e.g., skills or task difficulty) or unstable causes (e.g., effort or luck). Similarly, Table 1 below summarizes the features of Weiner's model.

Table 1

Weiner's Attribution Loci (Weiner 1974)

	Controllable		Uncontrollable	
	Stable	Unstable	Stable	Unstable
Internal Locus	Typical Effort	Malleable Ability	Innate Ability	Mood, emotion
External Locus			Task/Standard Difficulty	Luck

Attribution Theory and Education. Differences in race, student dispositions, pre-performance information, expectations, and background can lead teachers to make erroneous conclusions about how and why students perform the way they do (Riegle-Crumb & Grodsky, 2010; Daneshzadeh & Sirrakos, 2018). More importantly, a teacher's explanation of a student's performance, whether internal or external, can inform a teacher's subsequent expectation of future student performance, potentially leading to another shift in the standard used to grade student work (Kelley, 1967). For example, when a teacher judges the student's essay, and it does not fit within their predetermined model of student learning and dispositions, they may be inclined to seek out evidence that either supports that narrative or explains (i.e., makes an attribution about) the deviation from it.

Moreover, attribution error may occur when a teacher considers that a student's low exam score results from the student's disinterested attitude (internal and unstable dimensions) versus a low score resulting from difficult exam questions (external and stable dimensions). The same can be said about high exam scores. A teacher may incorrectly attribute a student's success to an

illusion of mastery (internal, stable), easy test questions (external, unstable), or even randomness (external, unstable).

Even more problematic is that studies show that White teachers tend to attribute White students' academic failure to situational factors (external), such as a challenging test, and tend to attribute Black and Latinx students' failures to within-child factors (internal), such as laziness or disinterest (Lewis & Diamond, 2015; Schmalor et al., 2021; Steele, 2018).

In a more recent study, Carson (2019) expanded the attribution theory by studying worker performance in organizations to include the idea of *relational attributions*, which he explains as "explanations made for outcomes or behaviors experienced or observed by a focal individual that locate the cause of an outcome within the relationship between two others (people, groups, organizations, or any combination) not including the focal individual" (p. 544). His findings suggest that said individual can also attribute the interpreted relationship between an individual and others to the cause of an outcome. Similar to Carson's study, Gardner et al.'s (2019) study connected attribution theory to student grading when it examined the critical roles of attributions for outcomes (e.g., student performance grades) across socially significant relationships (e.g., teacher-student), concluding that relationships may impact the interpretation of student work.

In sum, attempting to understand what causes certain things to happen is the attribution process. Teachers may tend to emphasize internal characteristics, motives, perceived attitudes, or personalities to explain a person's behavior rather than considering situational or school climate factors (Dobelli, 2013; Wilson, 2011).

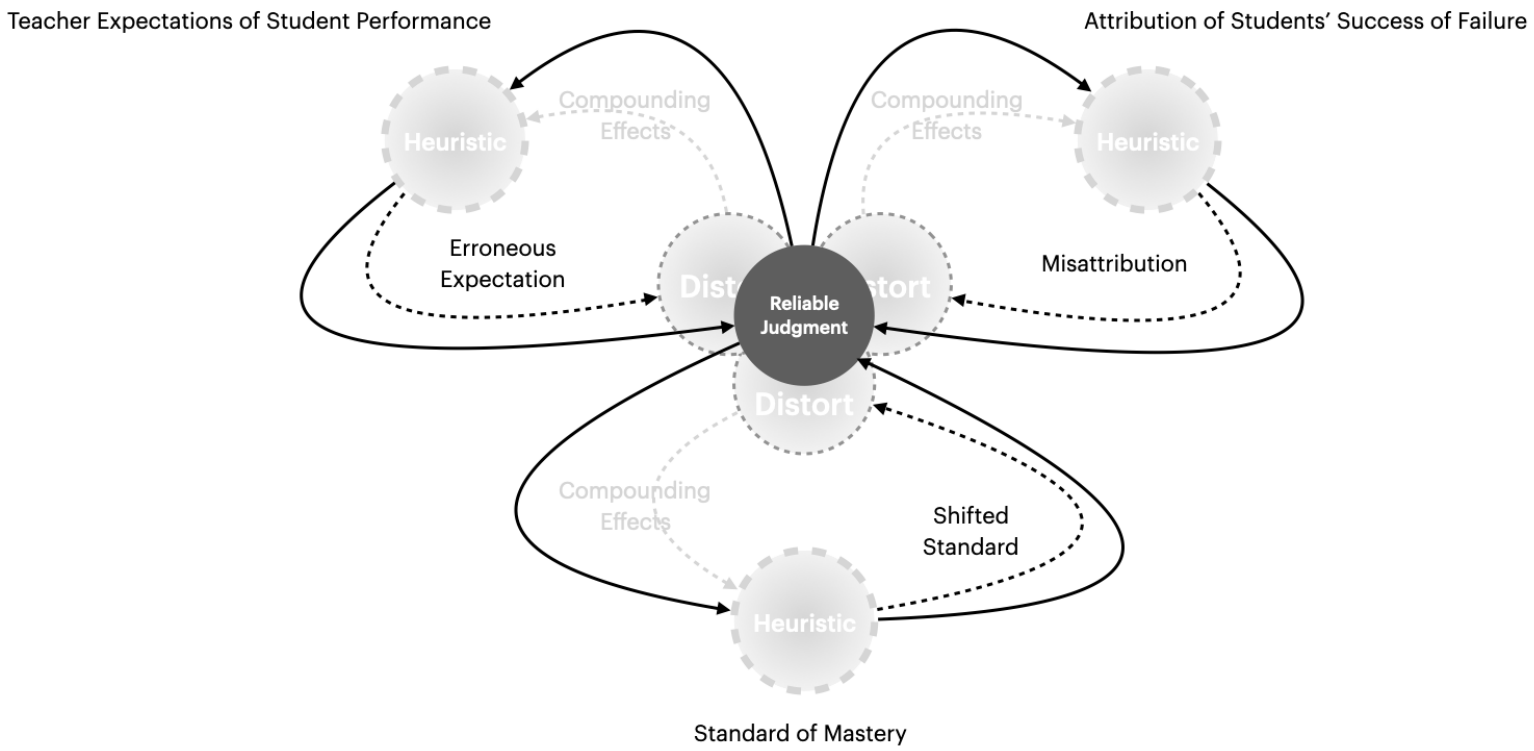
Intersectionality of Cognitive Biases in Education

People create mental maps of the world based on values, beliefs, and assumptions that help us understand concepts, their decisions, and our society (Spillane & Miele, 2007). As with

any phenomenon, one's understanding of the world can be distorted by personal biases, prejudices, and life experiences (Fischhoff et al., 2002). When people use a mental model based on biased thinking, they can distort their interpretation of reality, incorrectly predicting their future and misjudging their past (Weick et al., 2005). Problems often arise because we are overconfident in our understanding of the world and underestimate our objective ignorance (Taleb, 2007). In a foundational study, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Meehl (1954) describes the problems that emotional reactions to our judgments create in a similar way to other researchers, that the "satisfaction you feel with the quality of your judgment [could be] an illusion—the illusion of validity." (Kahneman et al., 2021, p. 115)

Considering the intersection of racial bias, shifting standards, expectancy bias, and attribution error is essential to making more informed decisions about implementing standards-based grading models. Figure 1 visualizes how biases can affect judgment and shows the intersection of these four theories in the classroom system. Teachers' judgments can often be distorted by cognitive biases such as racial prejudice, shifting performance standards, inaccurate expectancy, and incorrect attribution of success and failure (Gilovich et al., 2002). The repeating loops indicate the possibility of the teacher's decision-making being simultaneously affected by various cognitive biases. In other words, it shows the cognitive biases that can overlap and compound in a single judgment a teacher may make about student work or performance. This figure also shows the compounding effect biases can have since initial decisions lead to subsequent decisions, and any misrepresentation or distortion in the original decision can further distort judgment.

Figure 1: Intersection of Heuristics, Expectation Bias, Shifting Standards Theory and Attribution Error



This dissertation approaches the concept of accuracy in grading as a teacher's assigned rubric score being close to an accepted value for a given quality of student work, regardless of what other extraneous information the teacher is given when completing their scoring task. In this context, making an accurate judgment equates to making a reliable one.

Summary

In summary, there are several sources of variability in judgment (Kahneman et al., 2021; Wilson, 2011): level, pattern, occasion, and system-level variability. Level variability refers to the variability between teachers and complex and easy graders. Pattern variability occurs when a teacher is a hard grader in one context but not another; it is not always stable, but it can be and is typically the more significant contributor to variability in judgment. For example, a teacher prides themselves on growing students in writing; therefore, grading student writing is always more complicated. Occasion variability occurs when a teacher's judgment is affected by situational factors such as feelings and weather. System-level sources of variability are the "unwanted variability and judgments that should ideally be identical" (Kahneman et al., 2021, p. 21).

CHAPTER 3

METHODOLOGY AND RESEARCH DESIGN

This chapter discusses the methods used to carry out this vignette survey experiment. It begins by restating the purpose and the central research questions, followed by a description of the 3 × 2 factorial research design. Also, the chapter outlines the ethical considerations made and the process of recruiting the sample participants. A detailed description of the data instruments used and an overview of the data analysis are also included in the chapter.

Purpose and Research Questions

This dissertation focused on cognitive biases (racial bias, shifting standards, expectancy bias, and attribution error) and their effect on teachers' grade determination using standards-based grading models, particularly rubric grading of an essay. These phenomena are critical to investigate because standards-based grading is offered to lessen grading bias (Iamarino, 2014; Muñoz & Guskey, 2015). The overarching research question guiding this study was as follows: To what extent are teachers influenced by their cognitive biases when using standards-based rubrics to evaluate student performance?

The specific sub-questions were as follows:

1) To examine potential racial bias: How does a student's implied race affect the grade teachers award based on a standards-based rubric?
2) To examine potential shifted standards and expectation biases: How does a teacher's additional background information on the student, as presented in a student learning profile, affect the grade teachers award based on a standards-based rubric?

3) To examine potential attribution error processes: How does a student's implied race or learning profile relate to the content conveyed in teacher feedback?

Research Design: Factorial Vignette Experiment Study

Factorial survey experiments "ask participants to respond to hypothetical objects or situations (vignettes)" (Sauer et al., 2020, p.196), and are a common tool used in the social sciences, perhaps because "they combine experimental design features (i.e. randomization) with the advantages of heterogeneous respondent samples"(Sauer et al., 2020, p.196). Factorial designs featuring random assignment (explained in the sections to follow) improve upon correlational regression analyses from observational data because they can establish causation between studied variables (Mutz, 2011). In this dissertation, one independent variable tested for its effect on rubric scores was profile information (positive tone, negative tone, and name only), and the other independent variable was implied student race (names likely to be attributed to White and Black individuals). The dependent variable was the teacher's score of the student's writing mastery. Since there were two independent variables, one with three options and one with two options, there were six resulting conditions (see Appendix A, Table 10). The 3 × 2 factorial design was used to estimate the following: the main effect of profile information on student scores; the main effect of implied race on student scores; the mean scores given to negative, positive, and name-only profiles, and Jamal (Black student) and Jake (White student) profiles. Further, interaction effects between the two main manipulated variables (profile type and student race) were estimated. Including interaction effects in the statistical model means that one independent variable's effect on the dependent variable is theoretically expected to depend on the other independent variable. For example, it helped to understand if the effect of profile

tone on student grades depended on implied race or if the effect of implied race on student grades depended on the profile tone.

Name Selection in Experimental Studies

To ensure that the intended reaction to the name is evoked, the same criteria were used to select names as is with audit studies; selecting form names used in previous audit studies or using government population statistics (Butler & Crabtree, 2017; Gaddis, 2017a). The same criteria were used in this study. For student names, they were selected from lists containing first names historically associated with White males (Jake) (Levitt, 2014) or Black males (Jamal) (Bertrand & Mullainathan, 2004). To avoid overcomplicating the results only first names were used in this study.

Selection of Profile Information and Tone

Similar to Pager's (2003) famous audit study, *Mark of a Criminal Record*, this experiment used three potential profiles (see Appendix A, Table 11): one was the student's name only; the second profile included name and information about prior performance with a positive tone; the third included name and information about prior performances with a negative tone. This additional piece of profile information was included so that it could activate various biases as the teacher scored the essay. Thus, this experiment is designed to examine several potential sources of biases well as how they might interact with race.

Sampling

The study involved a sample of 219 educators who identify as White, have teaching experience, and are of employment age. My target sample size was 200 educators, based on a power analysis performed using G-Power. Based on a small minimum detectable effect size of 0.20, at a power of 0.80, one numerator degree of freedom (because the design features two

levels), six groups (i.e., assignment to one of six scenarios), and no covariates, the minimum necessary sample size was determined to be n = 199. I opted to stay close to this number due to cost constraints.

