ABSTRACT

Ruth Steidinger, A STUDY TO DETERMINE THE CONCURRENT VALIDITY BETWEEN COURSE GRADES AND STANDARDIZED ASSESSMENT RESULTS (Under the direction of Dr. Lynn Bradshaw). Department of Educational Leadership, August 2011.

This study sought to determine the concurrent validity between course grades and performance on standardized test in classrooms where nonachievement factors were separated from the reporting of achievement factors in a teacher's grading system.

In reviewing the literature, multiple studies indicated that the greatest culprit in the distortion of grades was the inclusion of non-achievement factors (e.g. penalties for late work, extra credit for bringing in supplies, effort or lack thereof, attendance). Therefore, this study examined the relationship between teachers' grading practices and the correlation established between their students' grades and their performance on the standardized test.

High school science teachers from two comparable schools were selected to participate in this study. They were selected because they claimed to have a grading system that separated nonachievement factors from achievement factors and they taught courses that culminated in a standardized assessment. The teachers participating in this study were asked to complete a survey to assess how well their grading system aligned with what measurement experts recommended, with regards to the removal of nonachievement factors. Results from each teacher's survey were compared to the correlations determined between his/her student's grades and performance on the standardized assessment.

Findings for all teachers but one revealed a strong relationship between students' grades and their performance on the standardized assessment. However, when looking at whether or not a teacher's grading system had an impact on the strength of the correlation, it was determined that no association of significance existed.

The outcomes of this study furthered the research on effective grading systems. Based on the findings, there continues to be a strong need for effective professional development on how to establish a valid and reliable grading system. Furthermore, issues on how to build an effective assessment were also revealed. Further studies addressing either topic are warranted.

A STUDY TO DETERMINE THE CONCURRENT VALIDITY BETWEEN COURSE GRADES AND STANDARDIZED ASSESSMENT RESULTS

A Dissertation

Presented to

The Faculty of the Department of Educational Leadership

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

by

Ruth Steidinger

August, 2011

A STUDY TO DETERMINE THE CONCURRENT VALIDITY BETWEEN COURSE GRADES AND STANDARDIZED ASSESSMENT RESULTS

by

Ruth Steidinger

APPROVED BY:

DIRECTOR OF DISSERTATION:_____
Lynn Bradshaw

COMMITTEE MEMBER:_____
William Rouse, Jr.

COMMITTEE MEMBER:_____
Lane Mills

COMMITTEE MEMBER:_____
James McDowelle

INTERIM CHAIR OF THE DEPARTMENT OF EDUCATIONAL LEADERSHIP:

_____
William Rouse, Jr.

DEAN OF THE GRADUATE SCHOOL:

_____
Paul Gemperline

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Since the late 1800s, schools in the United States have been in the practice of assigning grades to students to communicate their level of performance as it related to their academic studies (Kirschenbaum, Simon, & Napier, 1971). Parents, teachers, school administrators, and postsecondary institutions have come to expect a method for reporting a student's academic progress. Researchers have asked teachers, parents, and school administrators "What is the purpose of grades?" and they have found the responses to typically be one or all of the following (a) to communicate student achievement to students and parents, (b) inform post-secondary institutions and potential employers, (c) aid in educational and career planning, (d) provide information that students can use for self-evaluation, (e) evaluate the effectiveness of instructional programs, and (f) guide administrative decision such as a grade and class placement, graduation, promotion, and honors (Feldmesser, 1971; Frisbie & Waltman, 1992; Guskey, 2002; Linn, 1983). These findings suggested that grades play a significant role in the academic life of a student. In 2004, Brookhart found that grades have been shown to have strong and lasting effects on students' attitudes, behavior, and motivation to learn. This finding suggested that a student's grade can influence the way he/she approaches his/her schooling.

With the implementation of more stringent state and federal accountability measures, state-mandated tests and/or high-stake tests have grown in their significance in the past decade. Such tests have been used as a standardized measure of a student's academic achievement since the late 1970s/early 1980s (Airasian, 1988). Because standardized assessments impose certain controls on testing and are given under uniform conditions, they are considered by some researchers to be a meaningful basis for evaluating performance (Hopkins, Stanley, & Hopkins, 1990). Funds tied to legislation rely on these test scores (e.g. No Child Left Behind and Race to

the Top). Many districts require a student to score proficiency in several of these standardized assessments in order to graduate.

Measurement experts agreed that grades can be a reliable and valid tool for communicating a student's academic progress if the grading system was derived from valid assessment practices (Green & Emerson, 2007). In order for this to occur, measurement experts recommended that nonacademic factors such as attendance and behavior be reported separately (Friedman & Troug, 1999; Guskey, 2001; Guskey, 2009; Marzano, 2000; McMillan, 1999; O'Connor, 1995). Despite these recommendations, researchers have found that teachers continue to include nonachievement factors into their grading systems (Baron, 2000; McMillan, 1999; McMillan & Nash, 2000). No matter which method a teacher used for grading, "assigning grades and reporting student learning is an inherently subjective process" (Guskey, 2001, p. 32).

Statement of the Problem

In the United States, grades have continued to be the universal symbol for communicating a student's academic progress; however the development of a teacher's grading system has not been a consistent process. Historically, the creation of one's grading system has been a very private practice and, in most systems, unchanged since the late 1800s and early 1900s (Brookhart, 2009). With loose guidelines outlined in most districts' policies; teachers have been given the freedom to develop their own criteria for reporting grades resulting in panoply of grading systems within a school and within a department.

In North Carolina, teachers were required to sign a contract obligating them to abide by the rules and regulations set forth by their local Board of Education (see Appendix A). In Wake County, local Board of Education policy required teachers to align their instruction with the North Carolina Standard Course of Study (see Appendix B). What was also true in North

Carolina was that beginning in the 2006-07 school year high school students have been required to pass the North Carolina End of Course (NC EOC) test in five subjects in order to obtain a high school diploma. On October 7, 2010, the North Carolina State Board of Education removed this requirement for at the state level but gave local LEAs the option to maintain the requirement. Wake County Public Schools retained the passing of the NC EOCs as a requirement for graduation, increasing the significance of this assessment. The North Carolina End of Course tests were standardized tests designed to measure a student's mastery of the standard course of study for that particular course. If teachers were in fact fulfilling their contractual obligations and the grades they administered were a reflection of the student's learning of the content taught, the grades they administered should be a reliable predictor of a student's performance on the NC EOC.

Need for the Study

In March of 2010, Wake County Public School's Evaluation and Research department conducted a study exploring the relationship between high school course grades and exam scores. The study examined the trends in the percentage of students who failed the NC EOC test for Algebra I but passed the course and percentage of students who passed the NC EOC test for Algebra I but failed the course. The study also reviewed these same findings for the English I NC EOC and student course grades. The trends for both courses were observed over a four-year period. The study found that the alignment between NC EOC performance and course grades was stronger for Algebra I than for English I. Students had a higher likelihood of passing English I with a failing NC EOC test than for Algebra I. The relationship between course grade and exam grade evolved over the four-year period, with its growing stronger in one aspect (fewer students failing the test while passing the course) but weaker in the other (more students passing the test

but failing the course). These results suggested several things but of relevance to this study it suggested that the current methods for grading were not always a reliable predictor of student's performance on the NC EOC. The results of this study have also provided additional motivation for stakeholders of WCPSS to revisit the current grading system policy. A significant factor in which stakeholders disagreed was in the reporting of nonacademic factors. While all stakeholders believed in the value of reporting nonacademic factors, they differed in how they should be reported. There were two schools of thought (a) nonacademic factors should be included in a grading system; and (b) nonacademic factors should be reported separately. Additional research was needed to determine if one approach provided more accurate grades, as compared to NC EOC performance, than the other approach.

Purpose of the Study

This study was a follow up to the March 2010 study conducted in Wake County. At the conclusion of this study the question was raised of how much the *hidden curriculum* (student behaviors and attitudes) influenced the calculation of a student's final grade. This study sought to answer this question by examining the grading systems of teachers who chose to report nonacademic behaviors separately and the relationship of the student's grade to the NC EOC score. The researcher sought to identify a more reliable grading system in an effort to provide constituents with a more dependable predictor of performance on the NC EOC.

Research Questions

To explore this issue, this study sought to answer the following questions:

1. Of the grading systems reviewed which teacher's grading system aligned more closely with research best practice?

2. Did teachers' grading systems with the greatest alignment to research best practice produce a stronger relationship between students' grades and their NC EOC performance?

The null hypothesis tested in this research was teachers grading systems that were strongly aligned with research best practice will not make a difference in the strength of the relationship between their students' grades and their NC EOC performance.

Limitations

This research was highly dependent on the acquisition of a group of teachers from the same school, teaching the same subject, and claiming to have grading systems that were aligned with researched best practice. Because of this, the pool of teachers was limited. There was a clear limitation in this study in that there were only eight teachers. Given this, the study would have been strengthened if it included a qualitative piece allowing more insight into the beliefs of each teacher.

The data for Teacher G was diminished significantly when it was learned that he did not maintain numerical averages for his students for the 2009-10 school year. Therefore, the researcher was limited to the 2010 fall semester data for this particular teacher.

A substantial portion of this research is built around teacher responses to a survey with no way to verify the validity of the responses. Therefore, the possibility exists that respondents may have misrepresented their practices via their responses.

The answer to the second research question was contingent upon the outcomes of the series of Fisher's Exact Tests. This analysis does not conduct individual comparisons and when the data was reviewed in table format, teacher by teacher, there was evidence that suggested a

relationship might exist for some teachers. Individual analysis of some sort might have yielded additional findings.

Lastly, the survey was piloted prior to the year this study began. However, after the review of the literature modifications to the survey could have been made to make it a stronger instrument yet doing so would have delayed the study; therefore, the researcher decided to proceed with the study.

Methods for the Study

In answering these questions, a validity coefficient was established by examining the relationship between teacher grades and student scores on standardized tests. Teachers were asked to complete an on-line survey to establish their methods of practice. This measure was compared to the validity coefficient to determine which teacher's grading system yielded the strongest relationship.

Organization of the Dissertation

Chapter 1 provided an introduction to this study which explored the relationship between a student's grade and his/her NC EOC score and how a teacher's method of grading impacted that relationship. The guiding theoretical framework for this study was based in the assumption that grading systems that removed nonacademic factors from the system provided more accurate grades (Brookhart, 2009; Marzano, 2000; O'Connor, 2002). Chapter 2 synthesized the literature and research regarding grading systems and was organized by the following themes that surfaced in the review: (1) history of grades; (2) the purpose of grades; (3) recommendations of measurement experts regarding best grading practices, (4) validity and grades; and (5) constructing a valid grading system. The research design and methodology for the study were presented in Chapter 3, along with a description of the school district, two schools, measures,

and data bases used in this study. The research design for this study was quantitative. Chapter 4 consisted of the presentation of the data collected for this study. Chapter 5 presented conclusions and recommendations with regards to valid grading practices. Also included in this chapter were implications for public school policy regarding grading practices and possible areas for further study.

CHAPTER 2: LITERATURE REVIEW

Introduction

Grades have been shown to have strong and lasting effects on students' attitudes, behavior, and motivation to learn (Brookhart, 2009). This finding alone solicited the exploration of what actually should constitute a grade. In 2003 Smith found that students who aspired to pursue postsecondary education understood the impact of grades on their future decisions. As a result these students enrolled in the most rigorous courses and worked hard to attain the highest grade possible. Yet, when researchers study the reasons for students dropping out of school, poor academic performance and low grades were consistently noted as primary factors (Goldschmidt & Wang, 1999; Lee & Burkam, 2003). Each of these findings noted above were compelling reasons for educators to take a serious look at their grading systems.

Experts in the study of grading systems all agree that the primary purpose of grades should be to communicate student achievement (Brookhart, 2009; Guskey & Bailey, 2001; O'Connor, 1995). However, where the grading debate became heated was in the discussion of what that grade should represent. Should a grade reflect a student's effort, progress over time, attendance, attitude, conduct, *and* academic achievement? Should a grade only measure a child's performance against explicit standards? It would appear that given the primary purpose of grades was to communicate student achievement the response to the first question would be "no" and to the second, "yes". This was actually not the case. There are teachers, parents, and students who believed that nonachievement factors such as effort and ability were critical in developing the learning skills of students, therefore, should be reported (Guskey, 2002). Again, most stakeholders would not disagree with this statement; however, at this point the debate narrows to

"how do we report these nonachievement factors?" There were typically two schools of thought, calculate the nonachievement factors into the overall grade or report them separately.

Cross and Frary (1999) noted that when multiple factors were considered when calculating a student's grade, the message became distorted and the result was "*hodgepodge grading"* (p. 1). Ken O'Connor (1995) wrote that the most serious problem with our grades was that they try to serve so many functions that they have become almost meaningless in serving their primary purpose, communication. As a result, most systems developed policies and procedures which attempt to achieve all purposes for all stakeholders and, therefore, achieve none well (Austin & McCann, 1992).

History of Grades

It was widely accepted by researchers in the field of grading that the whole notion of assigning grades originated in higher education dating back to the mid 1600s. Smallwood's study of original documents at five institutions of higher learning (Harvard, William and Mary, Yale, the University of Michigan, and Mount Holyoke) yielded a great deal of the history of grading during this time period (Brookhart, 2009). A formal issuance of grades was first noted in the late 1700s in the form of grading scales. The first record of a grading scale was a set of categories at Yale (Smallwood, 1935). Students were classified as *Optimi, Second Optimi, Inferiors,* or *Pejores* with the progression noted being from best to worst. From here scales moved to numerical scales of varying forms. Different colleges used various scales; for example, scales of 1-4 including decimals, scales of 20, scales of 100, and rank scales with descriptive adjectives (Smallwood, 1935, p. 43). In 1850, the University of Michigan experimented with abolishing grades altogether and introduced a pass/fail system and after a few years extended this to include gradations of failure, *conditioned* and *incomplete* (Smallwood, 1935). As colleges and

universities began to allow electives, comparability of grades became a problem and they moved to a more uniform marking system that we know today.

While the university system introduced the notion of grading, it was not until the establishing of schools for children, via the common school movement, in the 1800s that K-12 institutions began to wrestle with grading and reporting (Brookhart, 2009). Initially, teachers reported student's learning progress orally to parents and then as the number of students increased teachers began to record the skills students were to master and their progress towards mastery was noted. Students were not permitted to progress until the skill was mastered. This is the earliest example of a narrative report card (Brookhart, 2009).

With the passage of the compulsory attendance law at the elementary level in 1852, the United States began to see an increase in the number of students entering high school. As a result there was a significant increase in the number of high schools and subject area instruction became increasingly significant. This specificity led to the introduction of percentage grading to certify a student's accomplishment in a specific subject. This method of grading was the beginning of the grading and reporting system we know today (Brookhart, 2009). It did not take long for educators to begin to question the reliability of this grading system.

Beginning in 1910, there were numerous studies conducted examining the reliability of grades. The studies of Starch and Elliott (1912) were probably the most famous from this time period. Starch and Elliott conducted a series of studies examining the reliability of percentage grading of high school examination papers in English, mathematics, and history. They set out to determine the extent to which a teacher's personal values and expectations influenced their grading. The first study conducted was the English study. They made 200 copies of two English examination papers written by students at the end of their first year of high school. The copies

were sent to 200 English teachers who taught first-year English and they were asked to grade the papers. Probably the most shocking finding in this study was that the range of variability in grades for the same paper was as large as 35 or 40 points. In addition to this Starch and Elliott found that different teachers focused on different aspect of the writing. For example, some reported an emphasis on neatness, spelling, and punctuation while others placed emphasis on how well the message was conveyed. Immediately, this study came under fire by critics who believed that writing was a highly subjective judgment. To counter this criticism, Starch and Elliott conducted the exact same study but used geometry papers and submitted them to mathematics teachers. When they analyzed the results they found the variation amongst mathematics teachers to be even greater than the English teachers. There were variations as great as 67 points (Starch & Elliott, 1913).

In the early 1900s, there was a movement away from the use of percentage grading in large part due to the wide variations in grading found in the Starch and Elliott studies. There was a move towards the use of symbols and categories with larger scales and fewer options. For example a three point scale using *Excellent, Average,* and *Poor* were used as measures of performance. Also, during this time period there was the introduction of the categories of *A, B, C, D,* and *F* (Johnson, 1918; Rugg, 1918). While the introduction of these categories reduced the variation in grades it did very little to solve the problem of teacher subjectivity. In an effort to resolve the issue of fairness, norm-reference testing was introduced. Educators believed that by evaluating students via a comparison of student performance, grades would be less subjective in nature and there would be a fairer distribution of grades (Brookhart, 2009). This was known as *grading on the curve*, referencing the bell-shaped curve distribution of grades.

