The Association Between Standards-Based Grading and Standardized Test Scores as an Element of a High School Reform Model

MARTY POLLIO

Jeffersontown High School

CRAIG HOCHBEIN

Lehigh University

Background/Context: *From two decades of research on the grading practices of teachers in secondary schools, researchers discovered that teachers evaluated students on numerous factors that do not validly assess a student's achievement level in a specific content area. These consistent findings suggested that traditional grading practices evolved to meet the variety of educational stakeholder expectations for schools, teachers, and students.*

Purpose/Objective/Research Question/Focus of Study*: The purpose of this study was to examine the role of standards-based grading in a high school reform by assessing the relationships between differing grading approaches and standardized test achievement.*

Setting*: The study examined student performance from 11 high schools operating in a large metropolitan school district.*

Population/Participants/Subjects*: The sample of students included two cohorts of 1,163 and 1,256 11th grade students who completed an Algebra 2 course and the state standardized test.*

Intervention/Program/Practice*: Each of the high schools implemented a locally designed reform known as Project Proficiency. A key component of the reform included utilizing standards-based grading to assess student proficiency of the content.*

Research Design: *This study utilized a non-equivalent control group design and quantitative analyses to compare the association between classroom grades and standardized test scores.*

Data Collection and Analysis: *The data for the study included the students' final grades, standardized test scores, and basic demographic information.*

Findings/Results*: Results indicated that the rate of students earning an A or B in a course and passing the state test approximately doubled when utilizing standards-based grading practices. In addition, results indicated that standards-based grading practices identified more predictive and valid assessment of at-risk students' attainment of subject knowledge.*

Teachers College Record Volume 117, 110302, November 2015, 28 pages

0161-4681

***Conclusions/Recommendations**: The article demonstrates the benefits of using standards-based grading in reforms attempting to improve the academic performance of secondary schools, but also notes how restriction of grades to mastery of standards will challenge educators' perception of their abilities and students' efforts. The article also notes the methodological limitations of prior grading research and suggests the need for more robust studies assessing grading practices, student achievement, and school performance.*

INTRODUCTION

Classroom grades play a critical role in the experience of secondary students in the United States. Although evidence has suggested that primary school grades provide important information about future academic achievement (Balfanz, Herzog, & MacIver, 2007), students' grades in secondary school can have immediate and weighty consequences (Hiss & Franks, 2014). Honor societies, scholarship programs, and post-secondary institutions often utilize grades and grade point averages to inform admission and financial decisions. In addition, grades often serve as the gatekeeper for students' participation in extracurricular activities. The National Collegiate Athletic Association (n.d.) even utilizes a sliding scale that incorporates grade point average and standardized college admission test scores to determine the eligibility of potential student-athletes. For students, an insufficient slate of grades could preclude participation, admission, or award.

Although grades have served as a common and important measure for assessing students, grades have lacked a uniform or standard meaning. According to a wide array of research, secondary teachers relied on a variety of factors to determine students' grades (Brookhart, 1993; Cross & Frary, 1999; Guskey, 2009; McMillan, 2001; Stiggins, Frisbie, & Griswold, 1989). For example, teachers utilized assessment of processes such as effort, behavior, class participation, homework completion, ability level, and growth (Brookhart, 1993; Cross & Frary, 1999; Guskey, 2009). Cizek, Fitzgerald, and Rachor (1996) observed, "It seems that classroom assessment practices may be a weak link in the drive toward improving American education" (p. 162). From both the importance and subjectivity of grades emerged a movement in secondary education to grade students solely on achievement in key academic standards within a curriculum (Guskey, 2009; Marzano, 2010).

A shift to standards-based grading requires deep and systematic changes to longstanding educational traditions. To facilitate these changes, reformers first need to demonstrate the benefits of standards-based grading to educators. The purpose of this study was to examine the role of standards-based grading in a high school reform by assessing the relationships

between differing grading approaches and standardized test achievement. Specifically, this study was designed to answer the following research questions:

1. Does a stronger association exist between standards-based grading and standardized test scores than with traditional grading practices?
2. Does a stronger association exist between standards-based grading and minority or disadvantaged students' standardized test scores than with traditional grading practices?

BACKGROUND

The concept of standards-based grading entails associations with a broad array of research topics and policy debates, including but not limited to accountability, standardized testing, teaching practices, and common curricula. Conducting extensive reviews of the empirical investigations and theoretical discourse in these tangential areas would divert focus from the processes and outcomes of classroom grading practices. Furthermore, this study did not evaluate the policies that currently govern public education, but rather focused on the success of an initiative operating within those parameters. However, given the complex endeavors of teaching and learning, as well as the complex associations of standards-based grading, we explicated our assumptions prior to reviewing the literature related to grading.

ASSUMPTIONS OF SCHOOLS, TEACHERS, AND GRADES

For decades, schools have performed a variety of functions. In addition to educating students about specific content knowledge, schools have offered numerous extracurricular activities, such as athletic teams, artistic performances, social clubs, and other non-academic endeavors (Duke, 1995). To meet the needs of students from economically disadvantaged backgrounds, many schools participate in subsidized student meal programs (Harwell & LeBeau, 2010), with a growing number of schools providing additional well-being services, such as health, dental, and counseling clinics (Moore, 2014). Furthermore, schools have grappled with greater societal issues, such as gender and racial integration (Grant, 2009; Ogbu, 2003; Reese, 2005; Tyack & Hansot, 1990). Although schools have attempted to achieve an array of important and meaningful objectives, we assumed student attainment of subject knowledge, such as literacy or numeracy, as the primary responsibility of a school.

Similarly, educational stakeholders, including taxpayers, parents, and administrators, have expected teachers to achieve multiple objectives.

In addition to the content of core subjects like reading, writing, arithmetic, history, and the sciences, teachers have been expected to provide instruction in non-academic skills. For instance, discussing report cards from New York City between 1920 and 1940, Cuban (1993) noted, "Space was provided for grades on effort, conduct, and personal habits" (p. 58). In the daily activity of schools and classrooms, teachers deliver instruction and guidance on critical skills like responsibility, creativity, resiliency, and others (Labaree, 2012). Although teachers often provide critical instruction in non-academic skills, we assumed students' mastery of subject knowledge as the primary responsibility of teachers.

For schools and teachers, grades have operated as the primary form of communicating the performance of students to educational stakeholders. Despite this ubiquitous practice of educational communication, grades have not encompassed a standard meaning. Grades might communicate student growth or diligence to external stakeholders (McMillan & Nash, 2000), as well as serve as rewards or sanctions for students (Brookhart, 1993). Such practices often relay beneficial or meaningful information (Bowers, 2009), but do not necessarily communicate students' academic achievement. To align with our expectations about the primary objectives of schools and teachers, we also assumed that grades validly and reliably represented students' mastery of subject knowledge.