I sought educators who identified as White, as this is a study in part about racial bias, particularly White individuals' potential racial bias toward Black individuals. With a White middle-class ideology often referenced as the status quo in U.S. educational institutions (Emdin, 2021), the direction of prejudice and racial oppression moves primarily from White to Black. Therefore, I included only White participants to understand if standards-based grading may not be as equitable of a grading system as is touted, partly because most teachers in the U.S. are White (Dee, 2004; Williams, 2008; Leonardo & Boas, 2021).

I designed a survey using Qualtrics software and created a survey flow that automatically divided the respondents into six randomly assigned groups (see Figure 1). The website Prolific recruited participants and sent them to the Qualtrics survey link. Further, I sought the participation of educators, hoping to attract those who had experience grading essays with rubrics, although this was optional. Educators from any subject area or level of teaching experience (primary or secondary) were allowed to participate. Last, the years of experience were not considered; any new or veteran educator could participate in the study. Prolific used these criteria to set up the recruitment parameters for survey distribution.

Procedure, Instrumentation, and Measures

The participants were randomly assigned one of the three profiles to read (positive tone, negative tone, or name only) and one of the two names, but all were given the same essay. The random assignment was balanced by the data system Qualtrics, resulting in six sub-groups with a total of approximately 35 participants in each group (See Table 10, Appendix A). The goal of

random assignment and balancing was to control the effects of gender, rubric experience, subject taught, and age on how a participant would grade these essays. A balanced sample would indicate that the randomization process was effective and helped rule out variances among respondents (individual differences/characteristics/demographics) as the reason for their responses.

Mastery Rubric

To score the hypothetical student essay, each participant received the same standards-based grading rubric developed by Stevenson High School, Lincolnshire, Illinois. Advanced Placement (A.P.) teachers at this school have developed a four-point mastery rubric using the A.P. curriculum for Grade 11 Language Arts (see Figure 4 Appendix E). During the A.P. junior English curricular team meetings, the professional learning community collaboratively developed the mastery scale and associated success criteria. Participants were asked to fill out the rubric for one essay and then comment on how the student might do on a hypothetical second essay (see Appendix D). This study's rubric contained the critical components of effective rubrics (Brookhart & Chen, 2015; Hasan, 2022; Marzano, 2011). Further, the essay used in this study is written by an actual student from the 2019 AP English Language course.

Measures

Independent variables. These were the two "manipulated" – that is, varying and randomly assigned – variables: student race, signaled using two student names (Jake or Jamal), and profile type with the student's previous performance information (name only, negative tone, or positive tone).

Outcome measures. There were four ways the rubric measured student performance: mastery scores on writing (on a scale of four, exceeds to still developing), thesis (on a scale of

two, developed or not developed), evidence (on a scale of two, sufficient or not sufficient), and sophistication (on a scale of two, yes or no).

Descriptive Statistics

Table 2 provides descriptive statistics on the variables and measures.

Table 2. Sample Descriptives

	Full Sample Mean (SD)	Jake Mean (SD)	Jamal Mean (SD)
Respondent Characteristics			
Male	.48 (.50)	.41 (.49)	.54 (.50)
Rubric Experience	.53 (.50)	.49 (.50)	.56 (.50)
Age Range			
18-25 years old	.34 (.48)	.28 (.45)	.40 (.49)
26-35 years old	.42 (.49)	.44 (.50)	.39 (.49)
36-50 years old	.20 (.40)	.21 (.41)	.18 (.38)
51+ years old	.05 (.21)	.07 (.26)	.02 (.13)
Subject taught[1]			
ELA	.27 (.44)	.31 (.46)	.23 (.42)
Math	.15 (.35)	.09 (.29)	.21 (.41)
Science	.14 (.34)	.15 (.36)	.12 (.33)
Social Studies	.08 (.27)	.07 (.26)	.08 (.28)
Other	.30 (.46)	.30 (.46)	.30 (.46)
Outcome Variables			
Mastery Level (range 1-4)	2.63 (.77)	2.65 (.67)	2.60 (.86)
Name Only	2.52 (.69)	2.57 (.64)	2.47 (.74)
Positive Profile	2.93 (.80)	2.91 (.64)	2.94 (.95)
Negative Profile	2.43 (.73)	2.45 (.66)	2.40 (.78)
Criteria Scores			
All Criteria	.46 (.50)	.46 (.50)	.46 (.50)
Criteria 1 – Thesis	.60 (.49)	.61 (.49)	.58 (.49)
Criteria 2 - Evidence	.51 (.50)	.51 (.50)	.51 (.50)
Criteria 3 – Sophistication	.28 (.45)	.26 (.44)	.29 (.46)
N	219	110	109

Notes. [1] Subject taught does not add up to 100 due to an N/A response category (e.g., these respondents may teach in an elementary school).

Table 2 shows that the sample was nearly equally split by gender and rubric experience, with 105 male participants (48%) and 112 female participants (51%) having prior rubric experience. 91 participants were 26–35 (42%), and the next largest age group was 18–25 (75 participants, 34%). Most participants had experience teaching English, with almost (59 participants, 27%), followed by math (33 participants, 15%), Science (30 participants, 14%), or Social Science (17 participants, 8%). 66 participants listed *other subjects* as their teaching experience.

The descriptive of the subsamples (i.e., those assigned to the Jake or Jamal scenarios) is similar, indicating that the random assignment effectively balanced out these parameters. For example, there were 46 males assigned to a Jake profile (41% of 110 participants in this subgroup), 54 of them had prior rubric experience (49% of 110 participants). The majority (71%) were under 35 (31 participants aged 18-25 and 48 participants aged 26-35). In comparison, in the group assigned a Jamal profile, there were 59 males (54% of 109 participants in this subgroup), 61 participants had rubric experience (56% of this subgroup), and the majority (70%) were under the age of 35 (44 participants aged 18-25 and 43 participants aged 26-35). Again, 59 participants had experience teaching English, 33 in math, and 30 in Science in both the Jake and Jamal profiles. There were 17 Social Studies teachers and 66 participants who listed *other subject* in their profiles. In sum, the subsamples differed the most regarding gender, but this is only because of the random assignment.

Quantitative Data Analysis

Ordered logistic regression is a type of statistical analysis that tests for the influence of independent variables and their interactions (e.g., name and profile tone) on a dependent variable

that is neither dichotomous nor continuous(e.g., mastery and criteria scores for student writing). Ordered logistic regression was deemed appropriate for this study's ordinal dataset (UCLA ARC, 2022). Several factors were considered in reaching this decision. Since examples of ordinal variables include scales similar to this study's rubric (e.g., "exceeds" through to "still developing"), it could be confidently concluded that the data set is ordinal. Second, one or more independent variables are ordinal or categorical. An example of categorical variables includes categories such as gender (e.g., male and female). In the study, the categorical variables are the student name and profile type. Third, there seems to be no strong intercorrelation among two or more independent variables in a multiple regression model, making it easier to understand the explanation of the dependent variable.

The following significance levels were adopted and are standard in other studies reviewed in Chapter 2 (e.g., de Boer et al., 2010; Gravetter et al., 2020):

1) A *p*-value of less than .05 for the implied race indicates that student race has a statistically significant effect on teachers' assigned scores;
2) A *p*-value of less than .05 for profile type indicates that that profile type has a statistically significant effect on teachers' assigned scores;
3) A *p*-value of less than .05 for the interaction between implied race and achievement profile indicates that the effect of student race depends on the profile type.

Using ordinal regression analysis, I could assess the effect of implied race and profile information on teachers' assigned scores while detecting any interaction between implied race and profile information.

Qualitative Data Analysis

Since it is difficult to perform a grading study without including an analysis of feedback, as the two are so intimately connected in the process of rubric and essay grading, I am using a mixed method approach to analyze the *counts* of different kinds of teacher feedback to get a sense of what is the most common comment teachers made to students about their writing. As Small(2011) explains in a methodological review piece, such counts can be a form of quantized qualitative data analysis. The qualitative data were drawn from open responses collected from participating teachers (2–3 sentences each). The data were collected as the last part of the student feedback, where the participants were asked to provide feedback to the student about a hypothetical upcoming essay.

To study attribution error this dissertation used qualitative codes to capture the essence and essential elements of the feedback segments (Saldana, 2021). The data were interpreted using Weiner's Attribution Model (Table 3). Two rounds of coding were performed to filter and highlight the salient features of the feedback (Coffey & Atkinson, 1996).

Table 3

Weiner's Attributions (adapted from Weiner 1974)

	Stable	Unstable
Internal	Fixed Ability, Inherent Skill	Malleable Ability
External	Rigor and Difficulty of Task/Standard	Occasion Noise, Luck

The next coding round was taking the Weiner quadrants and applying them to the themes in the teacher feedback. This was done to identify patterns in the codes to extract educational themes associated with Weiner's attributions. Table 4 below shows the teacher feedback coded with Weiner's attribution:

Table 4

Weiner's Attributions and Teacher Feedback

Attribution (Weiner 1974)	Teacher Feedback	Example
ES: Rigor and Difficulty of Task/Standard	Refer primarily to a standard of writing	(There is a standard for writing that you must achieve.)
IS: Fixed Ability, Inherent Skill	Refers to Static Characteristics and Inherent Ability	(You are a natural writer)
IU: Malleable Ability	Refers to Malleable Ability and Growth	(You can be a good writer)
EU: Occasion Noise, Luck	Refers to uncontrollable varied external factors	(It wasn't your day)

Limitations

Although this research design has the strength of isolating effects on the dependent variable and is based on a relatable scenario for teachers, it has a few limitations. First, the essay is a student sample from a retired AP exam and not written for the sole purpose of this study; hence, it may not align precisely with some of the rubric criteria, which could distort the final rubric grades. Meaning that the criteria descriptions may not be the language that a teacher uses for their proficient descriptions of thesis, evidence, and sophistication. Additionally, there are many different standards-based rubrics, and the one in this study may not be in line with the participating teacher's rubric experience. Also, the sample size and characteristics, which were limited due to cost constraints and concern about collecting the sample promptly, may limit the generalizability of the findings. Also, only participants who self-identified as White were recruited; some had limited experience with using rubrics. Further, the participants may not have interpreted the coded name or the profile information in an intended way. I chose names commonly associated with Black and White races to address this possibility (Conaway & Bethune, 2015; Gaddis, 2017a; Staats, 2016).

Also, unlike a true audit study, people knew they were being studied, which may have influenced the results. Further, there was no time for inter-rater reliability minimization. All people were from different schools and could not collaborate beforehand with the rubric. And there was no common definition of mastery or criteria presented to the participants before starting the engagement in the study.

One larger limitation of this study is that there is no way to know if the evaluators used a rubric or mastery scale similarly. "Individuals may differ in their interpretation of labels even when they agree on the substance of the judgment" (Kahneman, 2021, p. 186). In education,

teachers may agree on the proficiency of the performance but assign different scores because they use the rubric differently. In this study, the scoring of the essay might have been subtly affected because the participants may have weighed different aspects of the writing differently. For example, an evaluator may weigh creativity more than the writing form, whereas another teacher may weigh substantive detail over organization or tone.

As Fish (2017), who also used a factorial vignette design to investigate bias in teachers' assignments of students to special education or gifted classes, there are limitations to this research design. First, it "removes some of the complexities of human interaction […] such as […] the extensive information teachers often have about their students' learning from their daily interactions" (Fish, 2017, p. 331). Additionally, she asserts that "teachers participating in a factorial vignette survey face fewer constraints and incentives than in the complex school-based decision-making process. Thus, the quantitative estimates of the effects of this experiment might not accurately reflect a real-world context." (Fish, 2017, p. 331). This sentiment holds for this study's factorial vignette survey as it is a simplified simulation of a more complex process of teaching and grading students.

Researcher Bias

A person's perspectives, values, and lived experiences can influence the results of a research study, even in a quantitative study with rigorous controls in place (Zuberi & Bonilla-Silva, 2008). Instead of operating from an objective point of view, researchers may unknowingly allow their biases and positionality in society to influence their research process, data collection, and data analysis. I am a White male from a middle-class suburban background and this positionality may have unknowingly influenced how I constructed the research instrument and my interpretation of the results. Additionally, my educational background and expertise in

standards-based grading could influence my actions in this experiment's design and administering processes. Particular biases that I am concerned about in my dissertation methodology are as follows:

1. Design bias – selecting the sample essay and the data collection structure was my choice and may have been vulnerable to bias.
2. Confirmation bias – searching only for results that confirm my preconceptions.
3. Cultural bias – prioritizing my values and culture as I analyze the results

I mitigated these biases by conducting a pilot study with a diverse group of people (teachers who identified as White or as a person of color, both male and female), seeking their perspective during data analysis (educators and experienced researchers in this study's area of focus, both male and female), and re-evaluating the results throughout the research process. The perspective-seeking I engaged in during data analysis involved reviewing the data from multiple perspectives. During the data review process and before my write-up of the study's findings, I sought alternate points of view from the results. I discussed insights, knowing that dissonant viewpoints may emerge from these conversations.