In the 1920s and 1930s, researchers such as Herron and Proffitt established the fact that teachers consciously included nonachievement factors such as effort and attitude into their grade calculations. During this same time period, Crooks (1933) published a paper establishing an argument against the use of norm-referenced grading and advocated for the use of measurement against an absolute standard. This became known as *criterion-referenced* grading 30 years later when Glaser introduced the term in his landmark article in 1963.

From 1930 to 1960 we saw the introduction of the *scientific* multiple choice tests (Brookhart, 2009). Elementary schools were fairly successful in implementing standards-based grading but high schools were still grounded in norm-referenced grading due to the use of high school grades in college admission. It was believed that sorting students according to their academic accomplishments was helpful information for college admissions (Brookhart, 2009).

The decades between 1960 and 1980 were notorious for periods of student unrest (Brookhart, 2009). The research around grading mirrors this time period. One of the most noted pieces of research during this time was that of Kirschenbaum et al. (1971), *Wad-ja-get? The Grading Game in American Education.* The authors of this work were seeking to answer the question, "Is the traditional system of grading- the one most of us experienced throughout many years of schooling- the most educationally useful system of evaluation?" (p. 14). They conducted a comprehensive literature review but decided to report their findings in the form of a novel. The drama of Mapleton High School highlights many of the concerns found in real situations. At the conclusion of their novel Kirschenbaum et al. (1971), offered several examples of alternative grading systems that addressed some of the issues they brought to light in their novel (e.g. written evaluation, self-evaluating, contract systems, and mastery approach or performance

curriculum). The authors described each alternative system noting the advantages and disadvantages of each (see Appendix C).

With the 1980s came a concern of teachers' competence and their ability to adequately measure and assess the skills of students. This concern regarding teachers' ability to adequately measure student skills raised an additional concern of validity. These concerns brought about a new round of research on grading practices with teacher competency as the backdrop. Stiggins, Frisbie, and Griswold's (1989) study brought to light again the concern of teachers' use of nonachievement factors in grading, an issue raised in the 1920s. They analyzed the grading practices of 15 high school teachers against the recommendation of measurement experts. Their findings suggested that teachers' use of nonachievement factors in grade calculation brought to question the validity of the grade (Stiggins et al., 1989).

With the turn of the 21st century came the signing into law of the No Child Left Behind Act (NCLB). NCLB required states to establish content standards and to report student achievement on these standards. As a result, the most important change in grading during this time was a shift towards standards-based report cards. This method of grading was more prevalent in elementary schools than in high schools. Standards-based reporting brought its own set of challenges, communication being one of great concern (Guskey, 2004, 2008).

Purpose of Grades

Much of the dilemma around how to grade stemmed from the confusion regarding the purpose of grades (Brookhart, 2009; Guskey & Bailey, 2001; Marzano, 2000). Frisbie and Waltman (1992) noted that grades would not be a necessary part of educating students if it were not for the need to communicate to parents or other stakeholders. "Grades are not essential to the instructional process: teachers can teach without them and students can and do learn without

them" (p. 35). In a paper presented at the annual meeting of the American Educational Research Association, Feldmesser (1971) contended that grades do provide unique and useful information to the student and that they stimulate the making of other kinds of evaluations (e.g. grade placement, class ranking, honors courses). He found that when professors were allowed to evaluate work on a pass/fail method of grading, students complained that they received very little feedback as opposed to those students taking the course for a letter grade (Feldmesser, 1971, p. 6).

Researchers have asked teachers, parents, and students to identify the purpose of grades and their findings can be classified into six broad categories (Austin & McCann, 1992; Feldmesser, 1971; Frisbie & Waltman, 1992; Guskey, 2002) (a) to communicate the achievement status of students to parents and others, (b) to provide information that students can use for self-evaluation, (c) to select, identify, or group students for certain educational paths or programs, (d) to provide incentives for students to learn, (e) to evaluate the effectiveness of instructional programs, and (f) to provide evidence of students' lack of effort or inappropriate responsibility.

In 2002, Guskey sought to determine similarities and differences in teachers', students', and parents' perceptions of the purposes of grading and reporting. He had participants from all three groups rank order the six major purposes of grades noted above. The opinions expressed by all three groups were very similar. Each group ranked "communication to parents" and "feedback for students" as the two most important purposes and "lack of effort and responsibility" as the least important with "evaluate school programs" close behind (Guskey, 2002). The greatest variability in response was amongst the teacher group. Parent and student groups tended to be similar in their rankings. Guskey then evaluated the responses as they

progressed through the grades. He found that the higher the grade level the less important teachers found grades to be an incentive; yet, parents and students ranked the incentive value as increasingly important (Guskey, 2002). The lack of effort and responsibility became less important to teachers and parents at higher grades, but students ranked it as more important. Lastly and surprisingly, as the grade level increased, teachers, students, and parents tended to rank communication to parents less important but feedback to students as more important purpose. Even with these shifts in perspectives, it was evident that communication regarding the student's achievement, be it parents or students, was identified as an essential purpose of grading (Guskey, 2002).

When trying to identify the purpose of grades, researchers suggested that the competing purposes were problematic. Austin and McCann (1992) pointed out that too often school systems would work to develop a single policy that strives to achieve all purposes for all stakeholders and consequently, would achieve none well. Brookhart (2009) noted that "It is very difficult for one measure to serve different purposes equally well (p. 21)" and that "The main difficulty driving grading issues both historically and currently is that grades are pressed to serve a variety of conflicting purposes" (p. 31). Austin and McCann (1992) suggested that school systems should consider the development of multiple grading systems, "each with its own integrity and each serving well a given purpose" (p. 16).

Guskey (2007) suggested that there was a primary purpose and then there were secondary purposes in grading. Bailey and McTighe (1996) supported this suggestion in defining the primary purpose "to communicate student achievement to students, parents, school administrators, post-secondary institutions and employers" and that all other factors are secondary. For example, "providing teachers with information for instructional planning,… and

providing teachers, administrators, parents, and students with information for selection and placement of students" (Brookhart, 2009, p. 5) were noted by Brookhart as secondary purposes. Because these purposes can be in conflict with one another, Guskey (2007) also supported the idea of having separate measures for different purposes.

Grading Practices

If we considered the primary purpose of grades communicating student achievement to (a) students, (b) parents, (c) school administrators, (d) post-secondary institutions, and (e) employers, the next question to tackle was, what should a grade represent to effectively communicate this purpose? The opposing views of what a grade should reflect were noted in the conflicting opinions of the purpose of schooling. There were those that believed that the primary job of schools it to help students master certain knowledge and skills. This group wanted the grade to define the current status of a student's achievement against an explicit standard. Austin and McCann (1992) referred to this as measurement against *Achievement factors*. Others believed that the primary job of a school was to be developmental and wanted grades to reflect the effort and progress students were making – *Nonachievement factors* (Austin & McCann, 1992). Bishop (1992) suggested that the dual roles of advocate (nonachievement factors) and judge (achievement factors) were not compatible. He recommended that teachers yield the judging role to external assessment in order to function more fully as advocates, coaches, and mentors to their students (Bishop, 1992).

Achievement Factors

Tom Guskey, Doug Reeves, and Ken O'Connor were some of the current researchers who wrote articles indicating the need to have policies that required teachers to calculate grades solely on academic performance. Reeves (2008) wrote, "To reduce the failure rate, schools don't

need a new curriculum, a new principal, new teachers, or new technology. They just need a better grading system" (Reeves, 2008, p. 1). Yet, Cross and Frary (1999) noted in their research that even teachers who may agree with a grading system that only reflect academic performance, they continue to struggle to keep nonachievement factors out of their grading practices. In their study, Cross and Frary (1999) reported that measurement specialists recommended that achievement (what students know and are able to do) should be reported separately from other factors. When they asked teachers and students in their study if they agreed with this statement, 81% of the teachers and 70% of the students agreed or tended to agree with this statement (Cross & Frary, 1999). Yet, the practices of the teachers agreeing with this statement were not consistent with what they claimed to believe. Seventy-two percent of all the teachers in this study indicated that they raised the grade of low-ability students for factors other than achievement (Cross & Frary, 1999).

In Friedman and Troug's (1999) study they sought to learn how well 53 high school teachers' grading policies matched the recommendations of assessment specialists. While 77% of the teachers included achievement as a characteristic when grades were assigned, no one completely separated achievement from nonachievement factors (Friedman & Troug, 1999). Teachers agreed with achievement being the primary factor in calculating a student's grade, but they felt strongly that effort, behavior, and attendance should also be considered.

Ornstein (1994) noted that while recommendations may suggest including only achievement factors into a grade, teachers lacked the skills necessary to create objective test and to interpret the results. "Assigning grades to students' schoolwork is inherently subjective, regardless of the method used" (Ornstein, 1994, p. 3).

Because teachers were concerned with motivation, self-esteem, and the social consequences of giving grades, using student achievement as the sole criteria for determining grades was rare.

Nonachievement Factors

Nonachievement factors were considered by many to be value judgments, laden with subjectivity (Brookhart, 1993; Cross & Frary, 1999; Friedman & Troug, 1999; Guskey, 1994). To illustrate this point Brookhart (1993) asked the teachers in her study to assign a grade to two students based on different circumstances. David was a student in an average Algebra I class. The grade in this class was based on two tests for a grading period. David made an F on this first test and on the second test he obtained a low D. The teachers were asked to assign David a grade for the grading period. The choices were an overall F based on the average of the two exams or an overall D because he showed improvement in his performance. Seventy- three percent of the teachers chose the latter grade. The second situation was similar but Bernie obtained a B on his first test and a low A on his second test. The choices for a final grade were an overall grade of B which is the average of the two tests or an overall grade of A noting that there was improvement in his performance. In this scenario the same percentage of teachers chose the first grade, a B (Brookhart, 1993). While those with measurement instruction were more likely to be represented in the percentage that chose the average in both scenarios, the majority with or without measurement instruction chose to let nonachievement factors influence the way they graded one student but not the other.

Zoeckler (2007) focused his study on English teachers. He explored the moral aspects of grading, noted earlier as nonachievement factors. While the teachers resisted the notion that they made broad judgments about the character of their students when it came to grading them, all

admitted to allowing their judgment of a child's character to influence their decision between passing and failing a student when it came to assigning a grade at the end of a marking period (Zoeckler, 2007).

McMillan and Lawson (2001) found that teachers conceptualized two major ingredients when grading students: academic achievement (achievement factors) and effort, ability, and improvement (nonachievement factors). In reviewing the teacher ratings, it was clear the factors related to academic achievement were most important; however, the nonachievement factors noted did not differ much from the mean value of the achievement factors (McMillan & Lawson, 2001). When taking a closer look at the distribution of scores within these two areas, the variability was great. For example, when asked how much a teacher factors the ability level of a student into his/her grade the ratings were (a) 12 teachers – not at all, (b) 14 teachers – very little, (c) 25 teachers – some, (d) 29 teachers – quite a bit, (e) 17 teachers – extensively, and (f) 4 teachers – completely. This finding suggested that two students could take the same course with different teachers, perform equally as well on tests, yet because they have different ability levels they could end up with different grades. How much a student's ability factors into the teacher's grade was completely contingent upon that teacher's practice. This variability also existed when teachers rated their grading practices with regards to how much improvement and effort contributed to student grades. Ornstein (1994) suggested that allowing improvement to factor into a student's grade created an unfair advantage to low-achieving students as they have more opportunity to improve by simple regression to the mean.

Friedman and Troug (1999) also invited teachers to participate in a focus group. They included parts of their discussion in their study. When asked about how effort should factor into a grade, most of the focus group agreed that borderline cases should be decided using effort, but

only in the cases of students who were earning Ds or Fs (Friedman & Troug, 1999). This belief supported the findings in Brookhart's 1993 study. Teachers in Friedman and Troug's focus group also showed great support for including student behavior and attendance in their grading practices.

The review of this literature brought to light the difficulty educators have in removing nonacademic factors from grade calculation. Even in situations where teachers stated their belief against such a practice, they were found to include effort when faced with grading a low-ability student (Cross & Frary, 1999; Zoeckler, 2007).

These findings supported Austin and McCann's (1992) belief in the need for multiple grading systems, each with their own purpose. In this case, a system that measured academic performance against an explicit standard and a system that measured nonacademic performance against an explicit standard would be warranted.

Validity and Grades

Cross and Frary (1999) brought to light that even teachers who believed in a grading system that only reflected academic performance have difficulty with removing nonacademic factors. The inability to remove nonacademic factors introduced the concern of validity and teacher bias.

Historically, teachers' grading practices have been a private practice in which they have had total control. While the symbols teachers used may be the same, when the process for determining a grade varied from teacher to teacher the validity of the grade diminished, in large part because of the introduction of teacher biases. A simple example can illustrate this point. If you wanted to know if your students had mastered the concept of solving one-step equations you would not have assessed them with a test asking them to solve two-step equations. This would be

an invalid measure. If you assessed them using a test with questions involving one-step equations you would have created a valid measure. In scoring your valid test, if you marked a student's paper lower because you could not read his/her answers you would have just introduced your bias into the scoring of this assessment. It was no longer a valid measure. "Subjectivity in grading becomes detrimental to students when it translates to bias" (Guskey & Bailey, 2001, p. 33).

There were numerous studies illustrating the impact teacher's biases may have had on students academic performance. For example, Hills (1991) found that teachers' perceptions of students' behavior can significantly influence their judgments of academic achievements. In 1993, Bennett, Gottesman, Rock, and Cerullo reinforced this finding in determining that students with behavior problems rarely received a high grade regardless of their performance because their behavior problems overshadowed their performance. This was found to be particularly true for boys. Sweedler-Brown (1992) found that the neatness of a students' handwriting has been shown to have a significant impact on a teachers' judgment of the students' work. Lastly, in 1997, Spear found that a good paper tended to be graded more favorably if it followed a poor paper than if it preceded it. Spear (1997) also found that if the good paper followed two or more poor papers the evaluation of that good paper was even better.

In 1989, Samuel Messick published an article that introduced researchers to a new concept of validity, *consequential validity*. Consequential validity placed greater emphasis on the use of the score (grade) and the consequences of its use than earlier theories of validity. This article generated a flood of further classifications of validity. Of these classifications, construct validity (the test score does not equate with the construct it attempts to tap (Messick, 1989) was considered to be the most important aspect of validity in that it was the umbrella for all other

aspects. Criterion and content-related validity were also identified (Hopkins et al., 1990; Shepard, 1993). In 1995 Messick further investigated this concept of construct validity and identified two forms of errors that would have a negative impact on the interpretation of test scores or grades; construct irrelevant variance and construct under-representation. Construct validity was compromised by construct irrelevant variances. An example of this would be when a teacher has taken into consideration extraneous factors such as effort, ability, or behavior when assigning a grade, distorting the message of communicating what the student has learned. In other words, there was a disagreement in the student's knowledge and the grade they received. One would expect those to be strongly correlated if the method for assessing was valid. Construct under-representation was represented when a teacher used limited information to determine a student's grade. For example, a teacher used one final summative grade to report all that a student had learned for an entire semester. Smith (2003) would argue that this was potentially an issue of reliability as oppose to validity in that this assessment may not have provided enough information for the teacher to make a reasonable decision about this student's knowledge with regard to this realm of information.

As noted earlier, under the construct validity umbrella resided criterion-related validity. Concurrent validity has been identified as a subset of criterion-related validity (Hopkins et al., 1990) and was relevant to this review and study. Concurrent validity was present when one variable was highly correlated with another variable of which one would expect (e.g. a student's grade was strongly correlated to his/her NC EOC score).

As a review of the research indicated, it was difficult for teachers to remove nonachievement factors when grading. Even studies where teachers believed that grades should only be based on achievement factors, they were still found to include factors of

nonachievement. Willingham, Pollack, and Lewis (2002) stated that it was reasonable to assume that behaviors had an impact on student achievement and methods for influencing student behavior should be sought in order to positively impact student achievement. However, in their study they found that unless artifacts such as attendance and class behavior are taken into account when measuring validity "common statistics will give a misleading and unduly conservative picture of validity and fairness" (Willingham, Ollack, & Lewis, 2002, p. 28). The arguable challenge in the review of the literature was how to craft a valid grading system.

Constructing a Valid Grading System

Ken O'Connor (1995) proposed the following guidelines for constructing a valid grading system. These guidelines originated with Gronlund and Linn (1990) but have been modified considerably with Rick Stiggins having the greatest influence (O'Connor, 1995). O'Connor (1995) also noted that these guidelines were "consistent with the fair assessment practices codes proposed for use in both the United States and Canada" (p. 96). These guidelines framed the review of what constitutes a valid grading system.