Regardless of individual views on the purpose of schooling or responsibilities of teachers, federal and state governments currently hold schools and teachers accountable for student learning of specific standards. To demonstrate proficiency of these standards, students complete a battery of standardized state accountability assessments. One critical component of this educational reform may be the implementation of sound grading practices that directly measure student attainment of required standards. Without valid grading practices, students are likely to enter into their exam sessions with an invalid comprehension of their knowledge and abilities (Rosenbaum, 1997). Similarly, without valid grading practices, teachers might misjudge and therefore mismanage precious instructional time. To ensure a high-quality education for all students, grading reform must become pervasive throughout secondary education in America. As Guskey (2009) states,

> If grades are to represent information about the adequacy of students' performance with respect to clear learning standards, then the evidence used in determining grades must denote what students have learned and are able to do. To allow other factors to influence students' grades or to maintain policies that detract from that purpose misrepresents students' learning attainment. (p. 22)

TRADITIONS OF GRADING

To make systemic change within secondary education, measurement researchers stated that grades need to be based solely on levels of achievement within a class (Allen, 2005; Cross & Frary, 1999; Guskey, 2009). The vast majority of prior research on grading in secondary education indicated that most teachers do not focus grading on achievement. Brookhart (1991) initially described the grading process in secondary schools as a "hodgepodge of attitude, effort, and achievement" (p. 36). Most of these research studies involved surveying teachers on the various factors that they take into account when giving a student a grade in their class. For instance, Brookhart (1993) found that 84 surveyed teachers used the image of grades as currency to encourage student effort, participation, and appropriate behavior within the classroom. Cross and Frary (1999) further explored Brookhart's findings of the variety of factors used in the grading of secondary students. On the basis of their survey of 307 teachers, the researchers confirmed teachers' use of many non-achievement factors in grading students when they concluded:

> Because of the importance placed on academic grades at the secondary level, either for educational or occupational decisions, grades should communicate as objectively as possible the levels of educational attainment in the subject. To encourage anything less, in our opinion, is to distort the meaning of grades as measures of academic achievement, at a time when the need for clarity of meaning is greatest. (Cross & Frary, 1999, p. 56)

McMillan and Nash (2000) further investigated the influences on teacher decision-making with respect to grading and the justification that teachers gave when assigning grades. The authors surveyed 700 teachers and then interviewed a sample of these teachers. From the teacher responses, the researchers identified various classroom factors involved in grading. Although achievement, as defined by student understanding, was one of the primary categories, several other categories emerged. Such categories included the teachers' philosophy of teaching and learning, their desire to "pull for students," their accommodations for individual differences among students, and finally student engagement and motivation. Supporting Brookhart's (1991) assertions, McMillan and Nash (2000) concluded that teachers used grades as the main tool to encourage and monitor student engagement. Although teachers verbalized the need to measure student achievement through grading, "most teachers used a variety of assessments . . . including homework, quizzes, tests, performance assessments and participation" (McMillan & Nash, 2000, p. 26).

To better understand and explore the various factors used in grading, McMillan (2001) surveyed 1,483 teachers and identified four distinct factors most often seen in secondary grading practices. These factors included academic achievement, external benchmarks, academic enablers, and extra credit. In addition, McMillan discovered that teachers assessed higher-ability students in a motivating and engaging environment by measuring higher cognitive skills, while the same teachers gave lower-ability students more rote learning assessments, more extra credit, and less emphasis on academic achievement. Discovery of such differential grading suggested that grading practices in secondary schools maintained or possibly increased achievement gaps between student subgroups. Whereas teachers graded higher-ability students based upon achievement, they graded many at-risk students utilizing a wider range of factors. This wider range of factors potentially inflated students' grades, making them less valid indicators of standards' achievement, which subsequently obscured the students' needs for additional instruction, practice, or remediation.

VALIDITY OF GRADES

The results of the survey research of secondary teachers' grading practices exhibited that teachers used a variety of factors to grade students. Student achievement emerged as only one of the factors used by teachers to assess student work. Therefore, grades are not necessarily a valid measure of students' level of achievement in secondary education. Despite this lack of validity, educators utilize grades to make critical decisions about students' future, such as entry into elite clubs and organizations, access to scholarships, and admissions into college. If grades measure several factors, including a student's ability to navigate the social processes of school, and not just academic achievement, the validity of grades becomes a major concern in American education. For grades to be a valid measure of student achievement, teachers must assess students on their achievement based on required curriculum standards.

As a result of the variety of factors used by teachers to grade students, Marzano (2000) contended that in terms of measuring student achievement, "grades are so imprecise that they are almost meaningless" (p. 1). Allen (2005) summarized the critical nature of ensuring validity in the grading process in measuring academic achievement:

> Also, since many of these factors such as effort, motivation, and student attitude are subjective measures made by a teacher, their inclusion in a grade related to academic achievement increases the chance for the grade to be biased or unreliable, and thus

> invalid. The purpose of an academic report is to communicate the level of academic achievement that a student has developed over a course of study. Therefore, the sole purpose of a grade on an academic report, if it is to be a valid source of information, is to communicate the academic achievement of a student. (p. 220)

Guskey (2007) explored the perceived validity of teacher grades by surveying 314 educators in three different states. He asked educators to rank from 1 to 15 sources of evidence of student learning that "you trust to best show what students know and can do" (p. 21). The sources of evidence included standardized tests, various assessments, teacher observations, quizzes, homework completion, portfolio, students' grades, class involvement, and behavior and attitude. Statistical analyses of the data indicated that the participants gave a relatively low ranking to grades being an accurate indicator of student learning. Guskey (2007) concluded that the educators' low ranking of grades correlating with academic achievement resulted from both teachers' and administrators' recognition "that a variety of nonacademic factors, such as effort, attitude, participation, and class behavior, typically influence grades" (p. 22). Such factors also supported the discrepancies between student grades and standardized test scores (Allen, 2005).