Summary

This chapter included a discussion on this study’s purpose, the methodology of a survey vignette experiment, the design of the experiment, data collection procedures, and participants. A factorial vignette survey experiment was used to explore the vulnerability of a standards-based grading system to bias and the resulting influence on teacher evaluations of student performance.

CHAPTER 4

RESULTS

As discussed in Chapters 1 and 2, standards-based grading is often offered as a solution to inequitable grading practices (Berns, 2020; Buckmiller et al., 2017). Although there are many factors to consider when switching grading models, the psychology of human judgment and its impact on the interpretation of student performance is often underexplored. If products of our adaptive unconscious, cognitive biases (Wilson, 2011), operate when using a standards-based grading model, then it may continue to result in inaccurate grades and distort feedback, and even perpetuate stereotypes.

Quantitative Results

As a reminder, the following research questions were explored: 1) How does a student's implied race affect the grade teachers' award on a standards-based rubric? 2) How does a teacher's additional background information on the student, as presented in a student learning profile, affect the grade teachers award on a standards-based rubric? 3) How does a student's implied race or learning profile relate to teacher feedback?

The regression analyses that follow examine how the different experimental conditions of student race and student profile type affected respondents' mastery scores and criteria evaluations on the rubrics. Several additional models were used to control for the influence of other unrelated factors, such as respondent's past rubric grading experience, age, and subject area. Reviewing the distribution of scores assigned by the respondents across the different conditions is useful. Concerning mastery scores, the sample assigned an average score of 2.37 out of a possible of 4 points on the rubric (see Table 2, for sample descriptives in Chapter 3). For each of the three profile types (name only, positive, and negative), the average was 2.52 for name

only, 2.93 for positive tone, and 2.43 for negative. This shows that, on a descriptive level, positive profile types were associated with higher mastery scores. In the Jake subsample, the overall score for Jake was 2.35, and the name-only, positive, and negative profiles received 2.57, 2.91, and 2.45, respectively. For Jamal's subsample, the average mastery score was 2.39, with the name only being 2.47, the positive tone being 2.94, and the negative tone profile being 2.40.

Models

Tables 5-8 present the logistic coefficients for the ordered logistic regression models. The means and coefficients from these models allow for an assessment of the impact of the randomized independent variables (name and profile) on the teachers' essay scoring using a standards-based rubric. The more saturated models in each table include individual-level controls of participant age, subject matter experience, and whether they have used a standards-based rubric before.

The ordered logistic regression models presented in Table 5, Models 2 through 6, include the effects of the different profiles (name only, positive tone, negative tone). The coefficients in these models indicate that participants who received the negative tone profile gave lower mastery scores than participants who received the positive tone profile. These effects showed significance, *p*-values below and 0.001 (Model 2) and 0.01 (Models 3–6). Estimates in Model 2 of Table 5 reveal that name-only profile (Black versus White-implied name) had no significant impact on the scores, with the participating teacher not awarding any lower scores for Jamal than for Jake. The name-only logistic coefficients in Model 1 suggest that the name of the student alone did not affect the scoring of the essay; this was true in all models, no matter what additional variables were added to the models.

Table 5. Effect of Profile Type, Age, Rubric Experience, and Subject Taught on Mastery Rubric Scores for an Essay

	Model 1 Name	Model 2 Profile Type	Model 3 Name and Profile Type Interactions	Model 4 Interactions, controlling for teacher rubric experience	Model 5 Interactions, controlling for teacher age	Model 6 Interactions, controlling for subject taught
Jake	0.07 (0.25)		0.04 (0.44)	0.05 (0.44)	-0.04 (0.46)	-0.01 (0.44)
Profile Type						
Basic (name only)		0.25 (0.30)	0.14 (0.44)	0.15 (0.44)	0.78 (0.45)	0.11 (0.45)
Positive		1.31 (0.33)***	1.47 (0.48)**	1.48 (0.48)**	1.46 (0.48)**	1.45 (0.49)**
Negative (ref)		--	--	--	--	--
Jake*Basic Profile			0.21 (0.62)	0.19 (0.62)	0.34 (0.63)	0.27 (0.63)
Jake*Positive Profile			-0.28 (0.64)	-0.21 (0.64)	-0.24 (0.65)	-0.21 (0.65)
Prior rubric experience				0.35 (0.26)		
Age						
18-25 Years					0.06 (.62)	
26-35 Years					0.45 (.61)	
36-50 Years					0.07 (.66)	
51+ Years (ref)					--	
Subject taught						
ELA						-0.59 (.56)
Math						-0.49 (.61)
Science						-0.56 (.60)
Social Studies						-1.47 (.68)*
Other						-0.35 (.56)
Observations	219	219	219	219	219	219
Log Likelihood	-251.23	-242.07	-241.78	-240.90	-240.6	-238.82

*Standard Errors in Parenthesis + p<.10, *p<.05, **p<.01 ***p<.001; "ref" is omitted reference category.*

Model 6 of Table 5 which includes interactions between the two manipulated factors (student name and profile type), and controls for participant subject area, shows that in addition to the positive type main effect remaining significant as in the other models, having a social studies background was also negatively and significantly related to the mastery score. This means that social studies teachers were likelier to assign a lower score than other subject area teachers.

In sum, these results do not support the first part of the study investigating racial bias in using standards-based rubrics. Nevertheless, they do support the second part of the study, that is, profile information and its tone seem to affect the scoring in a standards-based model; the more positive information a teacher is given about a student before grading, the more likely they are to be favorable with their grading, and teachers are more likely to award lower scores when they are under the perception that the student is struggling. This effect was statistically significant even when controlling for participant age, standards-based rubric experience, and the subject taught.

Criteria Models

The influence of teacher bias in decision-making can be seen in the judgment success criteria on the rubric. Tables 6–8 show that the effect of profile tone on teacher judgment was statistically significant in some of the scorings of criteria (thesis, evidence, and sophistication).

Table 6. Effect of Profile Type, Age, Rubric Experience, and Subject Taught on Mastery Criteria Scores (Thesis) for an Essay

	Model 1 Name	Model 2 Profile Type	Model 3 Name and Profile Type Interactions	Model 4 Interactions, controlling for teacher rubric experience	Model 5 Interactions, controlling for teacher age	Model 6 Interactions, controlling for the subject taught
Jake	0.09 (0.28)		0.46 (0.48)	0.48 (0.48)	0.50 (0.50)	0.46 (0.48)
Profile Type						
Basic (name only)		0.10 (0.33)	0.39 (0.47)	0.42 (0.48)	0.37 (0.48)	0.47 (0.48)
Positive		0.60 (0.35)+	0.87 (0.49)	0.90 (0.49)+	0.89 (0.49)+	0.99 (0.50)*
Negative (ref)		--	--	--	--	--
Jake*Basic Profile			-0.58 (0.67)	-0.63 (0.68)	-0.54 (0.69)	-0.65 (0.68)
Jake*Positive Profile			-0.55 (0.69)+	-0.70 (0.71)	-0.63 (0.71)	-0.64 (0.71)
Prior rubric experience				-0.58 (0.29)*		
Age						
18-25 Years					0.67 (0.71)	
26-35 Years					1.07 (0.70)	
36-50 Years					0.50 (0.74)	
51+ Years (ref)					--	
Subject taught						
ELA						-0.60 (0.66)
Math						-0.89 (0.71)
Science						-0.51 (0.72)
Social Studies						-0.92 (0.78)
Other						-0.79 (0.66)
Observations	219	219	219	219	219	219
Log Likelihood	-147.50	-145.71	-145.24	-143.13	-143.13	-144.05

*Standard Errors in Parenthesis + p<.10. *p<.05, **p<.01 ***p<.001; "ref" is omitted reference category.*

The coefficients in Table 6, Model 4 show a marginally significant relationship between the positive profile and criteria score, with $p < 0.1$, while Model 6, which controls for the subject taught by the participants, shows a stronger relationship between positive type and criteria score, with $p < 0.05$.

From Model 4 in Table 6, which control for rubric experience and teacher age, one can conclude that participants with prior rubric experience reported lower criteria scores, on average, but that this did not mediate the marginally significant effect of positive profile type on criteria score, with $p < 0.1$. In sum, the estimates in Table 6 suggest patterns that positive profiles affect criteria scores more than, regardless of, the student's name. In contrast, the name-only profile showed no statistical significance in the criteria scores.

Table 7. Effect of Profile Type, Age, Rubric Experience, and Subject Taught on Mastery Criteria Scores (Evidence) for an Essay

	Model 1 Name	Model 2 Profile Type	Model 3 Name and Profile Type Interactions	Model 4 Interactions, controlling for teacher rubric experience	Model 5 Interactions, controlling for teacher age	Model 6 Interactions, controlling for subject taught
Jake	-0.02 (0.27)		-0.35 (0.48)	-0.35 (0.48)	-0.37 (0.49)	-0.15 (0.50)
Profile Type						
Basic (name only)		0.12 (0.33)	-0.40 (0.47)	-0.40 (0.47)	-0.40 (0.48)	-0.27 (0.50)
Positive		0.69 (0.34)*	0.75 (0.48)	0.75 (0.48)	0.75 (0.49)	1.02 (0.51)*
Negative (ref)		--	--	--	--	--
Jake*Basic Profile			1.01 (0.67)	1.01 (0.67)	1.08 (0.60)	0.94 (0.70)
Jake*Positive Profile			-0.07 (0.68)	-0.07 (0.68)	-0.02 (0.69)	-0.26 (0.71)
Prior rubric experience				-0.02 (0.28)		
Age						
18-25 Years					0.25 (0.72)	
26-35 Years					0.34 (0.71)	
36-50 Years					-0.25 (0.75)	
51+ Years (ref)					--	
Subject taught						
ELA						-1.89 (0.72)**
Math						-0.69 (0.77)
Science						-1.49 (0.77)+
Social Studies						-0.82 (0.84)
Other						-1.75 (0.72)*
Observations	219	219	219	219	219	219
Log Likelihood	-151.74	-149.24	-147.60	-147.59	-146.29	-139.96

*Standard Errors in Parenthesis + p<.10. *p<.05, **p<.01 ***p<.001; "ref" is omitted reference category.*

Table 7 shows the same patterns of significance on the rubric's *evidence* criteria as in Table 6. Model 2, however, shows that, again, the positive tone profile received higher evidence criteria scores, with a statistically significant *p*-value of .046, and a non-significant effect was observed for basic profiles. The ordered logistic regression coefficients in Models 4, 5, and 6, which controlled for participant age, rubric experience, and subject taught, were statistically significant. Experience in teaching Science was marginally significant, with $p < 0.1$. Additionally, ELA and" other subject" showed patterns of significance with p values of (.009) and (.015), respectively. The direction of the significance for Science, ELA, and "other subject" was negative, meaning that the teachers from these subject areas tended to give lower scores in the *evidence* criteria. Perhaps suggest that they were harder graders on this criteria.

Table 8. Effect of Profile Type, Age, Rubric Experience, and Subject Taught on Mastery Criteria Scores (Sophistication) for an Essay

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Name	Profile Type	Name and Profile Type Interactions	Interactions, controlling for teacher rubric experience	Interactions, controlling for teacher age	Interactions, controlling for subject taught
Jake	-0.15 (0.30)		-0.76 (0.66)	-0.76 (0.66)	-0.66 (0.67)	-0.75 (0.68)
Profile Type						
Basic (name only)		0.55 (0.41)	0.19 (0.55)	0.19 (0.55)	0.25 (0.56)	0.23 (0.58)
Positive		1.25 (0.40)**	0.95 (0.52)+	0.95 (0.52)+	0.95 (0.52)+	1.09 (0.55)*
Negative (ref)		--	--	--	--	--
Jake*Basic Profile			0.83 (0.85)	0.84 (0.85)	0.74 (0.86)	0.79 (0.87)
Jake*Positive Profile			0.71 (0.82)	0.74 (0.82)	0.65 (0.83)	0.73 (0.84)
Prior rubric experience				0.10 (0.32)		
Age						
18-25 Years					0.89 (1.12)	
26-35 Years					0.78 (1.11)	
36-50 Years					0.70 (1.14)	
51+ Years (ref)					--	
Subject taught						
ELA						-1.06 (0.66)
Math						-0.62 (0.70)
Science						-0.49 (0.70)
Social Studies						-2.01 (0.96)*
Other						-1.19 (0.66)+
Observations	219	219	219	219	219	219
Log Likelihood	-129.43	-123.88	-123.33	-123.28	-122.92	-119.33

*Standard Errors in Parenthesis + p<.10. *p<.05, **p<.01 ***p<.001; "ref" is omitted reference category.*

The tone of the profile information again impacted the sophistication criterion (Table 8). Model 2 of Table 8 shows that the positive tone profile received higher scores on the sophistication criterion at a statistically significant level, with a *p*-value of less than .01 and a non-significant effect for basic profiles. In successive models, the effect of positive profile type on sophistication score is marginally significant (models 3 through 5) and significant at the $p<.05$ level (model 6). The ordered logistic regression coefficients for Models 4, 5, and 6 show that teaching *social studies* had a statistically significant effect on the score among control variables, and "other subjects" had a marginally significant effect, with $p<.10$.