Limit the Valued Attributes Included in Grades to Achievement

O'Connor (1995) acknowledged that there was still considerable professional judgment involved in the development of a grading system and in the marking of assessments. However, attributes such as effort, participation, attitude, and other personal/social characteristics should be reported separately. This was a view supported by most measurement experts (Brookhart, 2009; Guskey & Bailey, 2001; Stiggins, 2000; Trumbull & Farr, 2000). Inclusion of nonachievement factors in a grading system was one of the most contentious points of discussion amongst stakeholder groups. As Marzano (2000) pointed out, most groups agreed that nonachievement factors should be included. He references two reports conducted by the Secretary's Commission

on Achieving Necessary Skills (SCANS) focused on identifying the skills that the business community, from blue collar to white collar, believed were necessary to be successful in the workplace. Both reports identified *life skills* such as effort, behavior, attendance, personal responsibility, working with others, and self-efficacy as essential for workplace success (Marzano, 2000). He also identified a poll conducted by the polling firm Public Agenda where parents were asked about what they believed should be taught in school. Their findings were reported in *First Things First: What Americans Expect from Public Schools* (Farkas, Friedman, Boese, & Shaw, 1994). The results of the survey indicated that 88% of those surveyed believed that schools should teach such skills as punctuality, dependability, and self-discipline (Marzano, 2000, p. 38). In another poll conducted by the Gallup Corporation, adults were asked to identify the competencies students should be required to master before high school graduation. The category of *life skills* (self-regulation and working well in groups) surpassed 13 traditional subject areas in the ranking (i.e. mathematics, science, history, language arts, physical education) (Marzano, 2000). So, was there a place for nonachievement factors in a grading system? Unequivocally, Yes, but where?

While separating nonachievement factors from achievement factors in a grading system would not completely eliminate the influence of teacher biases in grading, it was a start. In his work with teachers across the country, Robert Marzano (2000) had identified three nonachievement factors that seemed to resonate pretty strongly with teachers; effort, behavior, and attendance. Because teachers defined these factors slightly different, Marzano (2000) had suggested subdividing these factors (see Figure 1).

In his book *Transforming Classroom Grading,* Marzano (2000) went on to further define each of these subcategories to give teachers clearer direction on what was to be evaluated.

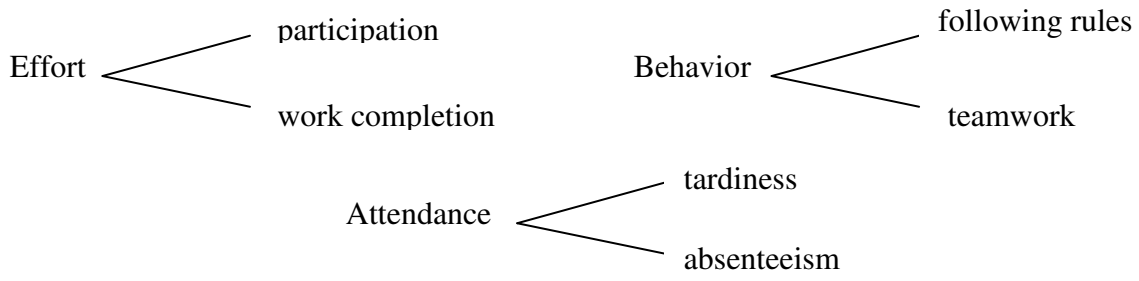

Figure 1. Sub-categories of nonachievement factors.

Because these factors were observables, researchers cautioned that for these informal observations to be useful, teachers needed to be real clear on the skills to be observed (e.g. use of rubrics). Brookhart (1994) indicated that most of our current grading systems did not take into account the teacher's need to manage classrooms and motivate students. The findings of William, Pollack, and Lewis (2002) supported the need for measures of "scholastic engagement" (p.26) as they did have an impact on student achievement. Teachers have known this to be true and consequently have constructed their own systems bringing into question the validity of their measures. Marzano (2000), Brookhart (2009), Guskey and Bailey (2001), and O'Connor (2002) were just a few of the researchers who provided examples of how to create rubrics/grading systems that assisted teachers in effectively measuring nonachievement factors in order to maintain the integrity of their grading system (see Appendices D,E, F, and G).

Sample Student Performance – Don't Mark Everything

This guideline supported the use of formative and summative assessments. As Brookhart (2009) noted, "Formative assessment means information gathered and reported for use (by the learner) in the development of (his/her) knowledge and skills, and summative assessment means information gathered and reported for use in judging the outcome of that development" (p. 45). In considering these definitions, one could have equated formative assessment with feedback given to the learner.

In 1992, John Haitte reviewed 7,827 studies from the past 30 years and generated a synthesis of the effects of schooling. He reported three major findings and of these three, feedback was shown to be the most powerful enhancer of achievement. "The effect-size for reinforcement was 1.13, remediation and feedback .65, mastery learning (which is based on

feedback) .50; more specifically, homework with feedback is much more effective than homework without feedback" (p. 9).

Black and William (1998) performed a meta-analysis of over 200 professionally reviewed literature documents on formative assessments. They found that improved formative assessment strategies resulted in an effect size growth between 0.4 and 0.7 standard deviation points. They also found that the effect on low-achievers was greater than that of high achievers. Lastly, students who were given explicit feedback performed better than those who were only given marks or grades.

Valerie Shute (2008) conducted a review of the research on formative feedback. She defined formative feedback as "information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning" (p. 154). In her review she identified some key elements to consider when providing formative feedback. Shute (2008) identified the historical review of Kulhavy and Stock (1989) as critical in identifying the type of information that was significant if one was providing the learner with effective feedback: *verification* (confirming whether the answer was right or wrong) and *elaboration* (addresses the response and also provided guidance).

Another feature to consider was the complexity and length of the feedback (Shute, 2008). The research on this particular feature got rather complex, but in general, if feedback was too long or too complicated, most learners would simply ignore it.

It was also recommended that feedback be goal-directed as opposed to being focused on an individual task. Research has shown that learners remained motivated and engaged if they were provided feedback on how they were progressing to the attainment of their goal (Fisher & Ford, 1998; Ford, Smith, Weissbein, Gully, & Salas, 1998).

The research regarding the relationship to timing of feedback to learning was inconsistent, but Shute (2008) summarized that immediate feedback could have both positive and negative effects. “The positive effect of immediate feedback can be seen as facilitating the decision or motivation to practice and providing the explicit association of outcomes to causes. The negative effects of immediate feedback may facilitate reliance on information that is not available during transfer and promote less careful or mindful behavior” (Shute, 2008, p. 166). Those providing feedback needed to remain cognizant of the balance between aiding and enabling.

The results of a study conducted by Davis, Carson, Ammeter, and Treadway (2005) supported the positive effects of feedback with high specificity on the performance of students with both high and low ability.

In summary, Shute (2008) found that the value of formative feedback was enhanced if it communicated whether the concept has been mastered or not, provided guidance, was concise, was easy to understand, was aligned with the learner’s goal, considers timing, and was explicit.

Grade in Pencil

This guideline was built on two premises: (1) A student’s knowledge and/or ability builds throughout a course; therefore, marks earned earlier in the year should be replaced by those earned later in the year; (2) Students should be given opportunities to retest in order to demonstrate mastery of the content. While grading experts generally favored this guideline (Brookhart, 2009; Guskey, 2001; Guskey & Bailey, 2001; O’Connor, 2002), there was not a lot of research to support the influence this practice has had on student’s learning and/or grades.

Relate Grading Procedures to the Intended Learning Outcomes

As referenced earlier, Shute's (2008) research on formative feedback indicated that students remained motivated and engaged in the learning process if provided feedback directly related to their learning goals. It was the method of reporting that became critical when considering the value of that feedback. Brookhart (2009) suggested the use of criterion-referenced scores for establishing grades. She defined criterion-referenced scores as "scores that will give students information about how well they hit the achievement target (standard) represented by a particular assessment" (p. 73). Two of the most common measures used in providing feedback on assignments measured against a set of standards were points or percents and rubrics (Brookhart, 2009). Marzano (2000) stated that the point method makes sense if a teacher addressed only one topic within a grading period. Typically, multiple topics were assessed within a grading period and more often than not, within a single assessment. When grades were based on performance assessments that measured varying topics inaccurate representations of student achievement were the result. Marzano (2000) suggested "A Topics-Based Grade Book" (see Appendix H). While Marzano and Brookhart both favored the use of rubrics, some educators questioned the accuracy of the rubric method because it was assumed that this method introduced an element of subjectivity. However, as Ornstein (1994) noted, "Assigning grades to students' schoolwork is inherently subjective, regardless of the method used" (p. 57). Marzano (2000) sought to dispel this belief by conducting a study in which he had two teachers, a 4th grade social studies teacher and an 8th grade mathematics teacher, score a specific assessment twice, once using the rubric method and the second time using the point method. The most interesting finding related to this guideline was that the correlation between the rubric score and an outside standardized assessment was much higher than the correlation

between the point score and the standardized assessment: ".65 versus .31 for the 4th grade teacher, and .71 versus .42 for the 8th grade teacher" (Marzano, 2000, p. 62). This study demonstrated the presence of concurrent validity.

Criterion-referenced grading was generally considered by teachers and students to be a more equitable and fair way of grading (Kovas, 1993); however, the grading criteria used by teachers may vary significantly. To assist in this method, Guskey (2001) suggested classifying these criteria into three broad categories: product, process, and progress. "Product criteria relate to students' specific achievements or levels of performance" (Guskey, 2001, p. 21). This type of measurement sought to determine the amount of knowledge a student had at a specific point in time. "Process criteria relate not to the final results but to how students go there" (Guskey, 2001, p. 21). Process criteria provided teachers a way in which to measure nonachievement factors such as effort, work habits, and other behaviors. "Progress criteria relate to how much students actually gain from their learning experiences" (Guskey, 2001, p. 22). This method of measurement supported the belief that learning was progressive and a grade should also reflect the knowledge gained over time. Few teachers based their grading system on one method. Often what was seen was a combination within a class as well as from student to student resulting in what Cross and Frary (1999) noted as hodgepodge grading. Making meaning of a student's grade became a challenge for stakeholders when criteria were not clearly defined.

Use Care "Crunching" Numbers

Ken O'Connor (2002) cited a quote from McMann (2003) that provided an excellent illustration of the problem with using the mean, "Whenever I hear statistics being quoted I am reminded of the statistician who drowned while wading across a river with an average depth of three feet" (McMann, 2003, n.p.). Calculation of the mean can distort a student's final grade as it

does not take into consideration outliers. A very simple example can illustrate this point. Teacher A adopted a grading system that was topics-based. She administered four assessments on the topic of estimation. Student A received the following scores with the earliest dated assessment listed first and so on: 25, 80, 89, and 97. In averaging the scores the student received a 72.7 which on a seven point grading scale was a D and on a ten point grading scale a low C. Measurement experts suggested the use of alternate measures of central tendency, median or mode when appropriate. It was not suggested that the mean never be used but to be mindful of the implications of your choice when averaging students' grades and to chose the measure that accurately reflected a student's knowledge. In the example above, if the median were used to determine Student A's grade, the grade would be 84.5.

Use Absolute or Preset Standards to Distribute Grades

Glasser (1992) noted that "No student's grades should ever depend on what other students do" (p. 108). This guideline refuted the use of grading "on the curve", the bell curve. This guideline also suggested that students should be evaluated on what they have learned and were able to do (criterion-referenced standards) not their relative standing amongst their classmates (norm-referenced standards). Competing against a preset distribution of grades proved to have negative effects on the relationships among students and the relationships between teachers and students (Krumboltz & Yeh, 1996). This competitive environment also proved to have a negative impact on student performance on achievement test (Johnson, Johnson, & Tauer, 1979). Krumboltz and Yeh (1996) also noted five basic ways in which the use of the bell-curve in grading had negative impact on teachers:

1. It turned teachers into students opponents,
2. It justified inadequate teaching methods and styles,

3. It trivialized course content,
4. It encouraged methods of evaluation that misdirect and inhibit student learning, and
5. It rewarded teachers for punishing students (p. 324).

In Zoeckeler's (2007) study, he found that the younger teachers, especially those without tenure, would "fix" their grades in order to keep their failure rates within an acceptable range so as not to draw the attention of their administrator.

It was not until the latter part of the 20th century that the American economy needed a large number of highly educated people (Canady & Hotchkiss, 1989). Sorting and selecting was an acceptable practice. Later on, this was no longer the case and teachers needed to "reflect on grading policies and practices that might be counterproductive for students at all ability levels" (Canady & Hotchkiss, 1989, p. 68).

Properly Record Evidence from Quality Assessment Instruments

O'Connor (1995) wrote, "Marks and grades are meaningful only if they are based on quality assessment instruments, that is, assessment instruments that are valid and reliable" (p. 100). Stiggins (1991) defined a person capable of developing such an assessment as *assessment literate*. He wrote, "Assessment literates know if and when an assessment appropriately reflects a clearly defined achievement target" (p. 537). Stiggins (1991) continued to define an assessment literate as someone who was capable of identifying factors that may delude the integrity of the results and mislead people in positions of making decisions. They were also capable of identifying appropriate conditions for assessing as well as if the results rendered from an assessment were produced in such a way as they communicated what was intended. Classroom assessment consumed a large part of a teacher's time and effort; however, as Austin and McCann (1992) identified in their review of policies and procedure from 144 districts, very few districts

(less than 10) gave teachers adequate guidance to ensure consistent grading, and none of the districts provided staff development to teachers to improve their grading practices. With regards to assigning grades, Ornstein (1994) wrote "Not only is a high degree of expertise required for being accurate in grading, one that eludes most teachers, but there is the false assumption that teacher-made tests reflect precisely what is being taught" (p. 57). This issue of the need for assessment training for teachers was recognized by Feldmesser in 1971. He noted this problem as a "remedial deficiency" (Feldmesser, 1971, p. 17) and colleges that train our teachers were to blame. Brookhart (1993) sought to determine if providing teachers with measurement instruction would influence the meaning and value judgments associated with grades. She found that measurement instruction had very little impact. Most notable were situations of pass/fail. In these situations, all teachers were more inclined to let extenuating factors influence their final decisions.

Describe Assessment and Evaluation Practices, Including Grading, to Students, at the Beginning of Instruction

The old adage, *doctrine of no surprises* was being suggested by this guideline. In other words, no pertinent facts should have been left untold. Brookhart (2009) pointed out students as the primary users of the information that communicated how well they are learning. In order to be successful the communication of the system used to assess their learning should have been clear and the intended purpose of all parts of this system should have been made known to the learner (O'Connor, 2002). As noted earlier, students will remain motivated in their efforts to reach their goal if the way in which they were assessed aligned with their learning goals (Shute, 2008).

Purpose of the Study

This chapter began by noting the impact grades had on a student's motivation to learn (Brookhart, 2009). Guskey (2006) conducted a study investigating the lasting effects on teachers and school administrators with regards to their experience with grades. He found that their experiences, both positive and negative, were directly related to the particular actions taken by their teachers or professors. "For better and for worse, teachers and professors obviously have powerful influence on their students' perceptions of grading, what grades mean, and what impact they will have" (Guskey, 2006, p. 12).

Studies have found problems with the validity of grading systems and have confirmed a gap between current practice and measurement theory (Brookhart, 1994; Willingham, Ollack, & Lewis, 2002). Hopkins et al. (1990) introduced the concept of concurrent validity. As noted earlier, this type of validity was present when one variable was highly correlated with another variable of which one would expect a close relationship (e.g. a student' grade and his/her standardized test score).

In March of 2010, Anne-Sylvie M. Boykin conducted an analysis of the relationship between high school course grades and end-of-course (NC EOC) scores in Wake County Public Schools. The analyses consisted of data from over a four-year period and looked at the relationship between grades and NC EOC scores in Algebra I and English I. The outcome of her analyses indicated a lack of concurrent validity as it related to the correlation between students' course grades and NC EOC assessments. There were several noteworthy findings:

1. The percentage of students passing the Algebra I NC EOC but failing the course rose from 1.9% in 2004-05 to 4.0% in 2007-08.

2. The percentage of students failing the Algebra I NC EOC but passing the course saw a slight decline from 2004 -05 (3.5%) to 2006-07 (2.6%) but had a substantial increase in 2007 -08 (4.1%).
3. The percentage of students passing the English I NC EOC but failing the course had a slight decline from 6.6% in 2004-05 to 5.9% in 2007-08. There was actually a substantial decline in 2006 -07 (2.4%) only to see the noted increase in 2007-08.
4. The percentage of students failing the English I NC EOC but passing the course maintained a steady incline from 2004-05 (0.9%) to 2007-08 (4.3%).

Each of these findings supported the lack of concurrent validity between these measures. Boykin (2010) indicated that these results raised the question of how much nonachievement factors were taken into consideration when calculating a student's final grade. Was the student who was compliant, hardworking, and responsible awarded the extra point(s) necessary to achieve a passing course grade? Was the non-compliant student who was capable of mastering the material penalized for his/her behavior, resulting in a failing course grade?

In a study to determine the consequential validity of grades, Baron (2000) evaluated the points of view of the senders of grades, teachers, and the receivers of grades, students, parents, high school counselors, and college admissions officers. Participants were asked to rank eight factors they believed should be included in grade composition. The items participants were asked to be rank included achievement and nonachievement factors. Senders and receivers ranked achievement factors the highest with nonachievement factors receiving the lowest weight.