In another study on the validity of grades, Bowers (2009) explored the relationship between teacher-assigned grades and standardized assessments. He found that schools used standardized test scores, in place of grades, to make data-driven decisions. Administrators have consistently sought to remediate and intervene for low performance on standardized tests, when student grades should also be used to inform these decisions. Conceding that grades are not a valid measure of a student's academic achievement, Bowers (2009) suggested how schools could better use grades as a basis to provide critical safety nets to support student success:

> The hypothesis here is that rather than cast this hodgepodge nature of grades in pejorative light as data that is useful to schools because grades only moderately correlate to test scores, the theory presented here . . . points to the idea that grades appear to assess both academic knowledge . . . as well as a student's ability to perform well at the social tasks of the schooling process, such as behavior, participation, and attendance. (p. 622)

RELATIONSHIPS BETWEEN GRADES, TEST SCORES, AND STUDENTS

Few empirical research studies have rigorously investigated the relationship between grades and standardized test scores. Little evidence exists on the impact of standards-based grading on standardized test scores. Welsh, D'Agostino, and Kaniskan (2013) even stated, "However, as far as we know, the linkage between SBPR (standards-based progress reports) and standards-based assessment scores has not been explored in the academic literature" (p. 26). Yet, as a result of schools' increased accountability for improving standardized test scores, several research studies have attempted to determine the relationship between grades and test scores. If recent increases in school accountability have led to changes in teacher grading practices, then an association should exist between grades and standardized test scores. If the use of standards-based grading methods has led to a decrease in the use of a medley of factors to assess student learning, and grades were more of a valid indicator of student achievement, then a strong correlation should have existed between grades and test scores.

Conley (2000) first examined the relationship between grades teachers give their students and proficiency scores given to the same students by external raters. Conley found little correlation between teachers' grading system and student proficiency. He specifically noted that students judged proficient through an analysis of their work by external raters were not necessarily the students with high grades. "The stepwise regression analysis examines teacher grading systems and student proficiency scores and found very little relationship between the grading system a teacher used and whether or not a student was proficient" (Conley, 2000, p. 18). Conley surmised that the low correlation suggested that separate constructs besides standards-based achievement were used in grading. Specifically, he noted that homework in mathematics classes and in-class assignments in English classes comprised a significant portion of a student's grade, although these assignments might not measure proficiency on mandated standards.

This relationship between test scores and grades was the topic of several research studies over the past decade. Lekholm and Cliffordson (2008) studied the grades of nearly 100,000 students from Sweden and their association with students' scores on national tests. Although results from their analysis indicated that the greatest variance in grades came from actual achievement levels in the subject area, other factors outside of achievement influenced the grades given to students. One of the most significant findings of their research revealed that schools with students from lower socio-economic levels assigned grades that were higher than the students' standardized test scores. Therefore, the at-risk students in these schools evince a lower correlation between grades and test scores.

Two other studies examined the correlation between grades and standardized tests, as well as the differences in the association between grades and test scores for minority, low socio-economic, and non-minority students. Brennan, Kim, Wenz-Gross, and Sipperstein (2001) and Haptonstall (2010) discovered modest correlations between teacher-assigned grades and standardized state assessments. However, both studies found a lower correlation between grades and standardized test scores for minority students, English language learners, and low socio-economic students than their counterparts. The findings suggested not only that grades did not strongly correlate with achievement scores on standardized tests, but that minority students and low socio-economic students were possibly given higher grades than their achievement levels warranted.

Together, the findings of Brennan et al. (2001), Haptonstall (2010), Lekholm and Cliffordson (2008), and McMillan (2001) supported a theory of grade inflation with minority and disadvantaged students. As a result of teachers including factors such as effort, behavior, and attendance, minority and disadvantaged students earned grades that overestimated academic attainment. According to Brennan et al. (2001) and Haptonstall (2010), the practice of grading at-risk students on factors other than achievement level supports the existence of a significant achievement gap between minority students and their White counterparts. Despite intense focus on the elimination of the achievement gap in American secondary schools, few education leaders have examined grading policies as a potential source of the problem.

Based on two decades of research on the grading practices of teachers in secondary schools, researchers found that teachers evaluated students on numerous factors that do not validly assess a student's achievement level in a specific content area. These consistent findings suggested that traditional grading practices evolved to meet the variety of educational stakeholder expectations for schools, teachers, and students. However, research data are limited on the correlation between grades and achievement scores on standardized tests at the secondary level. Even less research is available on the impact of standards-based grading on the correlation between grades and achievement scores. To address these data and research deficits, this study examined the association between standards-based grading and achievement as measured by standardized tests, for students in general and for minority and low socio-economic students in particular.

METHODOLOGY

PROJECT PROFICIENCY

In 2010, the Kentucky Department of Education (KDE) enacted Race to the Top legislation that forced the state to identify its lowest-performing schools and implement a federally approved turnaround model. Within Jefferson County Public Schools (JCPS), KDE identified 18 schools as persistently low-achieving (PLA) because of their low mathematics and reading scores on state accountability assessments. In an effort to improve reading and mathematics scores, JCPS educators implemented Project Proficiency (PP), a district-wide instructional program to standardize both mathematics and reading curricula and instruction at the secondary level (JCPS, 2011a). During the 2010–2012 school years, 11 JCPS secondary schools implemented PP to improve academic achievement. Although we limit our description of PP to aspects of the reform with direct influence on standards-based grading, the *Project Proficiency Guide* (JCPS, 2011a) provides more detailed information about the specific design of the initiative. In addition, Baete and Hochbein (2014) and Burks and Hochbein (2013) provide supplemental information about the implementation of PP across the district and within the participating schools.

A central strategy of PP, to guarantee student competency of subject knowledge, required mathematics instruction to focus on student attainment of key course standards. The curriculum identified three key standards within a grading period for high school mathematics courses. Through intentional dissection of key standards, measureable learning targets guided daily instruction across each of the classrooms and schools. The key standards and learning targets focused instruction and aligned curriculum for each classroom within the school and across each of the 11 district high schools. Through a unified process of instruction and interventions measured by district-designed common assessments, teachers attempted to guarantee competency of every student.

Within the PP initiative, teachers graded students solely on their proficiency level for each of the key standards within the grading period. Students took both a diagnostic assessment in the middle of the grading period and a proficiency assessment at the end of the grading period. Each of these assessments accounted for 40% of the students' final grade. Student reflection on their proficiency within each standard accounted for the final 20% (JCPS, 2011a). Finally, for students who do not reach proficiency, PP required schools to make accommodations to remediate after the grading period in order to meet the key standards and retake proficiency assessments. Teachers relied on grades to identify

and implement interventions for students who did not demonstrate proficiency of key standards.

All 11 high schools participating in PP implemented the initiative in every Algebra 1, Geometry, and Algebra 2 classroom. Within each of these mathematics classrooms, teachers attempted to ensure overall student competency by focusing on their proficiency in key standards for every student through a standards-based grading approach. As a result, grades became the key indicator to identify which students needed interventions and the additional instructional support needed to reach competency in the key standards. The design and implementation of PP also fostered teacher collaboration to develop and implement successful instructional strategies, as well as interventions for students performing below competency benchmarks.