Interpretation of the coefficients shown in Table 8 indicates that teachers are more likely to give higher scores on this criterion when they are informed, through the profile information, that they are grading the essay of a student who had previously performed well on the essay. Again, the coefficients for race reflect no relevance of the implied race of the student on this criterion score. These results align with the effects of a student profile and scoring on grading, though they did not show racial bias.

Qualitative Results

Feedback and grading are deeply intertwined and, thus, should be studied together. In other words, it would be hard to study one without the other as grades are a form of feedback and vice versa. Feedback is critical in grading because it often can replace a grade entirely, as seen in certain standards-based grading models (Guskey 2019; Koenka, et al., 2021; Scarlett, 2018).

To code the results, I looked at the more salient properties of the feedback phrase. Even though some of the comments could fit in several categories I opted to place them according to their more salient aspects. For example a comment could have implied growth, but did not directly state it but instead talked about the steps or specific aspect writing. Therefore I place it in

in external stable instead of internal unstable. Or if a small reference was made to one of the attributes, but a larger part of the feedback referenced another, I chose the larger amount to guide my coding. I acknowledge this process is subjective however the process of coding in general is subjective (Saldana, 2021).

Overall, teacher feedback referenced malleable ability, citing that students can change their writing ability with practice and effort (about 102 participants (47%) of the respondents gave feedback concerning of malleable ability; internal, unstable attribution). 47 participants (21%) of the respondents provided feedback aligned with the standard's rigor and difficulty (external, stable attribution). These teachers gave feedback about what the student could do to meet the targeted mastery and writing criteria, not necessarily discussing growth. The other totals for the other two types external unstable (occasion circumstances) and internal stable (fixed or natural ability) were (15%) and (17%) respectively.

Table 9

Feedback Results and Weiner's Attribution Model (1974)

	Stable, Internal Refers to Static Characteristics and Inherent Ability "You are a natural writer"	**Stable, External** Refers to the standard of writing "There is a standard for writing to achieve"	**Unstable, Internal** Refers to Malleable Ability "You can be a good writer"	**Unstable, External** Refers primarily to external factors out of the student's control "It wasn't your day"
	Weiner: (ability)	Weiner: (rigor and difficulty of task)	Weiner: (effort)	Weiner: (luck)
Overall (219)	17% (37)	*21% (47)*	**47% (102)**	15% (33)
All Jake Profiles (110)	14% (15)	*22% (24)*	**45% (49)**	20% (22)
All Jamal Profiles (109)	20% (22)	21% (23)	**49% (53)**	10% (11)
Name only (both) (74)	14% (10)	*22% (16)*	**51% (38)**	14% (10)
Name only (Jamal) (36)	14% (5)	*22% (8)*	**56% (20)**	8% (3)
Name only (Jake) (38)	13% (5)	*21% (8)*	**47% (18)**	18% (7)
Positive profile (both) (73)	*25% (18)*	19% (14)	**40% (29)**	16% (12)
Positive profile (Jamal) (36)	*31% (11)*	11% (4)	**47% (17)**	11% (4)
Positive profile (Jake) (37)	19% (7)	*27% (10)*	**32% (12)**	22% (8)
Negative profile (both) (72)	13% (9)	*24% (17)*	**49% (35)**	15% (11)
Negative profile (Jamal) (37)	16% (6)	*30% (11)*	**43% (16)**	11% (4)
Negative profile (Jake) (35)	9% (3)	17% (6)	**54% (19)**	*20% (7)*

Table 9 shows the quantified qualitative results for teacher attribution of the student's success or failure. Regarding the name-only profile, regardless of whether it was Jamal or Jake, Table 9 shows 51% of the respondents referred to the ability of the student to improve their writing skills between this essay and the next. As with the overall breakdown, we saw the same order, with the next highest being external stable (21%), followed by internal stable and external unstable at 22% and 14%, respectively. Based on the results above, it can be said that about half the teachers seem to have attributed the student's performance to an internal unstable phenomenon, indicating their belief that the student can grow and change into a better writer in the next essay.

Feedback Responses and Attribution Error

The following section reviews the feedback in the different profiles regarding attribution, starting with the most common category of internal unstable and concluding with the less-commonly reported external unstable. Before I present the feedback responses, I briefly revisit the description of each attribute. Internal unstable refers to within-person, mutable traits, external stable refers to outside-of-person, immutable traits, internal stable refers to within-person, immutable traits, and external unstable refers to outside-of-person, mutable traits. Examples of each are shown below:

Internal Unstable – Malleable Ability and Growth. Much feedback from teachers, regardless of the name or profile received, centers around their idea of malleable ability and growth, communicating to students that they can grow their skills and writing with practice, time, and effort. This makes sense since most educators are trained to develop students, not just evaluate them. The following feedback excerpts represent this attribution:

> "Why are you afraid about this? We are here to learn and after all these essays are not majority part of the grade. You should put more though (sic) and time on your next essay, if you want to score better."

This statement appears to promote the idea that with more time and thought, the student can learn to write well – signaling the belief that skills are malleable.

> "I would ask him to further develop his ability to argue his point of view and to try to read some essays on a topic that he might like."

Again, this statement implies that skills can be developed and offers the student the idea of reading other essays on topics of interest – suggesting a way to grow as a writer. A final example suggests ways to grow as a writer, research more comprehensively, and learn about the topic from multiple angles:

> "You should do wider research on the subject. Learning from multiple angles can give you a better understanding of it."

The language used in these feedback examples contains elements that refer to the potential for growth in the ability to write and demonstrates to the students that ability is not fixed. We see such examples throughout the study's feedback portion. These attributions comprised 47% (102 of the total 219 feedback responses).

External Stable - There is a Specific Standard for Writing. Regarding the attribution of stable external student ability, the standard of writing quality was explicitly referenced in the feedback. See the following example:

> "More specificity, more elaboration, more context all needed. The comparison of historical periods is not effective (sic) without development and anchoring to the topic. Thesis is not clearly stated nor is it supported."

In this example, only specific criteria of the standard of writing are referenced. There is no reference to growth in writing. It appears to be a checklist of items to complete to meet the writing standard.

> "The writing did not show significant insight. Where examples of people were given, there was not enough investigation into who the people were to make a strong enough argument."

Similarly, this example speaks from a deficient point of view, suggesting these items were not met concerning a standard.

These attribution examples include language referencing to the standard and criteria of writing - with little direct mention of *growth* in writing ability. These attributions comprised 21% (47 of the total 219 feedback responses).

Internal Stable - Fixed Ability. Concerning stable internal attribution, when the feedback contained information that the student should not worry because they are good writers or perhaps that it will work out due to their innate writing skill. Examples of this attribution are below:

> "You should feel relaxed, your last essay was ok and I'm sure you will succeed."
>
> "Sometimes is (sic) better to get worse (sic) score. That means you can learn and get better next time. There is not a reason (sic) to worry about the next essay."

These two examples include assurances (e.g., should feel relaxed) or confident predictions (that the student *will* succeed in the next essay). Further, by referencing past and current performances the teacher seems to subscribe to the idea that this student is a natural writer.

> "Trust in yourself and study accordingly."
>
> "It, (sic) is absolutely fine to be nervous. Just do your best, and that will be all."

These last two examples are other ways the participating teachers signaled a fixed ability attribution in their feedback. Generalized references to trusting oneself or doing one's best were usual with this type of attribution.

This feedback communicates to the student that they *need not worry because they are a natural writer* or *that their ability is so deeply rooted that they imply that there might not be any more growing to do.* In other words, they believe the student's success or failure on this essay to be of minimal importance since writing, according to them, writing may be more an innate ability. These attributions comprised 17% (37 of the total 219 feedback responses).

External Unstable - Luck or Situational Factors. Finally, with unstable external attribution regarding situational factors, luck or the context of the moment reassured the students that they had nothing to worry about since it was an external factor that likely contributed to their scores. Some examples of this attribution include:

> "You have nothing to be nervous about. Use the previous scores a a (sic) guide for the next one, (sic) now you know where you need to focus."

These two examples reference a prior successful performance and suggest that this essay is an outlier and need not factor into any judgment about their writing ability. The following example takes a broader swipe at this type of attribution:

> “To (sic) not be nervous and that grades generally do not matter in life."

These attributions suggested external forces might contribute to a person’s performance. These examples were 15% (33 of the total 219 feedback responses).

Summary

The quantitative findings suggest that teachers may be vulnerable to expectancy bias or shifting standards (Dobelli, 2013) since the participants who received a positive tone profile gave

higher scores than those who received a negative tone profile. Additionally, no significant racial bias was found when grading the essay. However, this does not mean racial bias does not exist in standards-based grading; thus, future research is necessary (see Chapter 5).

Regarding the qualitative component of the study, half the participating teachers gave feedback that attributed student ability to internal, unstable factors (malleable skill), with the other half attributing the students' performances to the other three factors (fixed ability, external fixed standard, or occasion circumstances). In every instance (type of profile), most feedback seemed to attribute performance success or failure to internal unstable reasons, meaning that if the student practiced more, they could improve their skill. This attribution is more common in the results and makes intuitive sense since teaching aims to develop students' minds and skills.

CHAPTER 5

DISCUSSION AND RECOMMENDATIONS

As discussed in Chapter 1, it is a common perception that standards-based grading models are more equitable and fairer due to the absence of mechanical performance calculations such as averaging, weighting, and grading scales (Knight & Cooper, 2019; Townsley & Wear, 2020; Veenstra, 2021). This study investigated this claim by testing how biases, such as racial prejudice, shifted mastery standards, expectancy theory, and attribution error can distort teacher judgments of student performance. This study explored the curiosities: 1) To what extent teachers are influenced by their cognitive biases when using standards-based rubrics to evaluate student performance? As schools consider shifting from conventional grading practices (points and averaging) to more standards-based models, a close examination of how teachers use their judgment to interpret students' performance is crucial. Biases that were studied in this experiment were shifting standards and expectancy theory (Rosenthal & Jacobson, 1968; Rubie-Davies, 2018), heuristics (Tversky & Kahneman, 1974; Kahneman et al., 2021), and attribution error (Weiner, 1974).

Summary of Findings

This study found that Jamal's and Jake's essays were interpreted and graded similarly, suggesting the absence of any significant racial bias in the evaluation of the essay. This finding differs from the results of many audit studies, which found racial bias triggered by names (Gaddis, 2017a; Ghoshal, 2018; Neumark, 2012). In educational research on racial bias, some studies have found limited or inconclusive effects on student grades (Quinn, 2020), while several others suggest its presence (Jacoby-Senghor, 2016; Bergh et al., 2010).

As such, the findings do not suggest racial bias's effect on the evaluation of student performance. However, they reflect an often significant, causal relationship between the profile information (positive tone profile vs. name only vs. negative tone profile) and the mastery and criteria scores assigned to essays by teachers. If a participant received a positive tone profile, they tended to assign higher scores to the essay than if they received a negative tone profile. When controlling for rubric experience, subject taught, and years of teaching experience, the results showed a similar effect of profile tone on rubric scores. This dissertation's findings suggest that teachers' decision-making in standards-based practices may be more vulnerable to many non-academic factors than is currently recognized by advocates of standards-based grading.

The findings were consistent with teacher expectancy studies (Rubie-Davies, 2018; Schuster et al., 2021). These studies show that teacher judgment can be distorted depending on an expectation a teacher might create of the student before grading their work (the profile's tone appears to influence the teacher's perception of the student before the event). For example, López-Pastor (2017) found that teachers often reward a historically higher performer with a benefit of the doubt in their grading and feedback and, in contrast, grade a student's work more critically if they have the impression that the student's previous performance was not good. In sum, just like anybody, teachers appear vulnerable to cognitive biases and thought distortions; standards-based rubrics do not mitigate them, and one might believe.

Implications and Recommendations

The findings suggest that standards-based grading may be influenced by the teacher's expectations of their students and its related distortions (excessive coherence and haloing), which is of specific importance. Students from whom the teacher had a high expectation (positive

profile) received higher grades, while those from whom the teacher had a lower expectation (negative profile) received lower grades. Further, the findings showed that about half of the teachers did not provide feedback related to malleable ability; instead, they suggested that the writing performance was either an outlier (citing stable internal attribution) or a result of external factors unrelated to the student's ability.

Training and professional development can help create awareness about the intersection of pedagogy and cognitive bias. However, more than training is needed to change professionals' practices (e.g., Kalev et al., 2006). Moreover, "implicit bias training has had some success in changing individual-level beliefs and actions, but meta-analyses (Bezrukova, et al. 2016) suggest it is largely ineffective in diminishing institutional inequities" (Pritlove et al. 2019, p. 504).

It is not suggested that teachers refrain from implementing standards-based grading models. Instead, educators should consider professional development on cognitive bias in terms of efficacy of practice and accountability. The following are several professional development recommendations for teachers working in standards-based grading systems or educators who are thinking about changing to it: 1) increasing interrater reliability; 2) interweaving grading rules with professional judgment; 3) narrative grade book formats; 4) attribute success or failure appropriately through feedback; 5) refining rubric evaluation of student performance, and 6) including students in the grading process.