Taking into consideration findings from both of these studies, this study sought to determine the concurrent validity between course grades and performance on the Biology NC

EOC and Physical Science in classrooms where nonachievement factors were separated from the reporting of achievement factors in a teacher's grading system.

CHAPTER 3: RESEARCH METHODS

The purpose of this study was to explore the relationship between student achievement data, student grades, and teacher grading practices as was reflected in the responses to a survey on grading practices. The following questions guided the design and study procedures:

1. Of the grading systems reviewed which teachers' grading systems aligned more closely with research best practice?
2. Did teachers' grading systems with the greatest alignment to research best practices produce a stronger relationship between students' grades and their NC EOC performance?

The null hypothesis tested in this research was teachers grading systems that were strongly aligned with research best practice will not make a difference in the strength of the relationship between their students' grades and their NC EOC performance.

Research Design

This was a quantitative study involving data from eight teachers and their students from two different high schools. Four of the teachers teach Biology at Millbrook High School and four of the teachers teach Physical Science at Garner High School. It should be noted that teachers from these two schools were selected to take part in this study because they have participated in professional development regarding effective grading practices and claimed to have established a grading system that eliminates nonacademic behaviors. In addition to this, each of these teachers taught a course that required students to take an NC EOC for their exam score; thus making this a convenience sample.

The first part of the study consisted of Biology teachers at Millbrook High School and Physical Science teachers from Garner High School participating in an online survey regarding

their grading practices. The survey results provided a more detailed description of the elements each teacher included or excluded from his/her grading system. Survey responses for each teacher were totaled and a mean value was calculated. Several analyses of the mean values were conducted to explore relationships between and amongst teachers and teacher groups.

For the second part of the study, a correlation between the students' NC EOC achievement level scores and semester grades from Biology teachers at Millbrook High School and NC EOC achievement level scores and semester grades from Physical Science teachers at Garner High School was established. The strength of the relationship was established by calculating the Pearson r, looking for a correlation greater than .5.

The final step in this research design was the use of the Fischer's Exact Test to determine if a relationship existed between the Pearson r value and the teacher's mean score.

Description of the District

At the time of this study, Wake County Public School District was the largest school district in the state of North Carolina. The district included twelve municipalities and a range of schools within suburban and rural areas. Student enrollment for the 2009-10 school year was 139,599 students and for 2010-11 it is 143,289 students. In 2009-10 there were 102 elementary, 30 middle, 23 high, and 4 special/optional schools totaling 159 schools. There were 4 new schools added in 2010-11 bringing the totals to 103 elementary, 32 middle, 24 high, and 4 special/optional schools, totaling 163 schools. Wake County offered three different calendar options, year-round and traditional at the elementary level; year-round, traditional, and modified at the middle school level; and traditional and modified at the high school level. In 1982, Wake County Public Schools introduced magnet schools to the district in an effort to improve the "healthiness" of some schools that were seeing a significant increase in poverty. Schools were

provided additional resources enabling students to participate in unique educational opportunities. Parents of students applied to attend a magnet school and selection was determined via a lottery. In 2010 -11, Wake County Public Schools had 22 magnet schools; 19 elementary, 8 middle school, and 5 high school. Table 1 reflects the demographic make-up of the district for 2009-10 and Table 2 for 2010-11.

Description of the Schools

Garner High School was an International Baccalaureate magnet school located in the eastern part of the county. It fell within the Garner municipality. In 2009-10 student enrollment was 2,311 and in 2010-11 enrollment declined slightly to 2,297.

Millbrook High School opened the 2010-11 school year seeking authorization to become an International Baccalaureate magnet school. At the time of this study, the school was scheduled to have its authorization visit in October of this school year. Prior to this, Millbrook was a traditional high school. This school resided within the Raleigh municipality and was located in the north central region of the district. The enrollment for the 2009-10 school year was 2,432 and in 2010-11 enrollment increased slightly to 2,443.

Tables 3 and 4 reflect the demographic make-up of the two schools included in this study, Garner High School and Millbrook High School for the 2009-10 school year and Tables 5 and 6 reflect the demographic make-up of the two schools for the 2010-11 school year.

Measures

The measures used in this study were student numerical grades, scale scores from North Carolina's student achievement assessment (NC EOC) and teacher scores on the Grading Practice Survey.

Table 1

2009-10 Demographic Information for Wake County Public Schools

Category	Number of Students	Percent of Total
Special Education	17,518	12.73
Free and Reduced	43,729	31.2
English as a Second Language	6,875	4.9
Limited English Proficient	12,467	8.9
Academically Gifted	24,187	27
American Indian	380	0.3
Asian	8,501	6.1
Black or African American	36,174	25.9
Hispanic or Latino	16,516	11.8
Multi-Racial	6,640	4.8
White or Caucasian	71,388	51.1

Note. Retrieved from http://www.wcpss.net/demographics/.

Table 2

2010-11 Demographic Information for Wake County Public Schools

Category	Number of Students	Percent of Total
Special Education	18,984	13.4
Free and Reduced	46,456	32.4
English as a Second Language	6,096	4.3
Limited English Proficient	11,388	7.9
Academically Gifted	26,230	28
American Indian	638	0.4
Asian	8,644	6.0
Black or African American	35,494	24.8
Hispanic or Latino	20,909	14.6
Multi-Racial	6,471	4.5
White or Caucasian	70,986	49.5
*Native Hawaiian or Other Pacific Islander	147	0.1

Note. Retrieved from http://www.wcpss.net/demographics/. * This category in previous years was subsumed under "Asian or Pacific Islander". This is the first year the category was split into two categories to distinguish students of Native Hawaiian or Pacific Islander race.

Table 3

2009-10 Demographic Information for Garner High School

Category	Number of Students	Percent of Total
Special Education	312	14
Free and Reduced	890	38.6
English as a Second Language	68	2.9
Limited English Proficient	132	5.7
Academically Gifted	315	14.1
American Indian	10	0.4
Asian	16	0.7
Black or African American	1,058	45.8
Hispanic or Latino	294	12.7
Multi-Racial	85	3.7
White or Caucasian	848	36.7

Note. Retrieved from http://www.wcpss.net/demographics/.

Table 4

2009-10 Demographic Information for Millbrook High School

Category	Number of Students	Percent of Total
Special Education	351	15
Free and Reduced	840	34.3
English as a Second Language	90	3.7
Limited English Proficient	185	7.6
Academically Gifted	405	16.9
American Indian	4	0.2
Asian	106	4.4
Black or African American	905	37.2
Hispanic or Latino	303	12.5
Multi-Racial	76	3.1
White or Caucasian	1,038	42.7

Note. Retrieved from http://www.wcpss.net/demographics/.

Table 5

20010-11 Demographic Information for Garner High School

Category	Number of Students	Percent of Total
Special Education	301	13.1
Free and Reduced	867	37.7
English as a Second Language	54	2.4
Limited English Proficient	97	4.2
Academically Gifted	387	16.9
American Indian	11	0.4
Asian	29	4.2
Black or African American	888	38.7
Hispanic or Latino	379	12.8
Multi-Racial	119	5.1
White or Caucasian	866	36.9

Note. Retrieved from http://www.wcpss.net/demographics/.

Table 6

2010-11 Demographic Information for Millbrook High School

Category	Number of Students	Percent of Total
Special Education	336	13.9
Free and Reduced	772	31.6
English as a Second Language	73	3.0
Limited English Proficient	149	6.1
Academically Gifted	515	21.2
American Indian	9	0.3
Asian	107	4.4
Black or African American	791	32.6
Hispanic or Latino	351	14.5
Multi-Racial	103	4.2
White or Caucasian	1,063	42.8

Note. Retrieved from http://www.wcpss.net/demographics/.

Criterion-Referenced or Standardized Assessment

At the time of this study, the NC EOC assessment was administered to students enrolled in Algebra I and II, English I, Biology, Chemistry, Civic and Economics, U.S. History, Physics, and Physical Science. Prior to the 2010-11 school year, students were required to pass five of these assessments in order to graduate from high school (Algebra I, English I, US History, Civics and Economics, and Biology). After this school year, while it no longer counted as a requirement for graduation, students were still required to take the NC EOC as it served as the child's exam grade and counted 25% of their overall grade for the course. The NC EOC was a multiple-choice assessment and student performance on the assessment was measured by both a scale score and achievement level. There were four achievement levels. Each level represented a different level of mastery in the subject area being tested. Each of the subject areas had its own set of descriptors. For the purposes of this study, Table 7 and 8 provide the descriptors for levels of achievement in Biology and Physical Science, respectively. Table 9 and 10 represent the NC EOC achievement levels by scale score ranges for Biology and Physical Science, respectively. A scale score that falls within the level III or level IV range was considered proficient. In Wake County Public Schools the scale score was converted by the Evaluation and Research department to a value that aligned with the system's grading scale. This provided teachers with a numerical grade they could be used in calculating a student's final course grade. This grade counted as the student's exam grade and was 25% of his/her final grade.

Student Grades

Semester grades for the spring and fall of the 2009-10 school year and fall 2010-11 school year were utilized for this study. The Biology teachers in this study taught honors and academic levels. Grades from both levels were retrieved to be included in the data analysis.

Table 7

Achievement Level Descriptors for Biology

Achievement Level	Descriptor
Achievement Level I	Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area. Students performing at Achievement Level I recognize basic biological concepts and require extensive remediation to successfully complete the course. They can inconsistently identify the steps involved in the scientific process. Students at this level may recognize biological terms but demonstrate minimal understanding of the application of the terms.
Achievement Level II	Students at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area. Students performing at Achievement Level II demonstrate limited understanding of biology concepts. These students can organize and interpret data with direct guidance. Students can identify basic cell structure/function, understand basic DNA structure, and solve simple Punnett squares. They can identify basic patterns of animal behavior and relationships among organisms in ecological systems.
Achievement Level III	Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well prepared for a more advanced level in the content area. Students performing at Achievement Level III demonstrate mastery of biology concepts. These students understand and can conduct scientific inquiry, interpret and analyze data, and make predictions pertaining to various biological processes. They can understand form and function of biological systems. Students at this level interpret factors contributing to patterns of inheritance, animal behavior, and human health. They can assess how biotic and abiotic factors influence homeostasis within changing systems.

Table 7 (*continued*)

Achievement Level IV	Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area. Students performing at Achievement Level IV demonstrate an advanced application of the scientific skills. These students demonstrate mastery of scientific inquiry skills including experimental design, identification of alternate interpretations of data, and application to real-world experiences. They can analyze the relationship between form and function of biological systems. They can also analyze and evaluate factors contributing to complex patterns of inheritance, animal behavior, and human health. Students at this level use their understanding of the universal genetic code to make predictions about unity and diversity of life.

Note. Table contents were retrieved from NC State Board of Education Policy GCS-C-010 http://sbepolicy.dpi.state.nc.us/.

Table 8

Achievement Level Descriptors for Physical Science

Achievement Level	Descriptor
Achievement Level I	Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area. Students performing at Achievement Level I show limited knowledge in the areas of forces and motion, energy and conservation, electricity and magnetism, structure and properties of matter, and regularities in chemistry. Students understand some components of the scientific inquiry process and may be able to conduct a simple experiment.
Achievement Level II	Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area. Students performing at Achievement Level II utilize some graphing and mathematical skills to solve problems in the areas of forces and motion, energy and conservation, electricity and magnetism, structure and properties of matter, and regularities in chemistry. Students show basic understanding of the scientific inquiry process by conducting simple experiments.
Achievement Level III	Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well prepared for a more advanced level in the content area. Students performing at Achievement Level III utilize graphing, mathematical, and interpretive skills to solve problems in the areas of forces and motion, energy and conservation, electricity and magnetism, structure and properties of matter, and regularities in chemistry. Students show understanding of the scientific inquiry process by designing and conducting simple experiments, then interpreting the results.

Table 8 (*continued*)

Achievement Level IV	Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area. Students performing at Achievement Level IV can interpret and analyze charts, graphs, data and mathematical formulas. Students replicate and evaluate scientific investigations and generate alternative interpretations related to the areas of forces and motion, energy and conservation, electricity and magnetism, structure and properties of matter, and regularities in chemistry.

Note. Table contents were retrieved from NC State Board of Education Policy GCS-C-010 http://sbepolicy.dpi.state.nc.us/.

Table 9

Achievement Level Ranges for North Carolina NC EOC in Biology

Test	Level I	Level II	Level III	Level IV
Biology	≤137	138-146	147-158	≥159

Note. Table contents were retrieved from NC State Board of Education Policy GCS-C-010 http://sbepolicy.dpi.state.nc.us/.

Table 10

Achievement Level Ranges for North Carolina NC EOC in Physical Science

Test	Level I	Level II	Level III	Level IV
Physical Science	≤139	140-148	149-159	≥160

Note. Table contents were retrieved from NC State Board of Education Policy GCS-C-010 http://sbepolicy.dpi.state.nc.us/.

There were no honors Physical Science course offered in Wake County. The district's grading scale was mandated by general statute, specifically GCS-L-004 (see Appendix I). The grading scale outlined in the policy was as follows: 93-100=A; 85-92=B; 77-84=C; 70-76=D; ≤69=F.

Nonacademic factors were purported to be eliminated from the grading system used by the teachers in this study. Therefore, teachers in this study claimed to have eliminated items such as extra-credit for nonacademic task (turning in an interim report on time), class participation, and student effort from grade calculation.

Teacher Grading Practice Survey

Teachers participated in an on-line survey that was developed using best practices as defined by the literature, (Brookhart, 2009; Guskey & Bailey, 2001; Marzano, 2000; O'Connor, 2007) (see Appendix J). The survey consisted of 29 items, each accompanied by a Likert scale response format of *always*, *almost always, varies, some,* and *little to none*. For the purposes of analysis, the response formats were recoded into rank numerical values from 1 to 5, with 5 representing *little to none,* 4 representing *some,* 3 representing *varies,* 2 representing *almost always,* and 1 representing the highest possible rating, *always.* The survey sought to assess a variety of grading practices, those supported by the research (e.g. excluding formative assessment scores from grade calculation) and those not supported by research (e.g. giving zeros for late work). This was done in order to eliminate the possibility of bias in the scores and to help avoid a repeated pattern of responses so that teachers would be thoughtful and deliberate in their responses. Items describing practices that were not supported by the literature were reverse coded in order that higher scores consistently indicated a use of a grading system that eliminated nonachievement factors. For example, item 4 on the survey asked teachers to respond to this statement, "As a teacher, I give a zero for cheating". The highest response of 5 (little to none)

was aligned with the research. However, in item 3 teachers were asked if they "accept late work for full credit". A response of 5 (little to none) was the highest value but not aligned with research. Therefore, prior to calculating the mean value for each teacher, items like this second example were reverse coded (5 changed to 1 and vice versa) in order for the highest score to consistently reflect a grading system most aligned with research.

There was also a section in the survey that sought to identify the amount of collaboration that existed amongst the teachers at each of the schools as it related to their grading practices. This information was used to address any noted trends in the data and/or similarities and differences in subject areas. The survey was constructed in such a way that teachers were required to provide a response for each of the items.

This survey was piloted during the 2008-09 school year. Teachers across Wake County's school district participated in the survey. Participation was voluntary. The results of the survey were used by school level administrators and teachers as a catalyst for conversation regarding grading practices at their respective schools.

The survey was divided into several themes; use of zeros, partial or full credit for different circumstances, use of extra credit, assigning of grades to group task, specific questions regarding grade calculation, and lastly there was a section that asked participants to identify how they incorporated items that were difficult to measure (e.g. student effort, academic growth, and professional judgment).

Description of the Database

NC EOC scale scores were provided by the district's Evaluation and Research (E&R) Department. Student grades and demographic information were retrieved from the NCWISE

Student Information System also provided by the E&R department. The Grading Practices Survey was deployed through Zoomerang, an on-line assessment tool used by the district.

Data Analysis

The data analysis consisted of three steps. First, a relationship between NC EOC scale scores and student grades in Biology and a relationship between NC EOC scale scores and student grades in Physical Science was determined. For the Biology classes, a relationship for honors and academic levels were established separately. Teachers included in this study were those who claimed to assess students in their classes on achievement factors alone. In Wake County, as mandated by GCS –C-003 (see Appendix I), when the final exam in a course was an NC EOC, the score for this exam counted 25% of the student's final grade. Because this study was trying to determine the relationship between a student's NC EOC scale score and his/her final grade the researcher recalculated the students' final grades removing the exam grade from the average to eliminate the redundancy. The Pearson r was calculated to establish the strength of the relationship between the two data sets, a student's final numerical grade and his/her NC EOC scale score. Calculating the correlation established an effect size that fell within a 0-1 range with 0 indicating no effect and 1 representing a perfect effect. A p value of .01 was established.