PARTICIPANTS

This study included participants from 11 high schools that implemented PP for the 2010–2011 school year. Educators implemented PP in an effort to improve the rate of proficient scores earned by students on the Kentucky Core Content Test (KCCT) in mathematics. All 11 high schools operated as part of JCPS in Louisville, Kentucky. As the 26th-largest school district in the nation, JCPS served 100,474 students in 150 schools. The demographic composition of the district consisted of 51% White, 37% Black, and 12% Other students. Nearly 62% of the student population qualified for the federal free/reduced lunch program (Table 1).

From the high school population, we identified two separate cohort groups (Table 1). In Kentucky, students take KCCT assessments in mathematics (11Math) and science (11Science) during their junior year of high school. Students included in this study completed an Algebra 2 course during the 2010 or 2011 school years and had grade 11 KCCT results in mathematics (11Math) and science (11Science) during the same year. One cohort consisted of 11th grade students from the 11 high schools during the 2011 school year. Each of these students completed an Algebra 2 course and received PP within the Algebra 2 course. The second cohort consisted of 11th graders during the 2010 school year from the same 11 high schools, but who did not receive PP within their Algebra 2 course. As a result of testing and district reform, juniors from the 11 high schools experienced PP in math, but not in science. Therefore, we analyzed science scores as a non-equivalent control group (Shadish, Cook, & Campbell, 2002) to compare the effects of PP between the same students in two different courses.

Table 1. Demographic Characteristics of Cohort Participants (N =2,419)

	non-PP Cohort (2010)		PP Cohort (2011)	
Characteristic	n	%	n	%
Gender				
Male	585	50.3	610	48.6
Female	578	49.7	646	51.4
Race/ethnicity				
Caucasian/White	509	43.8	552	43.9
Minority (non-White)	654	56.2	704	56.1
Free/reduced lunch	822	70.7	920	73.2

For purposes of this study, we defined Algebra 2 as a course in the Kentucky Program of Studies that met the Kentucky Algebra 2 graduation requirement. In JCPS and the 11 high schools, these courses included Algebra 2, Algebra 2 Honors, and Algebra 2 Advanced. Algebra 2 classrooms were identified in each of the study's schools through an evaluation of each master schedule. The 2010 cohort sample contained 1,163 (n = 1,163) students across 11 high schools. The 2011 cohort sample contained 1,256 (n = 1,256) students across the same 11 high schools. The three utilized courses included students who qualified for special education services. However, students who did not complete an Algebra 2 course or who did not have KCCT results in mathematics and science during the same academic year were excluded from the study.

MEASURES

We obtained student data for each of the cohorts from the JCPS Data Warehouse. For security reasons, when Kentucky administered the mathematics and science tests, proctors distributed multiple versions of the tests. The KCCT Test Administration Guide identified average Chronbach's Alpha measures of .89 and .84, respectively, for the six versions of the mathematics and science tests. Item and description indices were identified by the Kentucky Department of Education for each test version and converted to mean scale scores (MSS) from 0–80. Kentucky mean scale scores correlated to four performance-level descriptors: novice, apprentice, proficient, and distinguished (Table 2).

Table 2: Grade 11 Kentucky Core Content Test (KCCT) Mean Scale Score Range and Performance Descriptors

	Scale Score Range			
Content	Novice	Apprentice	Proficient	Distinguished
Mathematics	0–19	20–39	40–63	64–80
Science	0–19	20–39	40–62	63–80

Students' results on KCCT mathematics and science assessments served as outcome measures. We analyzed 11Math as the primary dependent variable, but utilized 11Science as an additional dependent variable to compare the effects of PP within the same treatment group. Specifically, the particular effect analyzed within this study was the association between standards-based grades within PP and 11Math scores. The use of standards-based grading in PP was only used with the 2011 cohort in Algebra 2 courses. In contrast, the 2010 cohort in mathematics and the 2011 cohort in science experienced a traditional grading approach, which was outlined in The Jefferson County Public Schools Student Progression, Promotion and Grading (SPP&G) (JCPS, 2011b) (Table 3). The SPP&G defined district policy concerning the components of an academic grade. The policy stated, "Academic grades must include a minimum of three of the following: portfolios, projects, discussion/problem solving, group work, classroom assignments, homework/journals/logs, quizzes, tests, participation, and teacher observation" (p. 8). Finally, district policy also mandated that "one component may not count for more than 40 percent of the total academic grade" (p. 8).

Table 3: Classroom Letter Grade Values and Ranges

Letter Grade	Value	Range
A (Exceeds Standards)	4.0	93–100
B (Meets Standards)	3.0	86–92
C (Marginally Meets Standards)	2.0	79–85
D (Below Standards)	1.0	70–78
U (Unsatisfactory Performance)	0.0	0–69

Note. Excerpted from JCPS "Student Progression, Promotion and Grading" (2011b)

The independent variable used for standards-based grading was the implementation of PP. Instead of using the traditional grading method, PP assessed students based on the standards-based grading approach. As a part of the standards-based grading process in 2011, teachers required their Algebra 2 students to become proficient in three key standards for

each six-week grading period. As a result, grades for mathematics in the 2011 cohort were based solely on a student's proficiency level in the three key standards for the six weeks.

DESIGN AND ANALYSES TO EVALUATE GRADING

A quasi-experimental non-equivalent control group design was implemented to analyze the association between standards-based grades and 11Math scores during the 2011 school year. Comparison of the association between grades and KCCT scores relied on two control groups. The first control group consisted of students who completed an Algebra 2 course and received an 11Math score in 2010, but did not experience the standards-based grading effects within PP. This control group provided a measure of association between two years in 11Math results. We also utilized a second control group, which provided a comparison of association between two groups of the same students. The second control group involved the same students as the treatment group. However, the students received standards-based grading as a part of PP in mathematics, but not in science. Therefore, we also analyzed the association between their science grades and 11Science scores and compared results to the association found in mathematics.

Inclusion of the non-equivalent control group, 11Science, reduced some of the most likely or greatest threats to validity. A cross-sectional single comparison of 2010 and 2011 results would be susceptible to several validity threats, including historical, selection, and compensatory issues (Shadish et al., 2002). For example, one administrator suggested that any improvement in 11Math resulted because of teachers "getting fire in their bellies." However, comparison of different students in the same subject, and the same students in different subjects, reduced the likelihood of such threats biasing results. Of course, not all threats could be eliminated, but as Shadish et al. (2002) stated, "Therefore, possible threats to validity are not always plausible ones" (p. 139).