Collaborative Grading of Student Work (Interrater reliability)

First, training evaluators (in this case, teachers) to implement system-level interrater reliability checks to ensure that their scoring practices align with other observers should be a priority (Bell et al., 2014). This training could focus on several central aspects of interrater reliability.

One aspect of interrater reliability is collaboratively grading evidence. For example, moving to a grading system in which various pieces of evidence are collected over time and reviewed with one's colleagues may help guard against cognitive bias or occasional noise. Mainly because of the frequency of review and the necessity of multiple collegial perspectives to determine the level of student learning, as suggested in Oakleaf's (2009) study on the interrater reliability in assessing information literacy.

Another aspect is collaboratively vetting criteria of mastery (e.g., thesis, sophistication, evidence): The more specific criteria teachers and their teams establish, the less likely is the scope of variability in scoring, and the less the teachers are vulnerable to biases since the teachers' team calibrates what mastery looks like (Andrade et al., 2020; Gawande, 2010). This recommendation aligns with Kahneman et al.'s (2021) insight that people can improve judgment by "breaking down the judgment problem into a series of smaller tasks" (p. 372). This process helps to decompose judgment into not only content criteria but also decision-making criteria. Professional development focused on team scoring criteria, proficiency discussions, simulations of rubric use, protocols for reviewing student work, and professional development on quality feedback may help to achieve this.

Similar to the collaborative grading of student work, teachers can conduct a noise audit: A noise audit (Kahneman et al., 2021) is where multiple individuals read and grade the same essay, discuss the mastery level and associated criteria, and finally, discuss the rationale for the score they assigned to the essay. If the variability in their scoring comes from the teacher's inability to judge student work effectively, then priority should be given to improving those skills. School leaders can support teachers to become more aware of the non-academic factors

that may influence their grading, which can minimize the chances of a student developing an inaccurate picture of self.

One other aspect of interrater reliability is self-reflection on biases. Teachers can be taught how to make decisions and interpret student data while becoming aware of any biases that student work might trigger or preconceptions (Cohen & Steele, 2002) the teacher may have before student performance evaluation (Samuels, 2013). “Decision-making hygiene” training (Kahneman, 2021, p. 9), achieved by conducting decision audits, should be a norm in teacher preparation.

Addressing interrater reliability in a school or district can help teachers deepen their awareness of the various factors used to evaluate student work and develop shared understandings of mastery. Considering the perspectives of other colleagues when grading may allow educators to acknowledge the multiple perspectives that student work can be seen through. Ideally, school leaders would want the same student work graded in the same way to arrive at the same grade, using the same rubric. This ideal alignment is only sometimes found to be the case, and continued training is needed in this area.

Interweave Simple Grading Rules to Supplement Professional Judgment of Student Work

Before teachers and leaders decide to change the grading practices to a standards-based approach, considering how teachers’ decisions can be affected by cognitive biases is crucial. By creating district-level policies and procedures on how teachers make decisions about the grades they award to students, schools can design institutional policies that avoid reliance on the mechanical process of grading (points and formulas) and instead embrace the more clinical aspects of judging student work (conversations, feedback, and review mastery evidence). To improve accuracy, we can model how we make judgments by forming simple rules and

implementing those rules in our grading process (Kahneman et al., 2021) that supplement the professional interpretation of student performance. An example of a standards-based grading policy that combines simple rules with professional judgment, as adapted from Adlai E. Stevenson High School District 125, Lincolnshire, IL, is shown below:

1) Rule 1: Compute an overall modal proficiency score from all the proficiency scores for each skill/standard for each student.
2) Rule 2: Compare recent proficiency scores in each skill/standard to its overall modal proficiency score and revise the overall proficiency score of each skill/standard.
3) Rule 3: Determine a final overall proficiency score based on Rules 1 and 2.
4) Rule 4: Calculate a final course grade (if necessary) by using the following algorithm:
    a) If all skills/standards have an overall proficiency score of proficiency or higher, then the student receives an A.
    b) If all skills/standards have an overall proficiency score of proficiency or higher, but one has an overall score of approaching proficiency, then the student receives a B.
    c) If all skills/standards have an overall proficiency score of proficiency or higher, but multiple skills/standards have an overall score of approaching proficiency, then the student receives a C.
    d) If one skill/standard has an overall proficiency score of failing to develop regardless of the remaining skills/standards scores, then the student receives a D.
    e) If multiple skills/standards have an overall proficiency score of failing to develop regardless of the remaining skills/standards scores, then the student receives an F.

In the example above Rule 4 is a simple rule, meaning that once the teacher has judged the mastery level of the standards associated with students' work.

Narrative Grade Book Formats

Teachers often give pause to the idea of professional development about grading in general. However, with training on bias and grading, teachers voice stronger convictions, perspectives, and feelings. This reality should allow school leaders to hold such training sessions, as they are critical.

To help lessen the likelihood of implicit bias in grading processes, teachers can use grade books to communicate stories of student learning instead of mathematical averaging of earned points (Reibel & Thede, 2020). An example of a process of formatting a grade book format that communicates students' stories of learning, inspires conversations, promotes inquiry, and may lessen bias is shown below:

Learning Story #1: How is the student growing? This story describes each student's progress toward proficiency in each target and standard. The sections communicate information about the students' progress—the rate at which they are "moving toward mastery" in course skills or knowledge standards. The students' weekly growth scores indicate that they are developing proficiency and growing as expected. Conveying a growth story in your grade book creates a sense of the students might be in their learning and where they might be going. A teacher can communicate the trajectory of student learning as positive or negative.

Learning Story #2: How is the student performing? This story describes each student's current level of proficiency in each standard and target. This area of the grade book communicates mastery of standards. Competency discussions between students and teachers about the student's body of work in course skills competency can help plant the seeds of a better

relationship. Exceeds: The students submitted evidence of learning indicating they have a skill proficiency level that exceeds the standard. Meets: The students presented evidence of learning indicating that they have a skill proficiency level that meets the standard. Approaching: The students submitted evidence of learning indicating they have an undeveloped yet emerging proficiency level that needs to meet the standard. Developing: The students' learning evidence suggests they are developing foundational knowledge and prerequisite skills.

Learning Story #3: How is the student behaving? This story describes how a student's learning habits and behaviors are of concern or well done. This grade book section offers insight into how a student behaves in the classroom. It is intentionally separated from the other sections not to cloud the interpretation of the student's overall competency in class standards. Teachers should not communicate compliance behavior with competence evidence; they are two different stories. If you place them together in the grade book, it may be hard to understand if the student is getting a particular grade because their behavior is good or if they are earning the grade because they are competent (Guskey 2014).

Learning Story #4: How is the student preparing? This story describes how a student's ability to build foundational skills and knowledge undergirds the successful performance of a standard. It is also vital to communicate the level to which a student is prepared to engage in a learning experience. To do this, you can use preparation scores. Teachers use these scores to communicate how well-prepared the student is. Teachers can use these codes to report the integrity of the homework preparation, checks for understanding, or other instructional activities. They are unique codes, as they do not communicate the status of a student's competency or their compliant behavior. Instead, they report the level to which students are building the foundational skills and knowledge needed to demonstrate mastery of the skills.

Attributing Success or Failure Appropriately through Feedback

Professional development on the subtleties of attribution in teacher feedback is valuable time spent with one's staff since feedback can often change the trajectory of a student's learning (Yeager et al., 2014). Feedback can become biased when the communication between student and teacher results in misperception, misapplication of skills or knowledge, or even non-learning, leading to the absence of learning growth or negative growth. The following questions, adapted from the work of decision science and behavioral economics expert Paul J. H. Schoemaker (2011), outline how teachers can ensure high-quality feedback.

Does the feedback contain assumptions? Do you have any preconceptions about the student that may cause you to *assume* why they got an answer right or wrong? Prejudices make the feedback to be unreliable.

Does the feedback have a context? Is your feedback based on an unfamiliar or unrelated context, or is it in line with what students expect to focus on?

Are the students ready? Do your lessons allow students to prepare to receive feedback? Best teaching practices clarify expectations early, provide formative feedback during the learning process, and create a growth-minded learning environment where students can give and receive supportive feedback.

What data does the feedback rely on? Does your feedback rely on the most readily available information or a complete data set of student performance? Partial or incorrect data can cause untrustworthy feedback.

Is the feedback clear? Do you use language that makes it clear how you want students to grow in learning, or do you use only deficiency language? In other words, do your expectations

frame growth potential by stating what proficient work *does* look like, or do your expectations state what proficient work *does not* look like?

Is the feedback balanced? Does your feedback evaluate both the performance of the skill and the quality of the learning?

Is the feedback relevant? Does the feedback relate to the learning target? Does it contain the aspects of proficiency from the learning target?

Any of these questions can guide a professional development session about minimizing the effects of bias in teacher feedback. As far as the outcomes from such training are concerned, we should see shifts in the language of teacher feedback. Two examples are listed below (Reibel, 2022):

Before Training

1. "When I write, I try and think about [detail]. Remember when I taught you the three-step process? No? The one that is in your textbook? That is the most effective process,"

After Training

2. "What did you think about when you wrote [this]? Seems like that interests you. I can see you are passionate about [that]. What are the first three things you did when you wrote [this]? That is an interesting place to start. Could I convince you to start here? No? Okay, that makes sense. You have to make this work for you."

Before Training

1. "You did not include [these details] about [person] in your essay. Try [these words],"

After Training

2. "Tell me more about these words [here]. I am interested to know why you think [this word] did not work instead. Oh, okay, that would work. You should add what you just said to your paragraph, and it was perfect."

Training teachers to provide constructive feedback may help mitigate the effects of cognitive biases in providing feedback on students' work.

Refining the Rubric Evaluation of Student Performance

As schools, including higher education institutes, adopted a more competency-based approach to learning (López-Pastor, 2012), have increased their use of rubrics, with "rubrics being one of the most innovative instructions to obtaining evidence regarding the acquisition of competencies" (Velasco-Martinez & Tojar-Hurtado, p.119). Further, educational institutions continue to adopt "as a means of mitigating racial bias scholars have proposed that teachers use rubrics delineating clear performance criteria." (Quinn, 2020, p.3). See studies by Malouff & Thorsteinsson (2016) and Payne & Vuletich (2018). With this trend comes the need for continued training on the design and use of rubrics to evaluate student work. Two types of rubric training that may prove useful in mitigating cognitive bias are 1) rubric design; and 2) rubric pedagogy (use of rubrics).

Rubric Design. Often rubrics are designed to evaluate students' mechanical application of knowledge and rote skill development practices to gain points or technical feedback. As it relates to this study, when participants treat writing as a process involving specific steps, they may reduce students' writing feedback to transactions, leaving them to perceive that teachers value their writing only when they mimic the prescribed writing procedures. To help minimize this design and banal use of rubrics, authors Jones et al. 2016 suggest establishing the following five-step pedagogy to improve rubric efficacy; these guidelines include "(1) deconstruction of

the rubrics and standardizing the marking method; (2) examples and exemplars; (3) peer review; (4) self-review; (5) reflective diary" (as cited in Velasco-Martinez & Tojar-Hurtado, 2018, p.119).

Rubric Pedagogy: Gathering First-Person Accounts with Rubrics. Teachers can use rubrics to gather their students' first-person accounts or reflections about their learning to understand better what is happening with students' mastery or knowledge. If there is more to the story, rubrics that activate first-person accounts of learning can help uncover it. Training how to use rubrics to gather first-person accounts of learning from their students is vital to minimize bias. Some ideas for how teachers can collect first-person accounts from rubrics include the following: 1) include segments that ask students to comment on how their current life experiences are affecting and shaping their learning and understanding; 2) encourage open dialogue using rubrics to ask students to express their thoughts, feelings, learning, progress, and growth, both in their lives and academic performance before or after an assessment; and 3) offer in-moment reflections about the assessment or task. Having students reflect on their learning during an assessment can give them valuable time to engage with their internal voice and give a teacher more information about their learning. Through rubrics, teachers can gain insights into how a student sees the world to adjust their instruction to meet students' needs better and potentially minimize biased grading or feedback.

Include Students in the Grading Process

One recommendation that does not necessarily stem from the findings but could complement efforts to reduce teacher bias is increasing interaction with students during the grading process. Grading is often a unidirectional action, from teacher to student. This singular direction can

increase the chances of one-sided interpretations of student work (biased interpretation) (Samuels, 2013).

Teachers can mitigate bias by including student voices in the grading process; in other words, they seek their perspectives and thoughts, which can lessen the chance for biased interpretation of student work. (Ross, 1994). In other words, *co-constructing* grades with students (Kilgour, 2020) may minimize the activation of teachers' subconscious biases; since the grading process is now a collaborative effort instead of a unidirectional process often seen in conventional grading where the teacher is the grade giver. Grading experts, such as Thomas Guskey (2014) and Tom Schimmer et al. (2018), continue to discuss ways to include students in the grading process through self-monitoring and self-evaluation of their work. One such method to involve students in the grading process (Reibel, 2022) includes: aspects of self-evaluation, self-appraisal assessments, and on assessments self-expression on assessments. The first aspects ask students to grade themselves on the mastery scale or rubric and use the associated criteria to justify their self-ratings (White, 2020). The second aspects ask students to comment on their states of thinking or emotion (White, 2020). The third aspects ask students to freely respond about their performance or even their lives outside of school (Emdin, 2021).