The second step was to analyze teacher survey responses. Teacher survey responses were assigned an ordinal value as noted earlier. These response values were totaled and a mean value was calculated for each teacher. Weighting of the responses was constructed in such a way that a higher score, a mean score close to five, indicated a grading system that was based primarily on achievement factors, whereas a lower score, a mean score close to one, indicated a grading system that incorporated nonachievement factors. Descriptive information regarding the overall trends and patterns of the survey was also provided to examine individual items and any

differences or similarities between the subject areas. For example, were all the teachers' scores from one school similar and if so, did this align with their responses to the questions regarding collaboration?

The final step of the analyses was to construct a Fisher's Exact Test in order to answer the second research question. This test was used to determine if there was a nonrandom association between two categorical variables (Weisstein, 2010). In this study, we examined the association between the correlation values of students' grades to their NC EOC scale scores to a teacher's mean score from the survey results. In order to conduct this test, scores from both sets of data had to be categorized. Teachers with mean scores greater than or equal to 3.5 were categorized as those who established grading systems predominately on achievement factors and values less than 3.5 were considered to represent grading systems that included nonachievement factors. The strength of the relationships between the students' grades and their NC EOC level score were categorized into two categories; those with $r \geq .7$ and those with $r < .7$. A p value of .01 was established. All calculations and data analyses were conducted using PASW Statistics 18, Release Version 18.0.

Summary

This chapter described the methods used in conducting this non-experimental, quantitative study design. Three different data sets were used in order to determine correlations and strength of correlations in an effort to establish a grading system that was a strong predictor of a student's performance on the NC EOC.

CHAPTER 4: RESULTS OF THE STUDY

The purpose of this study was to determine if grading systems that were aligned with researched best practices would yield more consistent results in determining how well students would perform on a standardized test. In seeking this answer, the grading systems of teachers who claimed to adhere to best practices with regards to their grading system were examined via a survey. A second component of this study examined the relationship between a student's grade and his/her performance on the North Carolina End of Course (NC EOC) test in each of the teacher's classes.

Research Questions

This study sought to answer the following research questions:

1. Of the grading systems reviewed which teachers' grading systems aligned more closely with research best practice?
2. Did teachers' grading systems with the greatest alignment to research best practices produce a stronger relationship between students' grades and their NC EOC performance?

The null hypothesis tested in this research was teachers grading systems that were strongly aligned with research best practice will not make a difference in the strength of the relationship between their students' grades and their NC EOC performance.

Review of Methodology

This quantitative study involved the disaggregation of data from eight teachers and their students from two different high schools. The number of student grades totaled to 1,050. The participating teachers were selected because they claimed to have established a grading system that was more in line with researched best practice. Each of the teachers had participated in

professional development regarding best practices for grading. In addition each participant taught a course where students were required to take an NC EOC. The courses reviewed for this study were Biology and Physical Science as these two courses require an NC EOC for the purposes of final examination. Data were retrieved from three semesters for six of the eight teachers (Teachers B – F, and H). Teacher A only taught Physical Science in fall 2010 therefore there was only one semester of data to be retrieved. The data from fall 2009 and spring 2010 for Teacher G could not be used as the teacher only recorded letter grades and could not access the numeric grades from NC Wise as the system does not store historical data.

To address the first research question, participating teachers were asked to complete an on-line survey regarding their grading practices. The survey consisted of statements that reflected proper grading practices supported by research as well as items that inquired about their perceptions regarding improper practices. This combination of item types was chosen in order to eliminate the possibility of researcher bias and to help avoid a repeated pattern in teacher responses. For example, in order to avoid all responses of 5 reflecting a grading practice supported by research some statements were worded in such a way that a response of 1 was considered to be the most acceptable practice. When the teacher survey averages were calculated these responses were reverse coded. Specifically, values of 1 were changed to 5, values of 5 were changed to 1, values of 2 were changed to 4, values of 4 were changed to 2, and values of 3 remained neutral. This was done in order to have the higher score consistently reflect a grading system supported by research. Survey responses for each teacher were totaled and a mean value was calculated. Individual item scores had a possible range of 1-5.

Additional analyses of the survey were conducted to determine how aligned the teachers' grading practices were with the research according to each of the statements. Are there practices

that teachers seem to follow more closely than others? This was done by calculating the mean value for all of the teacher responses around a specific practice. For example, questions one through three dealt with how a teacher handled the issue of late work. The mean value was calculated for these three questions by teacher to determine how aligned each teacher was with this specific practice. The teacher responses were then separated by their content area to determine if there was a difference amongst the teacher groups with regards to their alignment with the research.

To answer the second research question, Pearson r coefficients were computed to determine the strength of the relationship and then compared with the teacher survey results. It should be noted that in North Carolina the State Board of Education approved a policy stating that all NC EOC scores shall count at least 25 % of the student's final grade in the associated course (see Appendix J). In Wake County, the NC EOC counted as 25% of the student's final grade. Because this study was trying to determine how well a teacher's grading system measured student performance, as measured by the correlation between a student's grade and his/her NC EOC scale score, the student's grade was recalculated in order to remove the exam grade from the average. This process eliminated the redundancy in the comparison when the exam grade is the NC EOC score. The Biology classes consisted of both honors and regular. Correlations were calculated separately by course type and then calculated with both levels combined. All correlations were determined using PASW Statistics 18, Release Version 18.0

A series of Fisher's Exact Test were conducted to determine if there was a significant relationship between the two categorical variables, correlation of student's final grade and NC EOC performance and mean value from grading practice survey.

Results

Research Question # 1

Of the grading systems reviewed which teachers' grading systems aligned more closely with research best practice?

In order to answer this question teachers were asked to participate in an on line survey regarding their grading practices. While each of the teacher's claimed to have a grading system that aligned with best practice according to the research, responses to the survey helped to determine which teacher adhered most closely to a system recommended by grading experts and believed to report a more reliable grade. Participants were asked to rate their practices on a Likert scale ranging from 1 to 5 with 1 representing *Always and* 5 representing *Little to None.* As noted earlier, items representing inappropriate practices were reverse coded prior to calculating the average in order that higher scores always indicated a use of a better grading practice. The value of 3 remained neutral. Table 11 represents teacher responses to each of the statements. The statement responses that were reverse coded in order to calculate the mean are italicized. However, for the ease of interpreting teacher responses for the reader, two values are noted in Table 11. The original ratings provided by the teacher are noted outside the parenthesis and the reverse coded values used for calculating the mean are inside the parenthesis.

Two separate mean values were calculated. Statements 1 through 18 represent items that sought to determine to what degree a teacher removed nonacademic factors from his/her grading system. Therefore, the first mean was calculated using these scores only. The second mean included values from all twenty-two statements. While the remaining four statements were not directly related to nonacademic factors they were noted in the research as factors that should be considered when creating a grading system. Therefore, the second mean value was calculated to

Table 11

Teacher Survey Reponses and Mean: Value 1 = Always, 5 = Little to None

Statement	Teacher							
	A	B	C	D	E	F	G	H
1. As a teacher I give zeros for late work.	4	5	5	5	5	5	5	5
2. As a teacher I give partial credit for work that is turned in late.	1	1	1	1	5	1	2	3
3. As a teacher I accept late work for full credit.	4(2)	5 (1)	5 (1)	3(3)	5 (1)	1(5)	5(1)	3 (3)
4. As a teacher I give a zero for cheating.	1	3	3	3	5	5	4	4
5. As a teacher I allow a student to retake a test for full credit when caught cheating.	2(4)	1(5)	3(3)	2(4)	1(5)	3(3)	4(2)	5(1)
6. As a teacher I allow a student to retake a test for partial credit when caught cheating.	2	5	3	2	5	1	2	1
7. As a teacher I offer extra credit for academic tasks.	4	4	3	5	5	1	5	3
8. As a teacher I offer extra credit for non-academic tasks.	5	5	5	5	5	1	5	5
9. As a teacher, when assigning a grade to a group task I give the same grade to all members of the group.	3	2	1	3	5	5	2	5

Table 11 (*continued*)

10. As a teacher, when assigning a grade to a group task I give each member an individual grade as well as a group grade.*	5	1	5	3	5	5	5	5
11. As a teacher, when assigning a grade to a group task I only give individual grades.	3(3)	5(1)	5(1)	3(3)	1(5)	1(5)	4(2)	1(5)
12. As I teacher I lower grades for disruptive behavior.	5	5	5	5	5	5	5	5
13. As I teacher I give a class participation grade.	5	4	4	5	5	5	5	5
14. As I teacher I consider neatness when grading assignments.	5	3	5	4	5	5	5	4
15. As I teacher I consider a student's academic growth when assigning a grade.	5	3	4	2	5	1	5	3
16. As I teacher I consider a student's academic ability when assigning a grade.	4	3	4	4	5	1	5	3
17. As I teacher I consider a student's effort when assigning a grade.	4	4	5	4	5	5	5	3
18. As I teacher I use professional judgment to determine grades.	4	5	3	1	3	5	5	1
19. As a teacher I include the scores from common formative assessments in my grading scheme.	5	5	4	3	5	5	5	5

Table 11 (*continued*)

20. As I teacher I provide opportunities for students to retake assessments to improve their grades.	3(3)	1(5)	1(5)	1(5)	1(5)	1(5)	2(4)	1(5)
21. As I teacher I grade on a curve.	5	4	4	5	5	3	5	3
22. As I teacher I weigh graded work completed at the beginning of the term equally with graded work completed at the end of the term.	1	1	1	4	5	3	2	4
Mean Value for Qs 1-18	3.67	3.33	3.89	3.44	4.67	3.78	3.89	3.56
Mean Value for Qs 1-22	3.64	3.41	3.41	3.59	4.73	3.82	3.91	3.68

* A group mark is considered acceptable if it is based solely on non-academic performance and separated from the academic performance. Because this survey was based on how a teacher gathers grades for academic performance, it is not considered best practice to give a group academic grade. () represents values used in calculating the mean value for each of the teacher responses.

determine how aligned a teacher's overall grading practice was with what grading experts indicated was good practice. The higher value indicated a grading system that is more in line with researched best practices.

Mean values for the teachers ranged from 3.33 to 4.67 for questions 1-18 and 3.41 to 4.73 for questions 1-22. While none of the mean values indicated that any of the teachers were using a grading system that was in gross conflict with the research, Teacher E clearly had the strongest score with the second highest value .78 points away in one comparison and .82 points away in the second.

Analysis of Mean by Statement

After assessing the overall value of each teacher's grading practice, the researcher calculated means for each statement response to determine if there were practices that were more closely aligned than others by all the teachers. As a reminder a value of 5 represents *little to none* and a value of 1 represents *always.* Items had to be reverse coded when seeking mean values for multiple items. This was done in order to maintain consistency in that the higher value would represent the stronger score. Because this analysis was only on one item, no reverse coding was necessary. The results of this analysis are summarized in Table 12. Items in italics are those in which the majority of the teachers adhered to practices research would support. For this study values between 4 *(some, 10-39% of the time)* and 5 *(little to none, <10% of the time)* were accepted as in alignment as were values between 1*(always, >95% of the time)* and 2 *(almost always, 80-95% of the time),* where appropriate.

A review of mean values for individual statements revealed that fourteen out of the twenty-two statements were ones in which the majority of teachers adhered to practice supported by research ($\bar{x} \geq 3.5$). When reviewing the literature regarding the guidelines O'Connor (1995)

Table 12

Mean Value for Each Statement Response for All Teachers

Statement	Mean Value by Statement for All Teachers	Research would support a value of
1. As a teacher I give zeros for late work.	*4.875*	*5*
2. As a teacher I give partial credit for work that is turned in late.	2.375	5
3. As a teacher I accept late work for full credit.	3.875	1
4. As a teacher I give a zero for cheating.	3.500	5
5. As a teacher I allow a student to retake a test for full credit when caught cheating.	2.625	1
6. As a teacher I allow a student to retake a test for partial credit when caught cheating.	2.625	5
7. As a teacher I offer extra credit for academic tasks.	3.750	5
8. As a teacher I offer extra credit for non-academic tasks.	*4.500*	*5*
9. As a teacher, when assigning a grade to a group task I give the same grade to all members of the group.	3.250	5
10. As a teacher, when assigning a grade to a group task I give each member an individual grade as well as a group grade.	*4.250*	*5*
11. As a teacher, when assigning a grade to a group task I only give individual grades.	2.875	1
12. As I teacher I lower grades for disruptive behavior.	*5.000*	*5*
13. As I teacher I give a class participation grade.	*4.750*	*5*
14. As I teacher I consider neatness when grading assignments.	*4.500*	*5*

Table 12 (*continued*)

15. As I teacher I consider a student's academic growth when assigning a grade.	3.500	5
16. As I teacher I consider a student's academic ability when assigning a grade.	3.625	5
17. As I teacher I consider a student's effort when assigning a grade.	*4.375*	*5*
18. As I teacher I use professional judgment to determine grades.	3.375	5
19. As a teacher I include the scores from common formative assessments in my grading scheme.	*4.625*	*5*
20. As I teacher I provide opportunities for students to retake assessments to improve their grades.	*1.375*	*1*
21. As I teacher I grade on a curve.	*4.250*	*5*
22. As I teacher I weigh graded work completed at the beginning of the term equally with graded work completed at the end of the term.	2.625	5

proposed for constructing a valid grading system, the items in which these teachers had the greatest alignment were those that fell under the guideline *Limit the Valued Attributes.* Specifically, this guideline suggested the removal of nonacademic behaviors such as effort, participation, working well in groups, dependability, punctuality, attitude and other personal/social characteristics from a grading system. Statements 1-18 addressed these attributes either directly or indirectly. For example, statements 1-3 were trying to determine how a teacher addressed the issue of late work/punctuality. Were students penalized within their grade for not being punctual and if so to what degree? Or, as was trying to be determined in question eight, were students given extra points for nonacademic tasks such as bringing in a box of tissue? This is not only assessing effort but one's financial means to provide needed items.

There were eleven elements of a grading system being assessed in statements 1-18: (1) late work; (2) cheating; (3) extra credit; (4) group work; (5) disruptive behavior; (6) participation; (7) neatness; (8) academic growth; (9) academic ability; (10) effort; and (11) professional judgment. Of these eleven elements, the majority of the teacher responses were aligned with research in the areas of disruptive behavior, participation, neatness, extra credit and effort. However, inconsistencies emerged when reviewing the responses associated with late work and cheating. While the teachers agreed that assigning a zero for late work was not acceptable, there was not agreement on the amount of credit to award an assignment when turned in late. There was moderate agreement in the belief that a student should not be given a zero for cheating but when asked if the student should be allowed to retake the test for full or partial credit the responses differed. In order to take a closer look at how well each teacher adhered to certain attributes of a grading system, items that assessed the same attribute were clustered and mean values were calculated. This analysis is reviewed in the next section.

Analysis of Mean by Attribute

The researcher then clustered statement responses that addressed the same attribute and calculated the mean to determine how closely aligned each teacher was to specific attributes. As a reminder, the values were reverse coded for calculating the mean but were restored to the original values for ease of interpretation by the reader. Both values were represented in Table 11 with the reverse coded value in parenthesis. Items in italics represent the items that were previously recoded for calculating means.

Statements 1-3 addressed the issue of late work and Table 13 summarizes the findings. As noted in the literature review, it was recommended that a student be given full credit for the work he/she produced but should be punished separately for not being prompt in turning in the assignment. The closer the mean value was to five the more in line a teacher's practice was with the research. It was noted earlier that all the teachers agreed that assigning a zero for late work was not acceptable. However, where they began to differ was in how much credit to award a late assignment. None of the teachers, with the exception of Teacher F, was willing to give the student full credit for turning in a late assignment. All of the other teachers believed that there should be some consequence imposed on the student's academic grade for not turning the assignment in on time.

Statements 4-6 dealt with how the teacher handled the matter of cheating results of the analysis are in Table 14. With regards to this attribute Teacher E was consistent in his responses being aligned with research. Teacher B was somewhat inconsistent in his practice. He claimed to give a student full credit when retaking a test after being caught cheating yet in the question above he indicated that at times he may give a zero for cheating. Teacher A also struggled to provide a clear position on this issue. In question 4 he claimed to always give a zero yet his

Table 13

Mean Value for Statement Response for All Teachers on Statements 1-3 Late Work

Statement	Teacher							
	A	B	C	D	E	F	G	H
1. As a teacher I give zeros for late work.	4	5	5	5	5	5	5	5
2. As a teacher I give partial credit for work that is turned in late.	1	1	1	1	5	1	2	3
3. As a teacher I accept late work for full credit.	*4(2)*	*5(1)*	*5(1)*	*3(3)*	*5(1)*	*1(5)*	*5(1)*	*3(3)*
Average for Qs 1-3	2.33	.233	2.33	3.00	3.67	3.67	2.67	3.67

Note. () represents values used in calculating the mean value for each of the teacher responses.