EVALUATION AND ANALYSES OF GRADES

To analyze the influence of PP on the association between grades and test scores, we first tabulated basic descriptive statistics to determine the percentage of students in each cohort that scored proficient or above and received an A or a B in the corresponding content course. On the KCCT assessment, proficient or above is the level necessary for students to score in order for schools to avoid state and federal sanctions. Second, we evaluated students who received above average grades within the content course for each of the three cohorts. Students who received an A or a B in

the specific content course were considered above average in standard attainment. An analysis of variance (ANOVA) determined whether students who experienced standards-based grading and scored above average in their class scored higher on the corresponding KCCT assessment than students who experienced traditional grading.

Third, we determined the correlation coefficient (r) for all students in each of the three groups. By analyzing the correlation coefficient for each of the groups, the researchers determined whether the treatment group that received standards-based grading had a higher correlation to KCCT mathematics scores than the control groups that did not receive standards-based grading. A coefficient of determination (r^2) was determined for each of the groups, as well as for both minority students and students receiving free/reduced lunch (FRL) aid. Finally, a regression analysis measured the association of grades and the corresponding KCCT score between the groups. (1) Prior achievement as measured by eighth-grade mean scale scores, (2) FRL status, and (3) grades within the specific content course were all included as control variables to create a robust, yet parsimonious model.

RESULTS

CROSS TABULATION OF GRADES AND TEST SCORES

The researchers analyzed descriptive statistics of student grades and test scores to determine the percentage of students who received an above average grade in their mathematics or science course (A/B) and also scored proficient or distinguished on the corresponding KCCT test (Table 4). If students' grades were a valid indicator of their learning subject content, then students who scored an A or a B in their content class should have scored proficient or above on the state accountability assessment. With the students who experienced traditional grading methods, in both mathematics and science, this assumption did not prove true. In the non-PP Math cohort, 466 students (40%) received an A or a B in their Algebra 2 class, yet only 26% of them scored a proficient or distinguished on the 2010 KCCT mathematics assessment. Within the PP Science group, 514 students (40%) received an A or a B in their science class, of which 28% scored a proficient or distinguished on the 2011 KCCT science assessment. Within two traditional grading cohorts, success in the classroom as defined by grades did not translate into success on the KCCT assessment.

For students who experienced standards-based grading in PP Math, 568 (45%) received an A or a B in their Algebra 2 class, with 55% of them scoring proficient or distinguished on the 2011 KCCT mathematics

assessment. When teachers utilized standards-based grading methods, not only did the number of As and Bs increase, but the rate of passing the state assessment among students who earned these grades approximately doubled as compared to the two traditional grading cohorts. Despite this increase in proficiency among students who experienced standards-based grading, educators should be concerned that 45% of the students in the PP Math Cohort who achieved an A or a B in their Algebra 2 class still did not meet KCCT mathematics assessment proficiency.

Table 4. KCCT Performance Description as a Function of Subject Grade

Cohort	Grade	Novice		Apprentice		Proficient		Distinguished		Grade Totals	
		n	%	n	%	n	%	n	%	n	%
non-PP Math	A	35	3	90	8	62	5	9	1	196	17
	B	70	6	148	13	50	4	2	0	270	23
	C	113	10	139	12	34	3	1	0	287	25
	D	141	12	108	9	22	2	1	0	272	23
	U	85	7	44	4	9	1	0	0	138	12
Totals		444	38	529	46	177	15	13	1	1163	100
PP Science	A	37	3	82	7	69	5	3	0	191	15
	B	72	6	180	14	68	5	3	0	323	25
	C	108	9	147	12	34	3	2	0	291	24
	D	129	10	148	12	32	3	2	0	311	25
	U	58	5	69	5	13	1	0	0	140	11
Totals		404	33	626	50	216	17	10	0	1256	100
PP Math	A	7	1	48	4	128	10	18	1	201	16
	B	45	4	153	12	165	13	4	0	367	29
	C	74	6	164	13	94	7	4	0	336	26
	D	62	5	102	8	45	4	0	0	209	17
	U	62	5	57	5	22	2	2	0	143	12
Totals		250	21	524	42	454	36	28	1	1256	100

RELATIONSHIP BETWEEN GRADES AND TEST SCORES

Further data analysis assessed the association between grades and test scores. A Pearson correlation coefficient was calculated for the relationship between participants' grades and KCCT test scores in each of the three groups (Table 5). A weak but positive correlation was found in both

PP Math and non-PP Math, although the magnitude of the PP Math was greater. Overall, the correlation results indicated that students with higher grades tended to score higher on the KCCT mathematics test. However, for PP Sciences, little if any correlation between grades and test scores existed (Hinkle, Wiersma, & Jurs, 2003). Supplementary analysis revealed that correlations between grades and test scores among the minority and FRL student subgroups mirrored patterns found among the samples of all students. However, for each cohort, the correlation magnitudes for minority students were consistently less than both the FRL and All samples.

Table 5. Cohort Correlations Between Grades and Test Scores

	PP Math		**non-PP Math**		**PP Science**	
Correlation	*r*	r^2	r	r^2	r	r^2
All	.45	.20	.38	.15	.26	.07
Minority	.39	.15	.36	.13	.23	.05
FRL	.44	.19	.39	.15	.27	.07

Note. FRL = free or reduced lunch price eligible

Correlational analyses suggested that students who experienced standards-based grading had both stronger correlations between grades and assessment scores and stronger coefficients of determination than students who experienced traditional grading. Minority and FRL students who experienced standards-based grading in PP Math demonstrated greater correlations between grades and test scores as compared to the cohorts that used traditional grading methods. However, educators should again be concerned that regardless of subject or cohort, at best grades demonstrated weak positive correlations with achievement scores. Even with standards-based grading in PP Math, the association between grades and test scores was weak.

ANALYSIS OF VARIANCE IN CONTRAST OF GRADES

Further analysis of the data explored the mean KCCT test scores between each grade for the three groups. A one-way ANOVA compared KCCT test results based upon the earned grades of participants in each cohort. This contrast examined if groupings by subject grade revealed differences in KCCT test scores. For PP Math, non-PP Math, and PP Science, results indicated statistically significant differences between grades and KCCT scores (Table 6). Although correlational analysis revealed weak positive correlations between grades and test scores, ANOVA results indicated that students with higher grades earned statistically significantly

higher KCCT test scores. For example, students in PP Math who received an A (M = 47.31, SD = 13.77) scored higher on the KCCT assessment than students who received a B (M = 37.48, SD = 15.11). Further tests revealed that students continued to score lower on the KCCT assessment based on their specific grade.

Table 6. Mean KCCT Scores by Classroom Grade

Grade	PP Math[a]		non-PP Math[b]		PP Science[c]	
	Absolute	Centered	Absolute	Centered	Absolute	Centered
A	47.31	10.31	35.40	-0.60	34.26	-1.74
B	37.48	.48	28.69	-7.31	29.50	-6.50
C	31.18	-5.82	23.34	-12.66	24.69	-11.31
D	26.93	-10.07	19.96	-16.04	23.29	-12.71
U	22.34	-14.66	16.59	-19.41	22.04	-13.96

Note. [a]State mean = 37.00. [b]State mean = 36.00. [c]State mean = 36.00. Centered = Absolute – State mean.