Recommendations for Future Studies

Future research investigating the impact of cognitive bias and standards-based grading practices is needed. Although this study did not show any racial bias, these findings cannot be generalized to all educational settings, as one cannot ignore the fact that “race and racism exist in society; they are also present and prevalent in education and in the research and practice of education” (Milner, 2007, p. 391). The lack of differences observed in mastery scores between Jake and Jamal may have resulted from the non-activation of prejudicial bias, not because there

were none. Thus, future research can consider exploring this aspect in much more detail and a much larger context, for instance, by examining various names to clarify whether a name-only profile triggers racial bias.

Future research could consider replicating this study's design with a larger sample to assess its validity and generalizability better. It would be interesting to conduct this study in a school or district that has conducted professional development on inter-rater reliability and bias in grading, for example. Future replications could also compare other factors such as gender identification, dispositions, and performance in other subject areas, allowing a more precise understanding of how teacher judgments might differ over time (pattern variability). Subsequent studies could also examine races besides Black and white, with attention to whether the effects vary by teachers' perceptions of students' socio-economic background or learning supports.

Last, it is crucial to continue analyzing the feedback provided by teachers to students, specifically the reasons they attribute to a student's success or failure. Similar to the Gentrup et al. (2020) study cited in Chapter 2, future researchers might observe teachers during instruction as they provide feedback to students about their performance and code the types of language or behavior for evidence of other biases. Perhaps a larger sample size of respondents could better uncover the interplay between feedback, mastery scores awarded, and the presence of cognitive bias.

Conclusion

According to Kahneman et al. (2002, 2021), psychological biases are universal and produce variability in judgments because of individual differences in behaviors, experiences, values, and backgrounds. Their studies continue to find that "anything that reduces psychological biases can improve judgment" (Kahneman et al., 2021, p. 175). This study explored the extent to

which internalized cognitive biases influence teachers' evaluations of student work, specifically in a standards-based grading system. The participants in this study were given the same essay, mastery standard, and writing criteria. Yet, they interpreted the essay's quality differently depending on the student's profile (positive tone, name-only, or negative tone). The findings shed light on cognitive biases affecting teachers' judgment. These phenomena are critical to investigate as standards-based grading is being accepted in more and more settings (Guskey et al., 2019; Heflebower et al., 2019) since teachers in standards-based grading systems rely on their professional judgment to score and grade students. It is our liability to identify these cognitive biases and monitor them if we intend to achieve the goal of grading equity.

REFERENCES

Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 350-372.

Babad, E. Y., Inbar, J., & Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of educational psychology*, *74*(4), 459.

Bacchus, R., Colvin, E., Knight, E. B., & Ritter, L. (2020). When rubrics aren't enough: Exploring exemplars and student rubric co-construction. Journal of Curriculum and Pedagogy, 17(1), 48-61.

Bandura, A. (1997). Self-efficacy: The exercise of control. W H Freeman/Times Books/ Henry Holt & Co.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist*, *54*(7), 462.

Beckman, L., & Rodriguez, N. (2021). Race, Ethnicity, and Official Perceptions in the Juvenile Justice System: Extending the Role of Negative Attributional Stereotypes. *Criminal Justice and Behavior*, 00938548211004672.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in evaluator thinking. In T. Kane, K. Kerr, & R. C. Pianta (Eds.), Designing teacher evaluation systems: New guidance from the measures of effective teaching project (pp. 50–97). John Wiley. https://doi.org/10.1002/9781119210856.ch3

Berns, A. (2020, February). Scored out of 10: Experiences with binary grading across the curriculum. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education (pp. 1152-1157).

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991-1013.

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. American Educational Research Journal, 52(5), 861-893.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485.

Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment.

Biernat, M. (2012). *Standards and expectancies: Contrast and assimilation in judgments of self and others*. Psychology Press.

Bodenhausen, G. V., & Richeson, J. A. (2010). Prejudice, stereotyping, and discrimination.

Brault, M. C., Janosz, M., & Archambault, I. (2014). Effects of school composition and school climate on teacher expectations of students: A multilevel analysis. Teaching and Teacher Education, 44, 148-159.

Brimi, H. M. (2011). Reliability of grading high school work in English. Practical Assessment, Research, and Evaluation, 16(1), 17.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. Educational Review, 67(3), 343-368.

Brophy, J. E., & Good, T. L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of educational psychology*, *61*(5), 365.

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. Journal of educational psychology, 75(5), 631.

Brophy, J. E. (1985). Teacher-student interaction. In J. B. Dusek (Ed.), Teacher expectancies (pp. 303–328). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Buckmiller, T., Peters, R., & Kruse, J. (2017). Questioning points and percentages: Standards-based grading (SBG) in higher education. *College Teaching*, *65*(4), 151-157.

Buckmiller, T., Townsley, M., & Cooper, R. (2020). Rural high school principals and the challenge of standards-based grading. *Theory & Practice in Rural Education*, *10*(1), 92-102.

Bublitz, C. (2020). What Is Wrong with Hungry Judges? A Case Study of Legal Implications of Cognitive Science. Law, Science and Rationality (Maastricht Law Series). The Hague: Eleven, 1-30.

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*, *55*(3), 463-477.

Butler, D. M., & Crabtree, C. (2017). Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. *Journal of Experimental Political Science*, *4*(1), 57-67.

Butler, D. M., & Homola, J. (2017). An empirical justification for the use of racially distinctive names to signal race in experiments. Political Analysis, 25(1), 122-130.

Carson, J. E. (2019). External relational attributions: Attributing cause to others' relationships. *Journal of Organizational Behavior*, *40*(5), 541–553. https://doi.org/10.1002/job.2360

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology*, *41*(3), 131.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies*. Sage Publications, Inc.

Cohen, G. L., & Steele, C. M. (2002). A barrier of mistrust: How negative stereotypes affect cross-race mentoring. In *Improving academic achievement* (pp. 303-327). Academic Press.

Conaway, W., & Bethune, S. (2015). Implicit Bias and First Name Stereotypes: What Are the Implications for Online Instruction?. Online Learning, 19(3), 162-178.

Daneshzadeh, A., & Sirrakos, G. (2018). Restorative justice as a double-edged sword: Conflating restoration of black youth with the transformation of schools. *Taboo: The Journal of Culture and Education*, *17*(4), 2.

Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2016). Race and gender effects on employer interest in job applicants: new evidence from a resume field experiment. *Applied Economics Letters*, *23*(12), 853-856.

De Boer, H., Bosker, R. J., & van der Werf, M. P. (2010). Sustainability of teacher expectation bias effects on long-term student performance. Journal of Educational Psychology, 102(1), 168.

De Boer, H., Timmermans, A. C., & Van Der Werf, M. P. (2018). The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis. *Educational Research and Evaluation*, *24*(3-5), 180-200.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347.

Dee, T. S. (2004). The race connection: Are teachers more effective with students who share their ethnicity?. Education Next, 4(2), 52-60.

Dobelli, R. (2013). The art of thinking clearly: better thinking, better decisions. Hachette UK.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. Journal of Forensic Identification, 56(4), 600.

Dror, I. E., Scherr, K. C., Mohammed, L. A., MacLean, C. L., & Cunningham, L. (2021). Biasability and reliability of expert forensic document examiners. *Forensic science international*, *318*, 110610.

Dweck, C. S. (2018). Reflections on the legacy of attribution theory.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. Psychological Assessment, 8(4), 360–362.

Emdin, C. (2021). *Ratchetdemic: Reimagining academic success*. Beacon Press.

Erickson, J. A. (2011). A Call to Action: Transforming Grading Practices. *Principal Leadership*, *11*(6), 42-46.

Feldman, J. (2018). *Grading for equity: What it is, why it matters, and how it can transform schools and classrooms*. Corwin Press.

Feldman, J. (2019). Beyond standards-based grading: Why equity must be part of grading reform. *Phi Delta Kappan*, *100*(8), 52-55.

Ferman, B., & Fontes, L. F. (2021). Discriminating Behavior: Evidence of Teachers' Grading Bias. Available at SSRN 3797725.

Fischhoff, B., Kahneman, D., Slovic, P., & Tversky, A. (2002). For those condemned to study the past: Heuristics and biases in hindsight. *Foundations of cognitive psychology: Core readings*, 621-636.

Fish, R. E. (2017). The racialized construction of exceptionality: Experimental evidence of race/ethnicity effects on teachers' interventions. *Social Science Research*, *62*, 317-334.

Fisk, S. P., & Taylor, S. E. (1984). Social cognition: Topics in social psychology.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination.

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences*, *3*(April), 45–50.

Forgas, J. P. (2011). She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation. European Journal of Social Psychology, 41(7), 812-817.

Fox, L. (2015). Seeing potential: The effects of student–teacher demographic congruence on teacher expectations and recommendations. *AERA open*, *2*(1), 2332858415623758.

Frankel, M. E. (1973). Criminal sentences: Law without order.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. Social Forces, 93(4), 1451-1479.

Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. The ANNALS of the American Academy of Political and Social Science, 660(1), 282-299.

Gaddis, S. M., (2017a). "How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–89. https: //doi.org/10.15195/v4.a19.

Gardner, W. B., Karam, E., Tribble, L., & Cogliser, C. (2019). The missing link? Implications of internal, external, and relational attribution combi- nations for leader–member exchange, relationship work, self-work, and conflict. *Journal of Organizational Behavior*, *40*(5), 554–569. https://doi. org/10.1002/job.2349

Gawande, A. (2010). *Checklist manifesto, the (HB)*. Penguin Books India.

Gawronski, B., & Hahn, A. (2019). Implicit measures: Procedures, use, and interpretation.

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback, and student achievement. *Learning and Instruction*, *66*, 101296.

Ghoshal, R. (2018). Testing for Discrimination: Teaching Audit Studies in Quantitative Methods Courses. *Teaching Sociology*, *46*(4), 309–323. https://www.jstor.org/stable/26589047

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Gobble, T., Onuscheck, M., Reibel, A. R., & Twadell, E. (2017). *Pathways to proficiency: Implementing evidence-based grading*. Bloomington, IN: Solution Tree Press.

Graham, S. (2016). Attribution theory and motivation in school. In Handbook of motivation at school (pp. 23-45). Routledge.

Gravetter, F. J., Wallnau, L. B., Forzano, L. A. B., & Witnauer, J. E. (2020). Essentials of statistics for the behavioral sciences. Cengage Learning.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically minor effects of the Implicit Association Test can have societally large effects.

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of personality and social psychology*, *79*(6), 1022.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Guskey, T. R.. & Anderman, E. M. (2013). In Search of a Useful Definition of Mastery. Educational Leadership, v71 n4 p18-23 Dec 2013-Jan 2014

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

Guskey, T. R. (2014). On your mark: Challenging the conventions of grading and reporting. solution tree press.

Guskey, T. R. (2019). Grades versus comments: Research on student feedback. *Phi Delta Kappan*, *101*(3), 42-47.

Guskey, T. R., & Brookhart, S. M. (2019). What we know about grading: What works, what doesn't, and what's next. ASCD.

Harvey, P., Madison, K., Martinko, M., Crook, T. R., & Crook, T. A. (2014). Attribution theory in the organizational sciences: The road traveled and the path ahead. Academy of Management Perspectives, 28(2), 128-146.

Hanson, A., & Santas, M. (2014). Field experiment tests for discrimination against Hispanics in the US rental housing market. *Southern Economic Journal*, *81*(1), 135-167.

Hasan, A. A. A. (2022). Effect of Rubric-Based Feedback on the Writing Skills of High School Graders. Journal of Innovation in Educational and Cultural Research, 3(1), 49-58.

Heflebower, T., Hoegh, J. K., Warrick, P. B., & Flygare, J. (2019). *A teacher's guide to standards-based learning*. Marzano Research.

Heider, F. (1958). The naive analysis of action.

Henry, D. T. (2018). Standards-based Grading: The Effect of Common Grading Criteria on Academic Growth (Doctoral dissertation, Bowling Green State University).

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social psychology of education*, *20*(3), 471-490.

Hollyforde, S., Whiddett, S. (2002). *The Motivation Handbook.* Cromwell Press, Trowbridge, Wiltshire, UK.

Hornstra, L., Stroet, K., van Eijden, E., Goudsblom, J., & Roskamp, C. (2018). Teacher expectation effects on need-supportive teaching, student motivation, and engagement: A self-determination perspective. Educational Research and Evaluation, 24(3-5), 324-345.

Iamarino, D. L. (2014). The benefits of standards-based grading: A critical evaluation of modern grading practices. *Current Issues in Education*, *17*(2).