Table 14

Mean Value for Statement Response for All Teachers on Statements 4-6 Cheating

	Teacher							
Statement	A	B	C	D	E	F	G	H
4. As a teacher I give a zero for cheating.	1	3	3	3	5	5	4	4
5. As a teacher I allow a student to retake a test for full credit when caught cheating.	*2(4)*	*1(5)*	*3(3)*	*2(4)*	*1(5)*	*3(3)*	*4(2)*	*5(1)*
6. As a teacher I allow a student to retake a test for partial credit when caught cheating.	2	5	3	2	5	1	2	1
Average for Qs 4-6	2.33	4.33	3.00	3.00	5.00	3.00	2.67	2.00

Note. () represents values used in calculating the mean value for each of the teacher responses.

responses to 5 and 6 indicated that he often allowed a student to retake a test for full or partial credit. This set of statements yielded the greatest inconsistencies in responses. It appeared as if all but Teacher E struggled to take a position on this issue.

In looking at the statements that address extra credit this attribute yielded the greatest consistency amongst the teachers in their alignment with the research. There was really only one outlier who gave extra credit for either type of task, see Table 15. The research supported not providing extra credit for either type of task. All teachers, with the exception of Teacher F, were aligned with the research. Teachers D, E, and G had the strongest alignment.

Lastly, the researcher reviewed the mean values for statements 9-11. These statements addressed the issue of group work. Research supported assigning individual academic grades for a group task and evaluating a student's group participation via a nonacademic/behavior rubric. Therefore a mean value of 5 for these responses was most closely aligned with research. Teachers E, F, and H were explicit in that they only provided individual academic grades (see Table 16). What this did not tell us was how or if they evaluated the student's group participation (e.g. did they use a behavior rubric for feedback about group participation) but it was clear that this evaluation was not included in the student's academic grade. With Teacher B, it was clear that she never gave individual grades only. It appeared by her responses that she always gave both an individual and group grade and that the group grade was always the same. Teacher C was also very clear in his method. He did not give individual grades or a combination of an individual and group grade. He only gave one grade to each member of the group and it was the same grade for all the members. So, as a student, you could have essentially put in little to no effort and received an A if all the other members carried their weight. The reverse could also have been true. One member of the team could have worked very hard and all the other members

Table 15

Mean Value for Statement Response for All Teachers on Statements 7-8 Extra Credit

	Teacher							
Statement	A	B	C	D	E	F	G	H
7. As a teacher I offer extra credit for academic tasks.	4	4	3	5	5	1	5	3
8. As a teacher I offer extra credit for non-academic tasks.	5	5	5	5	5	1	5	5
Average for Qs 7-8	4.50	4.50	4.00	5.00	5.00	1.00	5.00	4.00

Table 16

Mean Value for Statement Response for All Teachers on Statements 9-11 Group Work

	Teacher							
Statement	A	B	C	D	E	F	G	H
9. As a teacher, when assigning a grade to a group task I give the same grade to all members of the group.	3	2	1	3	5	5	2	5
10. As a teacher, when assigning a grade to a group task I give each member an individual grade as well as a group grade.	5	1	5	3	5	5	5	5
11. As a teacher, when assigning a grade to a group task I only give individual grades.	3(3)	5(1)	5(1)	3(3)	1(5)	1(5)	4(2)	1(5)
Average for Qs 9-11	3.67	1.33	2.33	3.00	5.00	5.00	3.00	5.00

Note. () represents values used in calculating the mean value for each of the teacher responses.

could have offered little to no effort yet everyone received a poor grade, even the individual who worked very hard.

Analysis of Additional Guidelines for Grading

Statements 19-22 of the survey addressed other guidelines O'Connor noted as critical when constructing a valid grading system. They were not considered to be nonacademic evaluations therefore they were separated from the mean value of questions 1-18. For all four responses Teacher E was consistent in his adherence to researched best practice (see Table 17).

Analysis of Mean by Subject

The next analysis of question responses was conducted in order to determine how well the collective grading practices of teachers from the same school aligned with research. The two groups of teachers operated within a professional learning team (PLT). Rick DuFour is one of the leading researchers in the area of PLTs. In one of his books he cited the following six characteristics as essential components in effective PLTs: (1) shared vision and values; (2) collaborative culture; (3) collective inquiry into best practice; (4) a focus on results; (5) supportive and shared leadership; and (6) supportive conditions (DuFour & Eaker, 1998). In the high schools in Wake County, PLTs were established around a course taught in an effort to establish common practices within the same course. For example, all teachers teaching Algebra I in a high school would be members of the same PLT. In this study teachers A, D, E, and G operated within the same Physical Science PLT and teachers B, C, F, and H operated within the same Biology PLT. Their team average for each of the responses is noted in Table 18. The Physical Science PLT had thirteen out of twenty-two responses that were aligned with research and of the eleven attributes identified in statements 1-18, there was alignment amongst these teachers in seven of the eleven (extra credit, disruptive behavior, participation, neatness,

Table 17

Statement Response for All Teachers on Statements 19-22

	Teacher							
Statement	A	B	C	D	E	F	G	H
19. As a teacher I include the scores from common formative assessments in my grading scheme.	5	5	4	3	5	5	5	5
20. As I teacher I provide opportunities for students to retake assessments to improve their grades.	3(3)	1(5)	1(5)	1(5)	1(5)	1(5)	2(4)	1(5)
21. As I teacher I grade on a curve.	5	4	4	5	5	3	5	3
22. As I teacher I weigh graded work completed at the beginning of the term equally with graded work completed at the end of the term.	1	1	1	4	5	3	2	4
Average for Qs 19-22	3.50	3.75	2.50	4.25	5.00	4.00	4.00	4.25

Note. () represents values used in calculating the mean value for each of the teacher responses.

Table 18

Physical Science and Biology PLT Mean Values for Survey Responses

Statement	Mean Value by Statement for All Physical Science Teachers	Mean Value by Statement for All Biology Teachers	Research would support a value of
1. As a teacher I give zeros for late work.	*4.75*	*5.00*	*5*
2. As a teacher I give partial credit for work that is turned in late.	2.25	2.50	5
3. As a teacher I accept late work for full credit.	4.25	3.50	1
4. As a teacher I give a zero for cheating.	3.25	3.75	5
5. As a teacher I allow a student to retake a test for full credit when caught cheating.	2.25	3.00	1
6. As a teacher I allow a student to retake a test for partial credit when caught cheating.	2.75	2.50	5
7. As a teacher I offer extra credit for academic tasks.	*4.75*	2.75	*5*
8. As a teacher I offer extra credit for non-academic tasks.	*5.00*	*4.00*	*5*
9. As a teacher, when assigning a grade to a group task I give the same grade to all members of the group.	3.25	3.25	5
10. As a teacher, when assigning a grade to a group task I give each member an individual grade as well as a group grade.	*4.50*	*4.00*	*5*
11. As a teacher, when assigning a grade to a group task I only give individual grades.	2.75	3.00	1
12. As I teacher I lower grades for disruptive behavior.	*5.00*	*5.00*	*5*

Table 18 (*continued*)

13. As I teacher I give a class participation grade.	*5.00*	*4.50*	*5*
14. As I teacher I consider neatness when grading assignments.	*4.75*	*4.25*	*5*
15. As I teacher I consider a student's academic growth when assigning a grade.	*4.25*	2.75	*5*
16. As I teacher I consider a student's academic ability when assigning a grade.	*4.0*	2.75	*5*
17. As I teacher I consider a student's effort when assigning a grade.	*4.50*	*4.25*	*5*
18. As I teacher I use professional judgment to determine grades.	3.25	3.50	5
19. As a teacher I include the scores from common formative assessments in my grading scheme.	*4.50*	*4.75*	*5*
20. As I teacher I provide opportunities for students to retake assessments to improve their grades.	*1.75*	*1.00*	*1*
21. As I teacher I grade on a curve.	*5.00*	3.50	*5*
22. As I teacher I weigh graded work completed at the beginning of the term equally with graded work completed at the end of the term.	3.00	2.25	5

academic growth, academic ability, and effort). The Biology PLT also had an overall alignment in thirteen out of twenty-two of the responses, with nine of the items being the same as the Physical Science PLT. Out of the eleven attributes identified in statements 1-18 there was alignment in five (disruptive behavior, participation, neatness, effort, and use of professional judgment). When taking into consideration responses from all the teachers as compared to their PLT responses, question twenty-one was aligned with research when considered collectively, but when separated out it became evident that the Biology PLT's responses were not as aligned with research. There were three additional statements in which the members of the Physical Science PLT responded with answers aligned with the research, statements 15, 16, and 22. Thus when reviewing the separate PLT data and comparing it to the collective responses noted in Table 12 it appeared as if the Physical Science PLT enhanced the mean values in this table. While neither PLT had many responses that were 100% aligned, the Physical Science PLT had three statements, 8, 12, and 13 that were 100% aligned versus the one statement from the Biology PLT (statement 1). It was interesting to note that the two teams did not share any response that was 100% aligned.

Lastly, statements 23 and 25 of the survey had to do with how well the teachers from each PLT believed their group collaborated with regards to their grading practices. Table 19 reflects the mean value for each PLT's responses. A value of one was considered to be the most favorable response for each of these statements. While neither PLT was 100% aligned with regards to their grading practices, the Physical Science PLT appeared to have stronger alignment within their team. This corresponded with the results from the previous analysis wherein the Physical Science PLT reported to have more practices aligned with research and more statements in which they were 100% in agreement.

Table 19

Physical Science and Biology PLT Mean Values for Statements Regarding Collaboration: Value 1 = Always, 5 = Little to None

Statement	Mean Value by Statement for All Physical Science Teachers	Mean Value by Statement for All Biology Teachers
My PLT has a unified grading system	2	2
I use the same grading procedure as others in my PLT on graded work	1.5	2.5

Research Question #2

Did teachers' grading systems with the greatest alignment to research best practice produce a stronger relationship between students' grades and their NC EOC performance?

A Pearson r coefficient was calculated for each of the teacher's classes to assess the relationship between a student's final grade and his or her performance on the NC EOC. The correlation coefficient represents the strength and direction of this relationship at the $p = .05$ level.

Physical Science Teachers

Teachers A, D, E, and G taught Physical Science and claimed to utilize the same grading system. Table 20 presents the correlation coefficients for each Physical Science teacher indicating the relationship between their students' final grades and their performance on the NC EOC. For each of the teachers, the relationship was positive and all correlations were significant at the .05 level.

Biology Teachers

Teachers B, C, F, and H taught both regular and honors level Biology classes, and reported the usage of a consistent grading system across teachers and classrooms. Table 21 presents the correlations from these teachers, reporting honors and regular classes separately. The output, regardless of the level, showed that a student's final grade was positively related to his or her performance on the NC EOC at the .05 level.

Comparison of Survey Results to Correlations

The individual teacher mean values from the survey were then compared to the correlations. Table 22 displays the information that was used in the comparison. As a reminder, Teachers A, D, E, and G do not teach an honors level course.

Table 20

Correlation Between a Student's Final Grade Assigned by the Teacher and the NC EOC Scale Score for Teachers of Physical Science

Teacher	Number of Student Grades	Pearson's Correlation Coefficient
Teacher A	45	.822*
Teacher D	63	.599*
Teacher E	72	.565*
Teacher G	67	.755*

Note. *Correlation is significant at the 0.05 level (2-tailed).

Table 21

Correlation Between a Student's Final Grade Assigned by the Teacher and the NC EOC Scale Score for Teachers of Biology

Teacher	Number of Student Grades for Honors	Pearson's Correlation Coefficient for Honors	Number of Student Grades for Regular	Pearson's Correlation Coefficient for Regular
Teacher B	153	.746*	76	.741*
Teacher C	181	.587*	71	.588*
Teacher F	56	.610*	110	.765*
Teacher H	63	.560*	93	.469*

Note. *Correlation is significant at the 0.05 level (2-tailed).

Table 22

Comparison of Pearson r Correlations and Mean Values from Survey

Teacher	Pearson Correlation Regular	Pearson Correlation Honors Reg.	Pearson Correlation Hrs	$\bar{x}$ value Qs 1-18	$\bar{x}$ value Qs 1-22
Teacher A	.822			3.67	3.64
Teacher B	.741	.746	.833	3.33	3.41
Teacher C	.588	.587	.741	3.39	3.41
Teacher D	.599			3.44	3.59
Teacher E	.565			4.67	4.72
Teacher F	.765	.610	.765	3.78	3.82
Teacher G	.755			3.89	3.91
Teacher H	.469	.560	.652	3.56	3.68

A series of Fisher's Exact Tests were calculated using the information from Table 22 to determine if there is an association of significance between these two categorical variables; correlation between a student's final grade and NC EOC score and mean value of teacher's survey results. The researcher first divided the two variables into categories. All but one correlation value was greater than .5; therefore, it was decided that correlations between a student's final grade and NC EOC score with r values greater than or equal to .7 would be coded as a "yes" and those less than .7 would be coded as a "no". The second variable, mean value of teacher's survey results, was categorized into mean values greater than or equal to 3.5, coded as "yes", and mean values less than 3.5, coded as "no". Tables 23 – 28 represent the contingency tables for the six different comparisons.

The results of the Fisher's Exact Test for each of the contingency tables are noted in Table 29. There were no significant relationships in any of the comparisons. The null hypothesis was accepted.

Summary

While the Fisher's Exact test indicated that there was no significant association between the correlation of a student's grade with his/her NC EOC score and a teacher's responses on the survey for his/her respective class, the findings in the various analyses of the survey responses yielded some interesting results. Chapter 5 will provide conclusions based on results of these analyses as well as implications for educational leaders. Lastly, the next chapter will also provide recommendations for future research regarding this topic.

Table 23

Contingency Table for Mean Value of Teacher Survey Responses for All Statements and r Values with Honors and Regular Courses Combined

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	0	2	2
Yes	3	3	6
Total	3	5	8

Table 24

Contingency Table for Mean Value of Teacher Survey Responses for Statements related to Nonacademic behaviors and r Values with Honors and Regular Courses Combined

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	1	2	3
Yes	2	3	5
Total	3	5	8

Table 25

Contingency Table for Mean Value of Teacher Survey Responses for All Statements related to and r Values for Regular Courses Only

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	1	1	2
Yes	3	3	6
Total	4	4	8

Table 26

Contingency Table for Mean Value of Teacher Survey Responses for Statements related to Nonacademic Behaviors and r Values for Regular Courses Only

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	2	1	3
Yes	2	3	5
Total	4	4	8

Table 27

Contingency Table for Mean Value of Teacher Survey Responses for All Statements and r Values for Honors Only

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	1	1	2
Yes	2	0	2
Total	3	1	4

Table 28

Contingency Table for Mean Value of Teacher Survey Responses for Statements Related to Nonacademic Behaviors and r Values for Honors Only

	Strength of Relationship		
Grading System Aligned	Low	High	Total
No	1	1	2
Yes	2	0	2
Total	3	1	4

Table 29

Results of Fisher's Exact Test for Contingency Tables 23-28

Table	Chi-Square Value	p-value*	Significant Yes/No
-Mean value for statements 1-22 and r values for honor and regular combined (Table 23)	1.600	.464	No
Mean value for statements 1-18 (nonacademic) and r values for honor and regular combined (Table 24)	.036	1.000	No
Mean value for statements 1-22 and r values for regular courses only (Table 25)	.000	1.000	No
Mean value for statements 1-18 (nonacademic) and r values for regular courses only (Table 26)	.533	1.000	No
Mean value for statements 1-22 and r values for honors courses only (Table 27)	1.333	1.000	No
Mean value for statements 1-18 (nonacademic and r values for honors courses only (Table 28)	1.333	1.000	No

Note. *For all Fisher's Exact Test a significance level of .01 was used.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

Overview

This study built on an earlier study of grading practices in Wake County. Boykin (2010) explored the relationship between high school course grades and end-of-course (NC EOC) scores. Among other findings, Boykin (2010) found that the current methods for grading were not always a reliable predictor of student's performance on the NC EOC. This finding raised concerns about the validity of grading practices and the degree to which the hidden curriculum (student behaviors and attitudes) influenced the calculation of a student's final grade. The purpose of this study was to determine whether grading systems that minimized the hidden curriculum and were aligned with researched best practices would yield more consistent results in determining how well students would perform on an NC EOC. In order to establish these findings, participating teachers first completed a survey to determine how aligned their practices were with research. Then the strength of the relationship between student's grades and their performance on the NC EOC in each of the teacher's classes was established. The final analysis was the comparison of the teachers' survey score to the correlations established between their students' grades and their students' NC EOC performance.