Beyond statistically significant differences, ANOVA results also identified several differences of practical and pragmatic importance. First, only the average KCCT score from students who earned an A in PP Math qualified for designation as proficient by KCCT standards. Second, on average students from PP Math who achieved an A or a B earned a KCCT mathematics score greater than the state mean of 37.00. In contrast, students from non-PP Math and PP Science who earned an A in their content course on average scored less than the state mean on the corresponding KCCT assessment. Third, mean student scores from the three groupings demonstrated that PP Math students not only achieved greater test scores per grade, but also greater distinction between grades. For instance, PP Math students who earned an A averaged nearly 25 points higher on the KCCT mathematics assessment than students who failed the course, yet students who achieved an A in PP Science averaged 12 points higher on the KCCT science assessment than those students who failed the course. Similar to the descriptive and correlation results, these analyses indicated a greater association between grades and KCCT scores in standards-based grading than with traditional grading methods.

LINEAR REGRESSION RESULTS

Finally, a multiple linear regression estimated the variance accounted for in test scores utilizing several predictors, including grades. Prior academic achievement was the strongest predictor of KCCT achievement score in

mathematics or science for 11th grade students (Table 7). Grades within the specific content course were also significant predictors of achievement on the corresponding KCCT assessment. Although the grade within the course was significant in all three groups, it was a stronger predictor among students who experienced standards-based grading in PP Math than among students who experienced traditional grading in the non-PP groups. The PP Math regression model estimated the greatest standardized coefficient for grades (β = .25, $p < .001$), although the non-PP Math cohort (β = .22, $p < .001$) estimate was similar. Finally, unlike the non-PP Math and PP Science models, student FRL status was not a statistically significant predictor of test scores, above and beyond grades and prior achievement, in the PP Math model.

Table 7. Factors Accounting for KCCT Test Scores

Cohort	Factor	B	*se*(B)	β	*t*
non-PP Math	Constant	3.72	.94		3.96**
	FRL	-1.55	.70	-.05	-2.20*
	Achievement	.57	.02	.62	28.67**
	Grade	2.72	.26	.22	10.46**
PP Science	Constant	10.22	1.03		9.96**
	FRL	-1.77	.76	-.05	-2.34*
	Achievement	.48	.02	.56	24.54**
	Grade	1.80	.27	.15	6.61**
PP Math	Constant	9.85	.99		1.00**
	FRL	-.89	.73	-.02	-1.21
	Achievement	.59	.02	.61	29.94**
	Grade	3.38	.28	.25	12.23**

Note. *$p < .05$. **$p < .001$. B = unstandardized coefficient. β = standardized coefficient.

ANECDOTAL OBSERVATIONS

Unfortunately, a variety of constraints related to time and resources precluded rigorous analyses using qualitative methods. However, one of the authors participated in PP as the principal of an implementing school. Although he did not systematically interview or conduct focus groups, he actively observed the influence of PP on teachers and students in his school. In addition, his interactions with other district principals provided information about the broad perceptions of PP from the 10 other

implementing schools. As a principal who continued to utilize and expand the use of standards-based grading, his observations entail some bias. Thus, we limited the anecdotal observations to two primary concerns of standards-based grading: narrowed curriculum and teacher reactions.

The identification and emphasis of key standards holds the potential to narrow curricula and instruction to just "teaching to the test." However, hundreds of hours spent observing classrooms and interacting with teachers, students, and parents did not appear to substantiate this concern. Instead, teachers noted that the limited number of standards provided them with more planning and instructional time, in which they developed and delivered lessons superior to previous years. In addition, teachers positively noted that the identification of specific standards enabled them to focus on depth of content, instead of breadth. This time and focus likely contributed to teachers commenting that they felt their students gained a deeper understanding of the key standards when using PP and standards-based grading. Moreover, parent, teacher, and student discussions about grades appeared more meaningful and thoughtful. Instead of debating the number of points a student should have been awarded, more and more of these conversations focused on how the student demonstrated proficiency in a specific standard.

Although standards-based grading enabled teachers to focus instruction and conversations on subject matter attainment, this emphasis on attainment also challenged them when assessing diligent, but underperforming students. Teachers struggled with the notion of assigning low grades to students they perceived as "good." Teachers continued to fear that grades that did not match students' efforts would discourage and deter such students. Interestingly, the converse of this logic did not occur. For students who were less diligent or presented classroom disruptions, but demonstrated proficiency of the key standards, teachers did not begrudge high grades for these students.

This anecdotal evidence cannot conclusively refute or substantiate the quantitative results. Yet the insights gleaned from extensive work with educators and students experiencing standards-based grading add some descriptive nuance to the findings. In answering our research questions, we did not consider this anecdotal evidence. However, our considerations of implications for policy, practice, and research relied on our review of prior literature, the quantitative evidence, and the experiences of implementing PP.

SYNTHESIS AND CONCLUSIONS

Attributing the observed improvement in student achievement solely to the implementation of standards-based grading practices would be an inaccurate interpretation of this study's results. Standards-based grading was one component of the comprehensive PP reform. As demonstrated by Baete and Hochbein (2014) and Burks and Hochbein (2013), the package of curricular and instructional changes resulting from PP, which included standards-based grading, contributed to increases in student achievement. However, these prior studies did not specifically evaluate the influence of standards-based grading practices as part of PP. The methodology of the present study isolated the association between student grades and standardized test scores to compare standards-based with traditional grading practices.

With regard to the first research question, a stronger association existed between course grades and standardized test scores among students who experienced standards-based grading as opposed to students who experienced traditional grading methods. First, descriptive statistics found that more students who achieved an A or a B in their class scored proficient or above on state accountability testing when they experienced standards-based grading as opposed to traditional grading. Second, the magnitude of the Pearson correlation coefficient between grades and test scores was stronger for the PP Math cohort. Third, the analysis of variance found that students who achieved higher grades in their mathematics class also achieved higher scores on the KCCT assessment when they experienced standards-based grading. Finally, regression estimates indicated that grades in standards-based as compared to traditional grading accounted for a greater amount of variance in student test scores.