Irizarry, Y. (2015). Utilizing multidimensional measures of race in education research: The case of teacher perceptions. *Sociology of Race and Ethnicity*, *1*(4), 564-583.

Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. Journal of Experimental Social Psychology, 63, 50-55.

Jokić, A. (2017). Evaluation of Writing Assignments. European Journal of Multidisciplinary Studies, 2(7), 214-221.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review, 2(2), 130-144.

Joshi, E., Doan, S., & Springer, M. G. (2018). Student-teacher race congruence: New evidence and insight from Tennessee. *AERA Open*, *4*(4), 2332858418817528

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology*, *57*(3), 469.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. Personality and social psychology review, 9(2), 131-155.

Jussim, L., Robustelli, S. L., & Cain, T. R. (2009). Teacher Expectations and Self-Fulfilling Prophecies. In Handbook of motivation at school (pp. 363-394). Routledge.

Jussim, L., Careem, A., Honeycutt, N., & Stevens, S. T. (2020). Do IAT Scores Explain Racial Inequality?. In Applications of Social Psychology (pp. 312-333). Routledge.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel prize lecture*, *8*(1), 351-401.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Little, Brown.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise. *Harvard Bus Rev*, 38-46.

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. Journal of Educational Psychology, 109(6), 871.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American sociological review*, *71*(4), 589-617.

Kelley, H. H. (1967). Attribution theory in social psychology. In Nebraska symposium on motivation. University of Nebraska Press.

Kilgour, P., Northcote, M., Williams, A., & Kilgour, A. (2020). A plan for the co-construction and collaborative use of rubrics for student learning. *Assessment & Evaluation in Higher Education*, *45*(1), 140-153.

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.

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology*, *41*(7), 922-947.

Kukucka, J., Kassin, S. M., Zapf, P. A., & Dror, I. E. (2017). Cognitive bias and blindness: a global survey of forensic science examiners. *Journal of Applied Research in Memory and Cognition*, *6*(4), 452-459.

Kuklinski, M., & Weinstein, R. (2000). Classroom and grade level differences in the stability of teacher expectations and perceived differential teacher treatment. Learning Environments Research, 3(1), 1-34.

Legette, K. B., Halberstadt, A. G., & Majors, A. T. (2021). Teachers' understanding of racial inequity predicts their perceptions of students' behaviors. *Contemporary Educational Psychology*, *67*, 102014.

Leonardo, Z., & Boas, E. (2021). Other kids' teachers: What children of color learn from White women and what this says about race, whiteness, and gender. In Handbook of critical race theory in education (pp. 153-165). Routledge.

Levitt, S. D., & Dubner, S. J. (2014). *Freakonomics*. B DE BOOKS.

Lewis, A. E., & Diamond, J. B. (2015). *Despite the best intentions: How racial inequality thrives in good schools*. Oxford University Press.

López-Pastor, V. M., Fernández-Balboa, J. M., Santos Pastor, M. L., & Fraile Aranda, A. (2012). Students' self-grading, professor's grading and negotiated final grading at three university programmes: analysis of reliability and grade difference ranges and tendencies. *Assessment & Evaluation in Higher Education*, *37*(4), 453-464.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*(3), 245-256.

Marcelo, A. K., & Yates, T. M. (2019). Young children's ethnic–racial identity moderates the impact of early discrimination experiences on child behavior problems. *Cultural Diversity and Ethnic Minority Psychology*, *25*(2), 253.

Martinko, M. J., & Mackey, J. D. (2019). Attribution theory: An introduction to the special issue. *Journal of Organizational Behavior*, *40*(5), 523-527.

Marshall, H. H., & Weinstein, R. S. (1986). Classroom context of student-perceived differential teacher treatment. Journal of educational psychology, 78(6), 441.

Martinek, T. J. (1980). Stability of teachers' expectations for elementary school aged children. Perceptual and Motor Skills, 51(3_suppl2), 1269-1270.

Maryfield, B. (2018). Implicit racial bias. *Justice Research and Statistics Association*, 1-10.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale.

Marzano, R. J. (2011). Formative assessment & standards-based grading. Solution Tree Press.

Marzano, R. J. & Hardy, P. (2023) Leading a Competency-based Secondary School. Marzano Resources. Bloomington, IN.

McDermott, R. P. (1987). The explanation of minority school failure, again. *Anthropology & Education Quarterly*, *18*(4), 361-364.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of school psychology*, *46*(3), 235-261.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. Teaching and Teacher Education, 65, 48-60.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Milner IV, H. R. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. Educational researcher, 36(7), 388-400.

Muñoz, M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. *Phi Delta Kappan*, *96*(7), 64-68.
Mutz, D. C. (2011). Population-based survey experiments. In *Population-Based Survey Experiments*. Princeton University Press.

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *Journal of Human Resources*, *47*(4), 1128-1157.

Nosek BA, Bar-Anan Y, Sriram N, Axt J, Greenwald AG (2014) Understanding and Using the Brief Implicit Association Test: Recommended Scoring Procedures. PLoS ONE 9(12): e110938. doi:10.1371/journal.pone.0110938

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology*, *60*(5), 969-983.

Oates, G. L. S. C. (2003). Teacher-student racial congruence, teacher perceptions, and test performance. *Social Science Quarterly*, *84*(3), 508-525.

Ogbu, J. U., & Simons, H. D. (1998). Voluntary and involuntary minorities: A cultural-ecological theory of school performance with some implications for education. *Anthropology & education quarterly*, *29*(2), 155-188.

Oysterman, D., Elmore, K., & Smith, G. (2001). Self, Self-Concept, and Identity. Leary, Mark R., Price Tagney, June eds. *Handbook of Self and Identity*.

Pager, D. (2003). The mark of a criminal record. American journal of sociology, 108(5), 937-975.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. Educational Research Review, 30, 100329.

Paslay, C. (2021). Exploring White Fragility: Debating the Effects of Whiteness Studies on America's Schools. Rowman & Littlefield Publishers.

Payne, B. K., & Vuletich, H. A. (2018). Policy insights from advances in implicit bias research. *Policy Insights from the Behavioral and Brain Sciences*, *5*(1), 49-56.

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction*, *42*, 123-140.

Pigott, R., & Cowen, E.,(2000). Teacher Race, Child Race, Racial Congruence, and Teacher Ratings of Children's School Adjustment, Journal of School Psychology, Volume 38, Issue 2, 2000, Pages 177-195

Preston, J. (2007). Whiteness and class in education. Springer Science & Business Media.

Pritlove, C., Juando-Prats, C., Ala-Leppilampi, K., & Parsons, J. A. (2019). The good, the bad, and the ugly of implicit bias. *The Lancet*, *393*(10171), 502-504.

Quinn, D. M. (2020). How to reduce racial bias in grading. *Education Next*, *21*(1).

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, *42*(3), 375-392.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. Journal of Educational psychology, 76(1), 85.

Reardon, S. F., Weathers, E., Fahle, E., Jang, H., & Kalogrides, D. (2019). Is separate still unequal? New evidence on school segregation and racial academic achievement gaps.

Reibel, A. R. (2022). Embracing Relational Teaching: *How Strong Relationships Promote Student Self-Regulation and Efficacy (Strengthening student learning ownership with relational classroom practices)*. Solution Tree Press. Bloomington, IN.

Reibel, A. R. (2022). Teacher Decision-Making Matters: *The Influence of Teacher Choice on Student Learning*. Published on allthingsassessment.com on January 21, 2022. https://allthingsassessment.info/2022/01/21/1870-teacher-decision-making-matters/

Reibel, A. R. & Thede, M., (2020). Small changes, big impact : ten strategies to promote student efficacy and lifelong learning. Bloomington, IN :Solution Tree Press

Riegle-Crumb, C., & Grodsky, E. (2010). Racial-ethnic differences at the intersection of math course-taking and achievement. *Sociology of Education*, *83*(3), 248-270.

Rist, R. (1970) Student Social Class and Teacher Expectations The Self-fulfilling Prophecy in Ghetto Education. Harvard Educational Review, 40, 411-451.

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, *84*(2), 248-274.

Rogers, L. O., Rosario, R. J., & Cielto, J. (2020). The role of stereotypes: Racial identity and learning. In Handbook of the cultural foundations of learning (pp. 62-78). Routledge.

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24-37.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The urban review*, *3*(1), 16-20.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173-220). Academic Press.

Ross, R. (1994). The ladder of inference. *The fifth discipline fieldbook: Strategies and tools for building a learning organization*, 242-246.sing teacher voice and interpretation.

Rubie-Davies, C. M. (2017). Teacher expectations in education. Routledge, UK.

Rubie-Davies, C. M., Watson, P. W. S. J., Flint, A., Garrett, L., & McDonald, L. (2018). Viewing students consistently: how stable are teachers' expectations?. *Educational Research and Evaluation*, *24*(3-5), 221-240.

Rubie-Davies, C. M., & Peterson, E. R. (2016). Relations between teachers' achievement, over- and underestimation, and students' beliefs for Māori and Pākehā students. Contemporary Educational Psychology, 47, 72-83.

Saldaña, J. (2021). The coding manual for qualitative researchers. *The coding manual for qualitative researchers*, 1-440.

Samuels, N. (2013). Diversity and inclusion and the ladder of inference. *Handbook for strategic HR*, 147-151.

Sassenberg, K., & Moskowitz, G. B. (2005). Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology*, *41*(5), 506-514.

Sauer, C. G., Auspurg, K., & Hinz, T. (2020). Designing multi-factorial survey experiments: Effects of presentation style (text or table), answering scales, and vignette order.

Sbarra, D. A., & Pianta, R. C. (2001). Teacher ratings of behavior among African American and Caucasian children during the first two years of school. *Psychology in the Schools*, *38*(3), 229-238.

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.

Schmalor, A., Cheung, B. Y., & Heine, S. J. (2021). Exploring people's thoughts about the causes of ethnic stereotypes. *Plos one*, *16*(1), e0245517.

Schoemaker, P. J. (2011). *Brilliant mistakes: Finding success on the far side of failure*. University of Pennsylvania Press.

Scott, T. M., Gage, N., Hirn, R., & Han, H. (2019). Teacher and student race as a predictor for negative feedback during instruction. School Psychology, 34(1), 22.

Schimmer, T., Hillman, G., & Stalets, M. (2018). *Standards-Based Learning in Action: Moving from Theory to Practice*. Solution Tree. 555 North Morton Street, Bloomington, IN 47404.

Schuster, C., Narciss, S., & Bilz, J. (2021). Well done (for someone of your gender)! Experimental evidence of teachers' stereotype-based shifting standards for test grading and elaborated feedback. *Social Psychology of Education*, 1-26.

Schmalor, A., Cheung, B. Y., & Heine, S. J. (2021). Exploring people's thoughts about the causes of ethnic stereotypes. *Plos one*, *16*(1), e0245517.

Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual review of sociology*, *37*, 57-86.

Sorhagen, N. S. (2013). Early teacher expectations disproportionately affect poor children's high school performance. Journal of Educational Psychology, 105(2), 465.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, *67*, 415-437.

Spillane, J., & Miele, D. (2007). Evidence in practice: A framing of the terrain. *Teachers College Record*, *109*(13), 46-73.

Staats, C. (2016). Understanding implicit bias: What educators should know. American Educator, 39(4), 29.

Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers are people too: Examining the racial bias of teachers compared to other American adults. Educational Researcher, 49(4), 273-284.

Strand, S. (2014). Ethnicity, gender, social class, decision-making and achievement gaps at age 16: Intersectionality and 'getting it' for the white working class. *Research Papers in Education*, *29*(2), 131-171.

Steele, C. (2018). Stereotype threat and African-American student achievement. In *Social Stratification* (pp. 752-756). Routledge.

Taleb, N. N. (2007). The black swan: The impact of the highly improbable (Vol. 2). Random house.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minorities than for European American students? A meta-analysis. *Journal of educational psychology*, *99*(2), 253.

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*, *20*(4), 731-752.

Townsley, M., Buckmiller, T., & Cooper, R. (2019). Anticipating a second wave of standards-based grading implementation and understanding the potential barriers: Perceptions of high school principals. NASSP Bulletin, 103(4), 281-299.

Townsley, M., & Buckmiller, T. (2020). Losing As and Fs: What Works for Schools Implementing Standards-Based Grading?. Educational Considerations, 46(1), 3.

Townsley, M., & Wear, N. L. (2020). Making Grades Matter: Standards-Based Grading in a Secondary PLC at Work®. Solution Tree. 555 North Morton Street, Bloomington, IN 47404.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124-1131.

UCLA Advanced Research Computing, *CHOOSING THE CORRECT STATISTICAL TEST IN SAS, STATA, SPSS AND R* https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/. Accessed 6/15/2022.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. American Educational Research Journal, 47(2), 497-527.

Van den Broeck, L., Demanet, J., & Van Houtte, M. (2020). The forgotten role of teachers in students' educational aspirations. School composition effects and the buffering capacity of teachers' expectations culture. *Teaching and Teacher Education*, *90*, 103015.