Summary of Related Research

Experts in the field of grading systems tended to agree that the primary purpose of grades should be to communicate student achievement (Brookhart, 2009; Guskey & Bailey, 2001; O'Connor, 1995). Ken O'Connor (1995) provided guidelines for constructing a valid grading system. The guidelines originated with Gronlund and Linn (1990) but were modified considerably with Rick Stiggins having the greatest influence (O'Connor, 1995). O'Connor (1995) also noted that these guidelines were "consistent with the fair assessment practices codes

proposed for use in both the United States and Canada" (p. 96). There were eight guidelines but for the purposes of this study emphasis was placed on the first guideline when an analysis of the teacher survey was conducted. This guideline suggested that the teacher limit the valued attributes included in grades to achievement. In other words teachers should have excluded things such as evaluations of neatness, effort, working with others, and punctuality in the achievement grade. However, it should be noted that Marzano (2000) pointed out that most groups believed that the teaching of nonachievement factors should not be excluded from school. In a survey conducted by the polling firm Public Agenda, 88% of those surveyed believed that schools should teach such skills as punctuality, dependability, and self-discipline (Marzano, 2000, p. 38). The challenge was creating a system that measured both achievement and nonachievement factors but reported the two evaluations separately in order to have a clear picture of performance in both areas. Marzano (2000), Brookhart (2009), Guskey and Bailey (2001), and O'Connor (2002) were just a few of the researchers that provided examples of how to create rubrics/grading systems that assisted teachers in effectively measuring nonachievement factors in order to maintain the integrity of their grading system.

Each of the teachers in this study participated in an on-line survey that was developed using best practices as defined by the literature (Brookhart, 2009; Guskey & Bailey, 2001; Marzano, 2000; O'Connor, 2007). The analyses of the survey responses indicated how well each of the teachers in this study adhered to a grading system supported by these researchers.

Review of Methodology

The teachers participating in this study were selected because they claimed to have established a grading system that was more in line with researched best practice. In addition each participant taught a course where students were required to take an NC EOC. This criterion was

necessary in order to answer the second research question. Each of the teachers participated in a survey to determine how aligned their practices were with research. Statements 1-18 represented items that sought to determine to what degree a teacher removed nonacademic factors form his/her grading system and while the remaining four statements were not directly related to nonacademic factors they were noted in the research as factors that should be considered when creating a grading system. To examine teachers' survey responses, average scores were obtained from teacher responses for statements 1-18 and again for statements 1-22. These average scores established the first variable needed for the Fisher's Exact Test.

Additional analyses of the survey were conducted to determine if there were practices more commonly adhered to than others by all teachers or if the teacher practices from one school were more closely aligned than the teacher practices from another school. For example, three of the items in the survey addressed cheating. An average for these three items was obtained to look more closely at this specific practice. In looking at each of the averages calculated, values closer to five represented practices more aligned with research.

A second variable was established by calculating a Pearson r coefficient for each of the teacher's classes. This was done to determine the strength of the relationship between a student's grade and his or her performance on the NC EOC. Values greater than or equal to .5 were considered significant.

Once the two variables were established, a series of Fisher's Exact Tests were calculated to look at the relationship between the two categorical variables; correlation between a student's final grade and NC EOC score and mean values form the teacher survey responses. Because all but one correlation value was greater than .5 it was decided that values greater than or equal to .7 would be considered strong and those less than .7 were not. The second variable, mean value of

teacher's survey results, was categorized into mean values greater than or equal to 3.5, indicating a system more closely aligned with research, and mean values less than 3.5, less aligned with research.

Research Questions

The research questions addressed by this study were as follows:

1. Of the grading systems reviewed which teacher's grading system aligned more closely with research best practice?
2. Did teachers' grading systems with the greatest alignment to research best practices produce a stronger relationship between students' grades and their NC EOC performance?

The null hypothesis tested in this research was teachers grading systems that were strongly aligned with research best practice will not make a difference in the strength of the relationship between their students' grades and their NC EOC performance.

Analysis of the Data

The initial step in the analysis of data was to explore the teacher survey results using descriptive statistics. A series of mean values were calculated with averages ranging from 1-5, with 1 representing a grading system not closely aligned with research and 5 representing a grading system closely aligned with research. What follows is an explanation of the findings.

Research Question 1

Teachers' grading systems. When looking at the teachers' overall grading systems, the results indicated that the teachers in this study generally adhered to practices supported by the research. This finding was supported by the fact that the mean value calculated for each teacher's survey responses was ≥ 3.33. There was, however, one teacher whose score was considerably

stronger than the others, Teacher E. This teacher had a mean value of 4.73 out of a possible score of 5 for questions 1-22. The next highest mean value for questions 1-22 was 3.9. This score was from Teacher G. The initial review of survey results indicated that the grading system of Teacher E was more closely aligned with the research.

Elimination of valued attributes. One of the recommendations supported by most researchers in this study was the need to remove or limit the grading of valued attributes, or as noted in this study, nonacademic behaviors. Statements 1-18 in this study addressed to what degree the teacher removed these features from his/her grading system. In order to determine how closely aligned these teachers were with this guideline a second average was calculated for questions 1-18. It was found that the average for six of the eight teachers went down slightly. One teacher gained almost half a point when only considering these features of her grading system, Teacher C. The other teacher's value, Teacher A, was only slightly elevated by three hundredths of a point. When the questions dealing with the removal of nonacademic behaviors from a grading system were isolated the teachers grading system seemed to move further away from what research supports. This was true for Teacher E as well but he still maintained the highest average.

In order to take a closer look at each of the practices evaluated in this survey, mean values were calculated for each of the statement responses. In addition to this analysis those statements that addressed the same attribute were clustered and the mean value was obtained to see how closely aligned a teacher's practice was with the research for a particular guideline. For example when looking at the issue of late work item by item, teachers seemed to generally agree that giving a zero was not acceptable (mean value was 4.875). However, when it came to how much credit a student should receive for late work, assessed in statements 2 and 3, the mean

values moved further away from what research supported (mean value for statement 2 is 2.375 and the mean value for statement 3 is 3.875). When analyzing the issue of late work by teacher, the responses for all three items were averaged and it was determined that none of the teachers adhered closely with best practice. The highest average for this particular attribute was 3.67, scores for Teacher E and F. There were eleven valued attributes/nonacademic behaviors being addressed in statements 1-18: (1) late work (questions 1-3); (2) cheating (questions 4-6); (3) extra credit (questions 7-8); (4) group work (questions 9-11); (5) disruptive behavior (question 12); (6) participation (question 13); (7) neatness (questions 14); (8) academic growth (question 15); (9) academic ability (question 16); (10) effort (question 17); and (11) professional judgment (question 18) . The same analysis as indicated above for late work was conducted for each attribute.

As noted in Chapter 4, a review of the mean values of both individual statements and those clustered by attribute indicated that the majority of the teacher responses were aligned with research in the areas of disruptive behavior, participation, neatness, and effort. For these particular attributes there was only one item that assessed the teacher's practice therefore an additional mean score was not needed. It should be noted that Teachers E, F, and G all had scores of 5 for each of these items. A review of the remaining attributes yielded noticeable discrepancies in how the teacher's addressed the issues within his/her grading system.

There was considerable variability amongst the collective responses of these eight teachers as well as within individual teacher responses when reviewing the mean values that dealt with the matter of cheating. In looking at the mean values from the other attributes, even when a teacher's response may not be in line with the research it was at least clear on what position he/she was taking. All but one teacher really struggled to maintain a position on how to

deal with cheating as reflected in their responses. Teacher E was clear in that he did not give zeros for cheating and he always allowed a student to retake the test for full credit.

Responses to statements seeking to determine how a teacher dealt with extra credit and group work were most consistent as far as how well the practices aligned with the research. All but one teacher agreed that extra credit should not be awarded for nonacademic tasks. Where they began to differ was in the assignment of extra credit for academic tasks. When the review of all questions dealing with extra credit was conducted only three of the eight teachers' practices aligned with the research, Teachers D, E, and G. None of these teachers provided extra credit for academic or nonacademic tasks.

Most teachers believed that when assigning grades for a group task both an individual and group grade should be assigned. However, the research indicated that the group evaluation was a valued attribute, ability to work well in groups or collaborate, and therefore should be evaluated separate from the academic evaluation. Three teachers adhered strictly to this best practice, Teachers E, F, and H.

In examining the remaining statements regarding consideration of student's academic growth or ability when assigning a grade, or whether to use professional judgment when determining a grade it was found that the teachers were split down the middle on these issues. The same four teachers (Teachers A, C, E, and G) that did not consider academic ability either at all or the majority of the time also did not consider academic growth at all or the majority of the time. The remaining four teachers were inconsistent in their responses indicating no alignment with best practice. Three of the four teachers noted above taught at the same school (Teachers A, E, and G). While these four teachers were consistent in their responses to two of the statements they began to vary in their responses when it came to professional judgment. Only two of the

previous four noted did not consider professional judgment at all or the majority of the time when assigning grades, Teachers A and G. Research would suggest that none of these values be considered. There was one teacher that adhered to best practice in all three responses, Teacher G.

When looking at the overall findings, with the exception of the use of professional judgment, Teacher E consistently adhered to practices that were most aligned with the research. This was also supported by the overall average of his survey responses in that Teacher E had the highest average for questions 1-18 as well as questions 1-22.

Other guidelines for grading systems. Statements 19-22 of the survey addressed other guidelines O'Connor noted as critical when constructing a valid grading system.

O'Connor (1995) suggested that not everything a child produced in a classroom to demonstrate his/her learning should be graded. One such product would be an assessment that is formative. Shute (2008) and Brookhart (2009) indicated that formative assessments were intended to be a means of providing feedback to the learner for the purposes of modifying his/her thinking or behavior for the purpose of improving learning. Brookhart (2009) went on to note that it was the summative assessment that should be used to judge the outcome of the development of the learner. In short, formative should not be graded and summative should be graded. Statement 19 sought to determine which teachers adhered to this suggestion. Six out of eight of the teachers did not include formative assessments in their grades, Teachers A, B, E, F, G, and H. One of the eight teachers did some of the time (Teacher C) and the last teacher, Teacher D, varied from 40- 79% of the time as to when she included formative assessments in her grading system.

Statement number twenty addressed the guideline of grading in pencil, allowing students to retest in order to demonstrate mastery of the content (O'Connor, 1995). The responses to this

statement were similar to number nineteen in that six out of eight of the teachers, not the same six, allowed students to retake tests, one did some of the time (Teacher G), and the eighth teacher, Teacher A, varied as to when she allowed retakes.

The third guideline addressed in this set of statements was the use of a curve when assigning grades. A host of researchers (Glasser, 1992; Guskey, 2009; Johnson et al., 1979; Krumboltz & Yeh, 1996; O'Connor, 1995; Zoeckeler, 2007) believed this to be an unacceptable practice. It is believed that students should be evaluated on what they have learned and are able to do (criterion-referenced standards) not their relative standing amongst their classmates (norm-referenced standards). Grading on a curve has proved to have a negative impact on student teacher relations, student to student relationships and performance on achievement test (Johnson et al., 1979; Krumboltz & Yeh, 1996). A fourth of the teachers surveyed reported using a curve when assigning grades 10-39% of the time (Teachers B and C), another fourth used a curve 40-79% of the time (Teachers F and H) and the remaining half used a curve little to none of the time, less than 10% (Teachers A,D, E, G). When trying to understand why these teachers may have used a curve it is important to remember Zoeckeler's (2007) finding of the pressure teachers felt to keep their failure rate within an acceptable range.

The last guideline addressed in this section of the survey was based on the belief that a student's knowledge and/or ability builds throughout a course; therefore, marks earned earlier in the year should be replaced by those earned later in the year (O'Connor, 2007). This statement had the greatest number of teacher responses in opposition to best practice. Only one teacher adhered to this practice, Teacher E. Two teachers, teachers D and H, did this some of the time (10-39%) but four of the teachers did this little to none of the time, teachers A, B, C, and G. It should be noted that each of the teachers in this study used an electronic grade book provided by

Wake County Public Schools. While it can be done, setting up the electronic grade book to reflect this guideline would require a great deal of work at the system level.

In reviewing the responses from statements 19-22 Teacher E was the only teacher that consistently maintained practices that were aligned with the research. Therefore, this portion of the research study determined that Teacher E's grading system aligned more closely with researched best practice.

Responses by Professional Learning Teams (PLTs). Teachers within Wake County Public Schools have participated in professional development centered on the development of effective PLTs. One of the goals of an effective PLT was to establish a shared vision and values. Each of the teachers in this study had participated in PLT training and claimed to have established a shared vision and values with regards to their grading practices. This was one of the factors taken into consideration when they were chosen for this study. In order to ascertain how well these PLTs adhered to their claim of alignment of grading practices, mean values by statement were calculated for each PLT. In looking at the results, the Physical Science PLT had greater alignment with strong agreement in thirteen of the twenty-two statements, values ≥ 3.5, and of the eighteen that dealt with nonacademic behaviors there was strong agreement in ten of those statements. The Biology PLT had strong agreement in only nine of the twenty-two statements with seven of those statements falling within the first eighteen. The nine statements in which there was strong agreement within the Biology PLT were also strong within the Physical Science PLT: (1) not giving zeros for late work; (2) not providing extra credit for nonacademic tasks; (3) giving group and individual grades for group tasks; (4) not lowering a grade for disruptive behavior; (5) not giving class participation grades; (6) not considering neatness when grading; (7) not considering effort when assigning a grade; (8) not grading formative assessment;

and (9) allowing students to retake assessments to improve their grades. Yet, as noted earlier, the Physical Science PLT had agreement in four additional statements: (1) not offering extra credit for academic tasks, (2) not considering a student's academic growth when assigning grades, (3) not considering a student's academic ability when assigning a grade, and (4) not grading on the curve. It was also interesting to note that of those statements agreed upon within each PLT, there was only one statement in the Biology PLT in which the score reflected 100% agreement (allowing students to retake assessments to improve grades) and there were four with 100% agreement in the Physical Science PLT (not offering extra credit for academic tasks, not lowering grades for disruptive behavior, not giving class participation grades, and not grading on the curve).

The final analysis of the survey revealed how well the teachers within the PLT believed their PLT collaborated with regards to creating a unified grading system. Specifically the teachers were asked to rate the degree of alignment on the following statements: (1) My PLT has a unified grading system and (2) I use the same grading procedure as others in my PLT on graded work. While neither PLT believed their group to be 100% aligned the Physical Science PLT's scores reflected a stronger alignment. This corresponds with the earlier findings in that the Physical Science PLT had more practices aligned with research and more statements in which they were 100% in agreement. It should also be noted that Teacher E, the teacher with the greatest alignment to research, was a member of this PLT.

Research Question 2

To answer the second research question correlations were calculated to first establish the relationship between student grades and their NC EOC scores. The correlation was significant at the .01 level (2-tailed). Next a series of Fisher's Exact Test were calculated to determine if there

was an association of significance between a student's final grade and NC EOC score and the mean value of teacher's survey results.

Relationship of student grades to NC EOC scores. Correlations were calculated to examine the relationship between a student's final grade and his or her performance on the NC EOC. For the Biology courses, separate correlations were calculated for the honors levels and academic levels. There were no honors level courses for Physical Science. When looking at all the correlation values, not taking into consideration subject or level, all but one value was significant. The correlation value for Teacher H was not significant for her academic biology classes, $r = .469$. Teacher A had the strongest relationship, $r = .822$, for her academic physical science class.

While all of the correlation values for the physical science classes were significant, there were clearly two teachers with much stronger values than the other two members. Teachers A and G had r values of .822 and .755 respectively and Teachers D and E had r values of .599 and .565 respectively.

As mentioned earlier, all but one of the Biology classes yielded significant results. Teachers B and F had stronger correlation values in their academic biology classes, $r = .741$ and .765 respectively. For Teachers C and H, the correlation values computed to .588 and .469 respectively.

When comparing the academic biology classes to the honors biology classes there was no considerable difference in the correlation values for Teacher B and C but the correlation value for Teacher F dropped considerably for her honors class. The r value went from .765 to .61. For Teacher H the opposite effect was observed. The r value from her regular to honors class went from .469 to .56, moving it from a nonsignificant relationship to a significant one.

As a reminder, the analysis of the survey results clearly indicated that Teacher E demonstrated the greatest alignment to researched best in his grading practices. A review of these data showed Teacher A with the strongest relationship whereas the relationship between grades and NC EOC scores for Teacher E was the second lowest.

Comparison of survey results to correlations. Prior to calculating the series of Fisher's Exact Tests the information for these two variables, mean values from the survey and correlation values, was placed into a table. In comparing these values, some interesting observations were made. The teacher with the highest mean average for the survey responses, Teacher E, had the second lowest *r* value when taking into consideration all the academic level courses. The other interesting observation was made in the comparison of these variables for Teachers F and G. For each variable, both teachers maintained within a second or third ranking. Specifically, Teacher F had the second highest *r* value for her academic and honors biology classes (r = .765 and r =.61 respectively) and ranked third in the mean values for her survey responses (3.78 for questions 1-18 and 3.82 for questions 1-22). Teacher G had the third highest *r* value for her academic physical science class (r =.755) and ranked second in the mean value for her survey responses (3.89 for questions 1-18 and 3.91 for questions 1-22). In terms of consistency, the results of the correlation values for these two teachers were in line with their ranking on the survey.