In terms of the second research question, there was a stronger association between grades and test scores of minority or disadvantaged students in standards-based grading. The correlation coefficients of the standards-based grading cohort for minority and FRL students exceeded those of the traditional grading models. Statistics in Table 5 show that, among FRL students, the proportion of variance in the relationship between grades and test scores was over twice as high for PP Math students (19%) who experienced standards-based grading as compared to PP Science students (7%) who experienced traditional grading. The exact same at-risk students had higher correlation statistics when evaluated on standards-based grading as opposed to traditional grading. In addition, FRL status was not a statistically significant predictor in the PP Math model, but was statistically significant in the non-PP Math and PP Science models. This difference in models suggests that standards-based grading weakened the

negative association between socioeconomic status and student achievement. Together, these data support prior research, which suggested that teachers grade minority students less on achievement levels and more on a variety of additional factors (Brennan et al., 2001; Haptonstall, 2010).

Although the consistency of the findings supports the benefits of standards-based grading over traditional models, two limitations require discussion and potentially temper conclusions. First, the level of implementation of standards-based grading within each school and classroom was not explored in this study. Teachers in each of the 11 schools implemented PP within their mathematics classrooms, and this implementation required a certain level of fidelity with standards-based grading. The tenets of PP required that teachers "guarantee competency" of each of their students on three key standards for each six-week grading period. Without implementing a standards-based grading approach, teachers could not ensure that each student had met the three key standards. Schools and classrooms, however, could vary in their level of implementation with standards-based grading. This study did not take into account the level of fidelity of implementation with standards-based grading. Although science classes were used as a comparison group to measure the differences in a traditional grading approach and standards- based grading approach with the exact same students, teachers most likely varied in their fidelity of implementation of PP and specifically standards-based grading. As implementation data was not available, this research study could not account for these differences.

A second limitation of the study was the lack of correspondence between KCCT assessments and the tested course. The KCCT assessments in mathematics and science assess content over three courses throughout a student's high school career. In mathematics, Algebra 1, Geometry, and Algebra 2 contents are part of the KCCT mathematics assessment. Within this study, the grades students received on a standards-based grading approach only evaluate students on Algebra 2 content. Therefore, a student could have successfully mastered the content in Algebra 2, but the student's standardized assessment score could have suffered because of a deficiency in a previous mathematics class. In order to truly assess the association between grades and test scores, the standardized assessment should only cover content taught in that specific course.

Despite these limitations, the results of this research study indicated that the use of standards-based grading with PP classrooms increased the association between grades and standardized test scores among students within the 11 high schools that implemented the program. Students who were more successful in the content class that used standards-based grading were more likely to score proficient on the KCCT assessment than students

evaluated on traditional grading practices. The most significant finding to refute traditional grading methods derived from the 75% of students who received above average traditional grades in their specific content class, yet scored below proficient on the corresponding KCCT assessment. When evaluated by standards-based grading, nearly twice as many students scored proficient when successful in their core content class. These findings provided strong evidence to suggest that standards-based grading approaches should be central to an educational reform movement.

IMPLICATIONS FOR POLICY, PRACTICE, AND FUTURE RESEARCH

In an age of increased accountability and high-stakes testing, the implications for practitioners are important to consider. Educational stakeholders expect schools and educators to accomplish a multitude of tasks beyond teaching subject content (Labaree, 2012). Although critics question the validity of many accountability measures (Downey, von Hippel, & Hughes, 2008), current federal and state policies hold schools accountable for every student's proficiency in core content areas. In schools identified as persistently low-achieving, educators face relocation or termination of their post (Hochbein, Mitchell, & Pollio, 2013). Even in schools not identified as persistently low-achieving, state and district initiatives have measured and evaluated the added value of teachers (Guarino, Reckase, & Wooldridge, 2012), as well as rewarded or sanctioned teachers based upon their students' performance on state accountability assessments (Podgursky & Springer, 2007).

For schools and educators to ensure the learning of state standards by all students, our evidence would suggest that a standards-based grading approach may offer a more valid method than traditional grading practices. This research demonstrated that standards-based grading demonstrated a stronger association with state accountability test results than traditional grading practices. Furthermore, results indicated that students who performed above average in their class, and were evaluated with a standards-based grading approach as part of the PP reform, performed higher on state accountability assessments than similar students assessed by traditional grading. The results support the reasoning that the grades students receive in a core content class using standards-based grading actually reflect what the students know and can demonstrate on state proficiency assessments. This suggests that grades in a standards-based assessment system more validly reflect student learning (Allen, 2005).

Policymakers have worked to legislate a reduction in the achievement gap. This gap between minority and non-minority students and at-risk and not-at-risk students has been widely discussed. Little, if any, of the focus to

reduce the achievement gap has centered on changing grading practices (Welsh et al., 2013). Our evidence from this study suggests that student performance on standardized tests is associated with the use of standards-based grading. As part of PP, nearly twice as many students scored proficient on the mathematics assessment when they experienced standards-based grading in their Algebra 2 class. Furthermore, correlations between the grades and standardized test scores of minority and disadvantaged students were greater in standards-based grading classrooms than in traditional grading classrooms. As suggested by prior researchers, standards-based grading practices might be a necessary, but insufficient initiative to reduce the achievement gap in American education (Brennan et al., 2001; Haptonstall, 2010; Lekholm & Cliffordson, 2008; McMillan, 2001).

Most importantly, both practitioners and policymakers must grapple with ways to deal with the diligent student who is unable to master the key standards to attain a passing grade. With traditional grading systems, a student who is compliant with a teacher's policies and requests, completes all assigned tasks in a timely manner, and has a good attendance and behavior record almost always passes a core content class. In contrast, with standards-based grading, additional factors have little influence on a student's grade. If teachers grade solely on standard attainment, then the student who does not attain the standard must be given a failing grade.

Standards-based grading highlights the issue of the diligent and failing student, yet implementation of standards-based grading policies also offers some beneficial resolutions. Schools might return to policies that reported multiple grades (Cuban, 1993). Grades for diligence and conduct might supplement grades for standard attainment. Moreover, grades for fundamental skills like literacy or simple computation could help educators focus remediation efforts. For example, a student's inability to comprehend complex text might contribute to the failure of a history standard. Understanding this contributing factor could help tailor a student's remediation plan.

Similarly, schools might alter policies that require wholesale retaking of a failed course. By identifying the specific standard deficiencies, schools might again provide targeted and specific remediation efforts. Such targeted remediation would enable students not only to master the material, but also return to the prior academic pace without having to miss an entire semester or year. With the emphasis on having all students be college ready based on benchmark ACT scores, schools must be willing to use standards-based grading approaches to ensure that students are truly college ready. At the same time, schools are expected to graduate all students and decrease retention rates. Policymakers and practitioners must figure out ways to measure students on attainment

of key academic standards, while still providing necessary safety nets for students unable to achieve these standards.