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. Economics of Education Review, 30(5), 1045-1058.

Veenstra, K. (2021). The Effects of Standards-Based Grading and Strategies for Implementation: A Review of Literature.

Velasco-Martínez, L. C., & Tójar-Hurtado, J. C. (2018). Competency-Based Evaluation in Higher Education--Design and Use of Competence Rubrics by University Educators. *International Education Studies*, *11*(2), 118-132.

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in psychology*, *9*, 2305.

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. Educational Research and Evaluation, 24(3-5), 124-179.

Warikoo, N., Sinclair, S., Fei, J., & Jacoby-Senghor, D. (2016). Examining racial bias in education: A new approach. Educational Researcher, 45(9), 508-514.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409-421.

Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of educational research*, *42*(2), 203-215.

Weiner, B. (Ed.). (1974). Achievement motivation and attribution theory. General Learning Press.

Weiner, B. (1980). The role of affect in rational (attributional) approaches to human motivation. *Educational Researcher*, *9*(7), 4-11.

Weiner, B. (2012). An attribution theory of motivation. Handbook of theories of social psychology, 1, 135-155.

Weiner, B. (2018). Attribution theory in organizational behavior: A relationship of mutual benefit. In Attribution Theory (pp. 3-6). Routledge.

Westra, E. (2019). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, *196*(7), 2821-2846.

Williams, J. K. (2008). Unspoken realities: White, female teachers discuss race, students, and achievement in the context of teaching in a majority Black elementary school. Oregon State University.

Wilson, T. D. (2004). *Strangers to ourselves*. Harvard University Press.

Wood, D., & Graham, S. (2010), "Why race matters: social context and achievement motivation in African American youth," Urdan, T.C. and Karabenick, S.A. (Ed.) *The Decade Ahead: Applications and Contexts of Motivation and Achievement* (*Advances in Motivation and Achievement, Vol. 16 Part B*), Emerald Group Publishing Limited, Bingley, pp. 175-209. https://doi.org/10.1108/S0749-7423(2010)000016B009

Wormeli, R. (2018). Fair isn't always equal: Assessing & grading in the differentiated classroom. Stenhouse Publishers.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., ... & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, *143*(2), 804.

Zhan, X., Yu, X., Peng, C., & Xi, J. (2021, March). Interpreting Teacher Expectation in Teacher Feedback and Finding the Relationship with Student Learning. In *Society for Information Technology & Teacher Education International Conference* (pp. 991-998). Association for the Advancement of Computing in Education (AACE).

Zuberi, T., & Bonilla-Silva, E. (Eds.). (2008). *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

APPENDIX A
RESEARCH DESIGN

Table 10: Research Design: Vignette Survey Experiment using a 3 X 2 Factorial Design

	Implied White student name	Implied Black student name
Profile Type: Negative "Struggles as a student" Profile	"group 1" 30-35 people Student: Jake	"group 2" 30-35 people Student: Jamal
Profile Type: Name Only	"group 3" 30-35 people Student: Jake	"group 4" 30-35 people Student: Jamal
Profile Type: Positive "Is a good student" Profile	"group 5" 30-35 people Student: Jake	"group 6" 30-35 people Student: Jamal

Table 11: Research Design: Vignette Survey Experiment using a 3 X 2 Factorial Design

	Text presented to respondent for the implied White student name scenario	Text presented to respondent for the implied Black student name scenario
Profile Type: Negative "Struggles as a student" Profile	You are grading the essay of Jake S. Jake is in 11th Grade. Academically, Jake historically performs poorly. Jake did poorly on the previous essay. Overall, Jake has a moderate vocabulary.	You are grading the essay of Jamal S. Jamal is in 11th Grade. Academically, Jamal historically performs poorly. Jamal did poorly on the previous essay. Overall, Jamal has a moderate vocabulary.
Profile Type: Name Only	You are grading the essay of Jake S. Jake is in 11th Grade.	You are grading the essay of Jamal S. Jamal is in 11th Grade.
Profile Type: Positive "Is a good student" Profile	You are grading the essay of Jake S. Jake is in 11th Grade. Academically, Jake historically performs well. Jake did well on the previous essay. Overall, Jake has a great vocabulary.	You are grading the essay of Jamal S. Jamal is in 11th Grade. Academically, Jamal historically performs well. Jamal did well on the previous essay. Overall, Jamal has a great vocabulary.

APPENDIX B
IRB LETTER

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research & Innovation

Office for the Protection of Research Subjects
805 W. Pennsylvania Ave., MC-095
Urbana, IL 61801-4822

Notice of Approval: New Submission

February 10, 2022

Principal Investigator	Rachel Roegman
CC	Anthony Reibel
Protocol Title	*Investigating Racial Bias in Standards-based grading models*
Protocol Number	22689
Funding Source	Unfunded
Review Type	Expedited 7
Status	Active
Risk Determination	No more than minimal risk
Approval Date	February 10, 2022
Closure Date	February 9, 2027

This letter authorizes the use of human subjects in the above protocol. The University of Illinois at Urbana-Champaign Institutional Review Board (IRB) has reviewed and approved the research study as described.

The Principal Investigator of this study is responsible for:

- Conducting research in a manner consistent with the requirements of the University and federal regulations found at 45 CFR 46.
- Using the approved consent documents, with the footer, from this approved package.
- Requesting approval from the IRB prior to implementing modifications.
- Notifying OPRS of any problems involving human subjects, including unanticipated events, participant complaints, or protocol deviations.
- Notifying OPRS of the completion of the study.

APPENDIX C
SAMPLE ESSAY

Hypothetical student-written essay text presented to all respondents

In his book, In Defense of Elitism, William A. Henry III wrote about his views regarding elitism in our society and why it's beneficial. Elitism is the idea that the elite in our society should be valued and praised, and Henry believes that elitism stimulates social progress in our society and encourages self-improvement. However, elitism should not be valued because the elite do not always use their status for good and the non-elite do not get access to the same resources as the elite. Our society should not value elitism because many elites use their power for selfish reasons.

One historical example, is the Robber Barons. Rockefeller, Carnegie, Vanderbilt, and many more men were all businessmen who created successful long-lasting businesses. While they did have high social status they were also cruel and they constantly mistreated workers, made them work in poor conditions, gave them low pay, and they also created a monoply out of their businesses, which is illegal. Contrary to Henry's beliefs, these elites did not use their resources to make society better. A more recent example of corupt elites is the Varsity Blues scandal.

Many colleges require, or at least encourage, a good score on the SAT/ACT. However, children from a low-income household may not go to a high school that has sufficient enough education to give students high test scores. However, standardized test tutoring is quite expensive, and the tests themselves range from $50-$60, quite expensive for low-income families. While Henry believed that elitism would encourage self-improvement, it actually discourages hard work for the elites and prevents non-elites from improving at all and for example, Kim Kardashian is an upper-class woman who is part of a very successful family. However, her status only comes from her mother, who was already rich, and Kim Kardashian adds her daughter, Stormi, to her elite status as well. Stormi, and elementary school girl, gets many more resources than the general public, just because of her elite status. Henry's idea of elitism is flawed because elitism doesn't value self-improvement it only cuts down hard work and equality.

APPENDIX D

RESEARCH INSTRUMENT

Part 1: Respondent Reads Student Profile Information (see appendices A-B for text that appears).

Part 2: Grade a Student's Essay

Please indicate the student's overall mastery of writing a literary argument.

	Exceeds	Meets	Approaching	Beginning
The student can make a literary argument.	o	o	o	o

Please rate the thesis in the essay based on the following standard (choose one):

	Responds to the prompt with a defensible thesis that presents an interpretation in response to the prompt.	The thesis is more of a restatement of the prompt and at times provides more of a summary than a claim.
Thesis	o	o

Please rate the evidence used in the essay by the following standard:

	Provides specific evidence, the evidence is relevant, and/or the evidence strongly supports the argument.	Provides general evidence, evidence is at times irrelevant, and/or the evidence sometimes minimally supports the argument.

Evidence	o	o

Please rate the sophistication evident in the essay by the following standard:

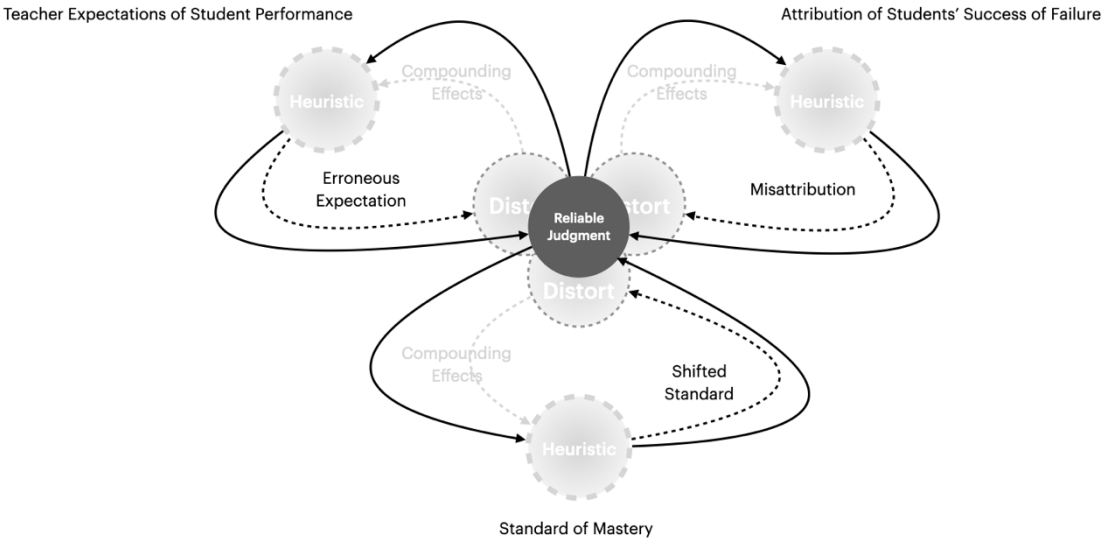
	The student showed nuanced thinking and/or develops a complex literary argument.	The student showed simple, mechanical thinking and/or develops a basic literary argument.
Sophistication	o	o

Part 3: Feedback and Attribution

This student approaches you after getting their rubric scores and says they are nervous about the next essay. Write 2-3 sentences describing what you would say to this student?

APPENDIX E
INTERSECTION OF COGNITIVE BIAS THEORIES

Figure 2: A Conceptual Model Showing the Intersection of Various Influences of Cognitive Biases on Decision-making



APPENDIX F
EXAMPLES OF STANDARDS-BASED RUBRICS

Figure 3: Example of a completed standards-based rubric using a Speaking proficiency-based learning target

Learning Target	4	3	2	1
Independently create an appropriate spoken message in familiar and unstructured situations.	Independently create an appropriate spoken message in unfamiliar and unstructured situations.	Independently create an appropriate spoken message in familiar and unstructured situations.	Independently create an appropriate spoken message in familiar and structured situations.	Independently attempt to create an appropriate spoken message in familiar and structured situations.

Communication Strategies	Student Reflection	Teacher Feedback
Engagement	I used good vocabulary and details. I was a little nervous, which could be why I didn't have the best eye contact and delivery. I forgot to add in the content piece we talked about during the formative practice speech.	Your vocabulary was good, however, you could have included a few more stretch words and terms that were on your vocabulary worksheet. Also, I am hoping that you include more than basic details in your speech. If you read the homework each night, you will gain confidence in your speaking because you can use more details. This will help you improve your body language and delivery.
Delivery		
Risk-Taking		
Body Language		
Supporting Skills		
Vocabulary		
Context Details		
Connections		
Evidence		

Figure 4: Type of Standards-based Grading Rubric from AP Junior English Team at Adlai E. Stevenson High School, Lincolnshire, IL

Junior Accelerated English

Writing Argument Rubric

4 - Exceeding	3 - Meeting	2 - Approaching	1 - Still Developing
I write effective, insightful, logical arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.	**I write arguments to support claims in an analysis of substantive topics or text s, using valid reasoning and relevant and sufficient evidence.**	I write arguments to support claims in an analysis of substantive topics or texts, using partially apparent reasoning and/or limited evidence.	I attempt to write a claim that relies on relevant support in an analysis of substantive topics or texts with questionable reasoning and/or evidence.

Success Criteria	How Well Am I Doing?	Teacher Feedback
Clear focus		Writer takes a clear position in the essay. Other:
Purposeful structure with effective transitions		Structure contributes to the purpose of the essay. Writer guides the reader through the essay with subtle, effective transitions. Other:
Appropriate Development and Support		Writer supports the position through establishing a pattern of reasonable, convincing evidence. Other:
Language, Citations, Voice		Writer demonstrates a command of language, using a formal or informal style that is appropriate to the assignment, appropriate word choice and sentence structure, and suitable rhetorical techniques. Other:

APPENDIX G
HISTOGRAM OF MASTERY AND CRITERIA SCORES

Figure 5: Histograms of Mastery and Criteria Scores