Finally, researchers should turn their attention toward the effectiveness of standards-based grading practices in schools. Most prior research has centered on overall grading practices, and little empirical research exists to support a movement toward standards-based grading. Researchers must build on the data within this study to establish a strong empirical research base for the widespread implementation of standards-based grading. This research requires both quantitative and qualitative analyses of the influence of standards-based grading practices on instruction and student achievement. Future research can use new end-of-course assessments to measure the association between grades and test scores within this new format. This research should also focus on the impact of standards-based grading on the instruction and achievement of minority students and at-risk students. As the standards-based grading movement continues to grow in secondary schools, researchers should explore the potential influence on the achievement gap as a result of new grading practices.

REFERENCES

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House, 78*(5), 218–223.

Baete, G. S., & Hochbein, C. (2014). Project Proficiency: A quasi-experimental assessment of high school reform in an urban district. *Journal of Educational Research.* Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00220671.2013.823371#.U5G0bvldW7o.

Balfanz, R., Herzog, L., & MacIver, D. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist, 42*(4), 223–235.

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

Brennan, R. T., Kim, T., Wenz-Gross, M., & Sipperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher assigned grades: An analysis of the Massachusetts Comprehensive Assessment System. *Harvard Education Review, 71*(2), 173–216.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35–36.

Brookhart, S. M. (1993). Teachers grading practices: Meaning and values. *Journal of Education Measurement, 30*(2), 123–142.

Brookhart, S. M. (1994). Teachers grading: Practice and theory. *Applied Measurement in Education, 7*(4) 279–301.

Burks, J. C. & Hochbein, C. (2013). The students in front of us: Reform for the current generation of high school urban high school students. *Urban Education.* Retrieved from http://uex.sagepub.com/content/early/2013/10/22/0042085913507060.abstract.

Cizek, G.J., Fitzgerald, S. M. & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159–179.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: Government Printing Office.

Conley, D. T. (2000, April). *Who is proficient: The relationship between proficiency scores and grades.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cross, C. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53–72.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890–1990* (2nd ed). New York, NY: Teachers College Press.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education, 81*(3), 242–270.

Duke, D. L. (1995). *The school that refused to die.* Albany, NY: State University of New York Press.

Grant, G. (2009). *Hope and despair in the American city: Why there are no bad schools in Raleigh.* Cambridge, MA: Harvard University Press.

Guarino, C., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* (Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit, No. 6602). Bonn, Germany: Institute for the Study of Labor.

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measures: Issues and Practice, 26*(1), 19–27.

Guskey, T. R. (2009*). Practical solutions for serious problems in standards based grading.* Thousand Oaks, CA: Corwin Press.

Haptonstall, K. (2010). *An analysis of the correlation between standards-based, non-standards-based grading systems and achievement as measured by the Colorado student assessment program (CSAP)* (Unpublished doctoral dissertation). Capella University, Minneapolis, MN.

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as a SES measure in education research. *Educational Researcher, 39*(2), 120–131.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin.

Hiss, W. C., & Franks, V. W. (2014). *Defining promise: Optional standardized testing policies in American college and university admissions*. Arlington, VA: National Association for College Admission Counseling.

Hochbein, C., Mitchell, A., & Pollio, M. (2013). The influence of AYP as an indicator of persistently low-achieving schools. *NASSP Bulletin, 97*(3), 270–289.

Jacob, B. A. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761–796.

Jefferson County Public Schools (2011a). *Project proficiency guide*. Unpublished manuscript.

Jefferson County Public Schools (2011b). *Student progression, promotion, and grading*. Louisville, KY: Author.

Kentucky Department of Education. (2010). *No Child Left Behind (NCLB) interpretive guide 2010*. Retrieved from http://www.education.ky.gov/nr/rdonlyres/0a2e4cd2-7b79-476c-a16a-33415da5e2fe/0/2010_nclb_interpretive_guide.pdf

Labaree, D. F. (2012). *Someone has to fail: The zero-sum game of public schooling*. Cambridge, MA: Harvard University Press.

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school levels: Effects of gender and family background. *Educational Research and Evaluation, 14*(2), 181–199.

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

Marzano, R. J. (2010). *Formative assessment & standards-based grading*. Bloomington, IN: Marzano Research Laboratory.

McMillan, J. H. (2001). Secondary teacher's classroom and grading practices. *Educational Measurement: Issues and Practices, 20*(1), 20–32.

McMillan, J. H., & Nash, S. (2000, April). *Teacher classroom assessment and grading practice and decision making*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Moore, K. A. (2014). *Making the grade: Assessing the evidence for integrated student supports*. Bethesda, MD: Child Trends.

National Collegiate Athletic Association (n.d.). *NCAA Eligibility Center quick reference guide*. Indianapolis, IN: National Collegiate Athletic Association.

No Child Left Behind Act, 20 U.S.C. § 6319 (2001).

Ogbu, J. U. (2003). *Black American students in an affluent suburb: A study of academic disengagement*. Mahwah, NJ: Erlbaum.

Persistently Low-Achieving School and School Intervention Defined. Kentucky Revised Statutes 160.346 (2010).

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909–949.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*, 1394–1425.

Reese, W. J. (2005). *America's public schools: From the common school to "No Child Left Behind."* Baltimore, MD: Johns Hopkins Press.

Rosenbaum, J. E. (1997). College-for-all: Do students understand what college demands? *Social Psychology of Education, 2*(1), 55–80.

Sanders, M. G. (2012). Achieving scale at the district level: A longitudinal multiple case study of a partnership reform. *Educational Administration Quarterly, 48*(1), 154–186.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin.

Spillane, J. P., Diamond, J. B., Walker, L. J., Halverson, R., & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. *Journal of Research in Science Teaching, 38*(8), 918–940.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14.

Tyack, D., & Hansot, E. (1990). *Learning together: A history of coeducation in American public schools.* New Haven, CT: Yale University Press.

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

MARTY POLLIO is the Principal at Jeffersontown High School in Louisville, Kentucky. His research interests include high school reform and educational policy. Recent publications include: Hochbein, C., Mitchell, A., & Pollio, M. (2013). The influence of AYP as an indicator of persistently low-achieving schools. *NASSP Bulletin, 97*(3), 270–289.

CRAIG HOCHBEIN is an Assistant Professor of educational leadership at Lehigh University. His research interests include the longitudinal process of school change and the effectiveness of policies intended to improve schools. Recent publications include: Hochbein, C., & Cunningham, B. (2013). An exploratory analysis of the longitudinal impact of principal change on elementary school achievement. *Journal of School Leadership, 23*(1), 64-90; and Hochbein, C. (2012). Relegation and reversion: A longitudinal examination of school turnaround and downfall. *Journal of Education for Students Placed At-Risk: Special School Turnaround Issue, 17*(1-2), 92–107.