Fisher's Exact Test. In order to determine if there was an association of significance between these two categorical variables, correlation between a student's final grade and his/her NC EOC score and the mean value of a teacher's survey results, a series of Fisher's Exact test were calculated. In Chapter 4 the rationale for the categorical division for each of the variables was explained. In short, because the *r* values for this study were all high it was decided that values greater than or equal to .7 were considered as a strong relationship between students'

grades and NC EOC performance. Survey responses with mean values greater than or equal to 3.5 were considered aligned with research best practices. Six different comparisons were calculated: (1) Mean values of all teacher survey responses for statements 1-22 to r values of all courses; (2) Mean values of all teacher survey responses for statements 1-18 (nonacademic) to r values of all courses; (3) Mean values of all teacher survey responses for statements 1-22 to r values of regular courses only; (4) Mean values of all teacher survey responses for statements 1-18 to r values of regular courses only; (5) Mean values of all teacher survey responses for statements 1-22 to r values of honors courses only; and (6) Mean values of all teacher survey responses for statements 1-18 to r values of honors courses only. None of the comparisons yielded a significant result. In other words, there was no relationship that existed between a teacher's method of grading, as determined by the responses to the survey, and the correlation values that existed between their students' grades and their performance on the NC EOC. Any observed result can only be ascribed to chance alone.

Conclusions

The purpose of this study was to determine if grading systems that were aligned with researched best practices would yield more consistent results in determining how well students would perform on an NC EOC. This was explored by answering the two research questions posted. This study yielded clear answers to each of the research questions posted.

Research Question 1

The first research questions sought to determine which teacher(s) grading systems aligned closely with research best practice. It was clear that Teacher E had the strongest alignment in his practice.

Research Question 2

The second question sought to determine if teachers' grading systems with the greatest alignment to research would yield a stronger relationship between the grades of their students and their NC EOC performance. After calculating a series of Fisher's Exact Test it was determined that no association of significance exists between the two categorical variable noted. The null hypothesis was accepted.

Limitations

This research was highly dependent on the acquisition of a group of teachers from the same school, teaching the same subject, and claiming to have grading systems that were aligned with researched best practice. Because of this, the pool of teachers was limited. There was a clear limitation in this study in that there were only eight teachers. Given this, the study would have been strengthened if it included a qualitative piece allowing more insight into the beliefs of each teacher.

The data for Teacher G was diminished significantly when it was learned that he did not maintain numerical averages for his students for the 2009-10 school year. Therefore, the researcher was limited to the 2010 fall semester data for this particular teacher.

A substantial portion of this research was built around teacher responses to a survey with no way to verify the validity of the responses. Therefore, the possibility exists that respondents may have misrepresented their practices via their responses.

The answer to the second research question was contingent upon the outcomes of the series of Fisher's Exact Tests. This analysis did not conduct individual comparisons and when the data was reviewed in table format, teacher by teacher, there was evidence that suggested a

relationship might exist for some teachers. Individual analysis of some sort might have yielded additional findings.

Lastly, the survey was established after reviewing the literature of a limited pool of researchers. Although it was piloted prior to the year this study began, a more thorough review of the literature yielded ways in which the instrument could have been made even stronger.

Implications and Recommendations

Implications for Instructional Practice

Given that statistically significant relationships were found between a student's grade and his/her performance on the NC EOC in all classes but one suggested the importance of carefully considering what goes into the calculation of a student's grade.

For the past three years, staff of the Wake County Public School System has participated in professional development regarding effective grading systems. The emphasis has been at the secondary level; therefore, the teachers in this study have had the opportunity to participate in professional development centered on this topic. These teachers professed to adhere to grading practices that were aligned with research best practice; however, in many cases their survey responses indicated otherwise. For example, giving a zero for cheating was not supported by the research yet, there were only two teachers that consistently adhered to this practice. Similar results were seen for the practice of considering neatness when assigning a grade. Three of the teachers considered neatness when assigning grades whereas the research suggested that this should not be included in measuring academic performance.

The survey results indicated that there was considerable variability among teachers within the same professional learning team. For example, while agreement was established in not assigning zeros to work turned in late, teachers do not agree on the amount of credit to give this

assignment. Two students in two different classes poised in the same circumstance could receive different grades, even if the quality of their work was equivalent.

The variability between teacher practices within a school and across schools has a direct impact on the learner. A student's measure of academic achievement could vary depending on which teacher evaluates his/her performance. In effect, stakeholders could become less confident that the grade the student received was an accurate reflection of his/her learning.

Professional development is needed in helping teachers understand the purpose of grades, the implications of inconsistent grading for all stakeholders, and how a reliable and valid grading system is constructed.

Implications for District Leaders

While some of the comparisons between the two variables (teacher mean values on survey responses and *r* values for students' grades and NC EOC performance) appeared to be related when reviewing the data in Table 22, the results of the Fisher's Exact Test indicated that these relationships can only be attributed to chance. When considering the implications of this practice, leaving it up to chance, as well as the noted variability in teacher grading practices, working to establish a grading system policy that yields a more consistent and predictable result in grading would be prudent for district leaders.

Future Studies

While Brookhart (1993) found that measurement instruction had very little impact on the elimination of value judgments in grades, Black and William (1998) found the opposite to be true. Austin and McCann (1992) identified in their review of policies and procedure from 144 districts that very few districts (less than 10) gave teachers adequate guidance to ensure consistent grading, and none of the districts provided staff development to teachers to improve

their grading practices. A future study incorporating the professional development noted earlier would benefit the research in determining if the training had an impact on a teacher's ability to construct a valid grading system. This study did not provide this opportunity and establishing this knowledge with one group in an experimental design to determine the impact would further the research in this area. In addition, adding a qualitative component to the study would help provide further insight into the beliefs of a teacher with regards to his/her grading system.

Four of the teachers in this study taught honors and regular levels of the same subject. When reviewing the *r* values for these levels for each of the teachers two of the teachers had values with relatively little difference; however, the other two teachers had values that were noticeably different. One teacher had a stronger correlation amongst her honors students than her regular and the other teacher had a weaker correlation amongst her honors students than her regular. It would be interesting to explore how or if the expectations of certain students impacts the construction of a teacher's grading system.

William, Pollack, and Lewis (2002) stated that "educators might do well to worry less about the statistical indices of the grade-test relationship and more about what is assessed and how that can improve teaching and learning" (p.31). Is the NC EOC the best measure for which to compare student learning? A study looking at the relationship of a student's grade to his/her performance on other measures, such as ACT or SAT, would be of benefit.

Lastly, conducting a study of the efficacy of grading practices in a district with a policy that provided explicit researched based guidelines is warranted to determine if this has an impact on consistency of classroom practice and in the accuracy of reporting student academic performance.

Summary

This study provided information on what factors teachers considered when constructing their grading system and whether or not the teacher's grading system yielded students grades that were aligned with their performance on the NC EOC. When reviewing the survey items statement by statement, it was noted that statements with stronger moral implications tended to have greater variability in the responses. For example, none of the teachers lowered grades for disruptive behavior. However, the responses for how to handle the matter of cheating were widely varied. This seemed to substantiate the notion that a teacher's grading system was very subjective. While it is nearly impossible to remove all subjectivity from a grading system there is room for improvement that could come with professional development.

The results of the Fisher's Exact Test indicated that no relationship existed between a teacher's method of grading and the correlation values that were established between his/her students' grades and -their performance on the NC EOC; yet, there was some evidence that suggested otherwise when looking at the teacher by teacher comparisons in Table 22. It would be beneficial to students, parents, and educators at all levels to pursue the examination of this relationship and to establish opportunities for educators to learn methods for improving their grading system.

Grades are universal symbols used to communicate evidences of student learning. Because they have implications for a student's future, every effort should be made to establish an effective grading system.

REFERENCES

Airasian, P. W. (1988). The case of state-mandated, high-stakes testing. *Educational Evaluation and Policy Analysis, 10*(4), 301-313.

Austin, S., & McCann, R. (1992). Here's another arbitrary grade for your collection: A statewide study of grading policies. Paper presented at the Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 343944)

Bailey, J., & McTighe, J. (1996). Reporting achievement at the secondary level: What and how. In T.R. Guskey (Ed.), *ASCD Yearbook – 1996: Communicating student learning* (pp. 119-140). Alexandria, VA: Association for Supervision and Curriculum Development.

Baron, P. A. B. (2000). Consequential validity for high school grades: What is the meaning of grades for senders and receivers? Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 445051)

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347- 356.

Bishop, J. H. (1992). Why U.S. students need incentives to learn. *Educational Leadership, 49,* 15-18.

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-144.

Boykin, A. S. M. (2010). *The relationship between high school course grades and exam scores* (E&R Report No. 09.39). Retrieved from Wake County Public Schools:http://www.wcpss.net/evaluationresearch/reports/2010/0939course_exams.pdf

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

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

Brookhart, S. M. (2009). *Grading*. Upper Saddle River, NJ: Pearson Education.

Canady, R. L., & Hotchkiss, P. R. (1989). It's a good score! Just a bad grade. *Phi Delta Kappan, 71*(1), 68-71.

Crooks, A. D. (1933). Marks and marking systems: A digest. *Journal of Educational Research,* 27, 259-272.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12,* 53-73.

Davis, W. D., Carson, C. M., Ammeter, A. P., & Treadway, D. C. (2005). The interactive effects of goal orientation and feedback specificity on task performance. *Human Performance, 18,* 409-426.

DuFour, R., & Earker, R. (1998). *Professional Learning Communities at Work: Best Practices for Enhancing Student Achievement.* Bloomington, IN: National Education Service.

Farkas, S., Friedman, W., Boese, J., & Shaw, G. (1994). *First things first: What Americans expect from public schools.* New York: Public Agenda.

Feldmesser, R. A. (1971). The positive functions of grades. Paper presented at the 55th Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 049704)

Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology, 51,* 397-420.

Ford, J. K., Smith E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83,* 218-233.

Friedman, S. J., & Troug, A. L. (1999). Evaluation of high school teachers' written grading practices. *ERS Spectrum, Summer,* 34-42.

Frisbie, D. A., & Waltman, K. K. (1992). Developing a personal grading plan. *Educational Measurement, Issues, and Practice, 11,* 35-42.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Glasser, W. (1992). *The quality school.* New York, NY: Harper-Collins.

Goldschmidt, P., & Wang, J. (1999). When can schools affect dropout behavior? A longitudinal multilevel analysis. *American Educational Research Journal, 36*(4), 715-738.

Green,K., & Emerson, A. (2007). A new framework for grading. *Assessment & Evaluation in Higher Education, 32*(4), 495-511.

Gronlund, N., & Linn, R. (1990). *Measurement and Evaluation in Teaching* (6th ed.). New York: Macmillan.

Guskey, T. (1994) Making the grade: what benefits students? *Educational Leadership, 52*(2), 14-20.

Guskey, T. (2001). Helping standards make the grade. *Educational Leadership, 58*(1), 20-27.

Guskey, T. (2002). Perspectives on grading and reporting: Differences among teachers, students, and parents. Paper presented at the Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 464113)

Guskey, T. (2004). Stability and change in high school grades. Paper Presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 490411)

Guskey, T. R. (2006). "It wasn't fair!" Educators recollection of their experiences as students with grading. Presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 492005)

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

Guskey, T. (2008). The rest of the story. *Educational Leadership, 65*(4), 28-34.

Guskey, T. (2009). Grading policies that work against standards…and how to fix them. In T. R. Guskey (Ed.), *Practical solutions for serious problems in standard-based grading* (pp. 9-26). Thousand Oaks, CA: Corwin Press.

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

Haitte, J. A. (1992). Measuring the effects of schooling. *Australian Journal of Education, 36*(2), 80-82.

Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappan, 72*(7), 540-545.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Johnson, R. H. (1918). Educational research and statistics: The coefficient marking system. *School and Society, 7*(181), 714-716.

Johnson, R. T., Johnson, D. W., & Tauer, M. (1979). The effects of cooperative, competitive, and individualistic goal structures on students' attitudes and achievement. *The Journal of Psychology, 102,* 191-198.

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

Kovas, M. A. (1993). Make your grading motivating: Keys to performance based evaluation. *Quill and Scroll, 68*(1), 10-11.

Kulhavy, R. W., & Stock, W. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279-308.

Krumboltz, J. D., & Yeh, C. J. (1996). Competitive grading sabotages good teaching. *Phil Delta Kappa, 78*(4), 324-326.

Lee, V. E., & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal, 40*(2), 353-393.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179-189.

Marzano, R. J. (2000). *Transforming Classroom Grading.* Alexandria, VA: ASCD.

McMann, G. (2003, October 15). Letter to the editor. *Toronto Globe and Mail,* n.p.

McMillan, J. H. (1999). The devastating effect of zeros on grades: What can be done? Opinion paper. Richmond, VA: Virginia Commonwealth University.

McMillan, J. H., & Lawson, S. R. (2001). Secondary science teachers' Classroom assessment and grading practices. *Metropolitan Educational Research Consortium,* (ERIC Document Reproduction Service No. 450158)

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans. (ERIC Document Reproduction Service No. 447195)

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), pp. 13-103). New York: Macmillan.

Messick, S. (1995). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

O'Connor, K. (1995). Guidelines for grading that support learning and student success. *NASSP Bulletin, 79*(571), 91-101.

O'Connor, K. (2002). *How to grade for learning: Linking grades to standards.* Thousand Oaks, CA: Corwin Press.

O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades.* Portland, OR: Educational Testing Service.

Ornstein, A. C. (1994). Grading practices and policies: An overview and some suggestions. *NASSP Bulletin, 78,* 55-64.

Reeves, D. B. (2008). Effective grading. *Educational Leadership, 65,* 85-87.

Rugg, H. O. (1918). Teachers' marks and the reconstruction of the marking system. *Elementary School Journal, 18*(9), 701-719.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education, 19,* 405-450.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research 78*(1), 153-189.

Smallwood, M. L. (1935). *An historical study of examinations and grading systems in early American universities.* Harvard Studies in Education #24. Cambridge, MA: Harvard University Press.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26-33.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39*(2), 229-233.

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

Starch, D., & Elliott, E. C. (1913). Reliability of grading work in mathematics. *School Review, 21,* 254-259.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.

Stiggins, R. J. (2000). *Specifications for a performance-based assessment system in teacher preparation.* Paper commissioned by the National Council on Accreditation in Teacher Education (NCATE), Washington, DC.

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.

Sweedler-Brown, C. O. (1992). The effect of training on the appearance bias of holistic essay graders. *Journal of Research and Development in Education, 26*(1), 24-29.

Trumball, E., & Farr, B. (2000). *Grading and reporting student progress in an age of standards.* Norwood, MA: Christopher-Gordon Publishers.

Weisstein, E. (2010). *Fisher's exact test from mathworld.* Retrieved from http://mathworld.wolfram.com/FishersExactTest.html

Willingham, W., Pollack, J., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Educational Measurement: Issues and Practices, 39*(1), 1-37.

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

APPENDIX A: NORTH CAROLINA TEACHER CONTRACTS

CONTRACT FOR PROFESSIONAL SERVICE

North Carolina Public Schools

STATE OF NORTH CAROLINA **PROBATIONARY CONTRACT**

WAKE COUNTY

THIS AGREEMENT entered into between the Wake County Board of Education, School Administrative Unit and **«Full_Name»**, who now holds, or is entitled to hold, a North Carolina Certificate in **«Cert_Area»**, Employee Identification Number **«Employee_Number»**, now in force, in accordance with and subject to the provisions of the school law applicable thereto, which are hereby made a part of this contract.

WITNESSETH:

That said licensed person, having been duly elected to perform professional services in the public schools of said school administrative unit, subject to any specific conditions set forth below, agrees to discharge faithfully all the duties imposed on such persons by the Laws of North Carolina and by the rules and regulations of the Board of Education of said school administrative unit.

While this contract is for the 2010-2011 school year, your employment as a probationary teacher will continue from year to year without the issuance of a new contract until such time that the Board decides whether or not to offer you a career (tenure) contract, unless: (a) you receive notice of nonrenewal at the end of a school year, (b) there is a break in your employment, or (c) you no longer serve as a licensed, 100% Teacher in a permanent position.

That, in consideration of this agreement, said Board of Education promises to pay the above named person for services rendered during the life of this contract the sum of which s/he is entitled according to the State Salary Schedule plus the local supplement, if any, applicable thereto, with State-supported positions being subject to the allotment of personnel by the State Board of Education and subject to the condition that the amount paid from State Funds shall be within the allotment of funds made to said administrative unit for salaries, and with federal and locally-supported positions being subject to the availability of federal and local funds, and subject further to the condition that when the position for which the employee is employed, whether a State, federal, or locally-supported position, is terminated this contract shall be terminated.

The assignments to duties will be made by the superintendent of schools.

That said Board of Education has authorized in a regular or in a called meeting its Secretary to execute this contract.

Specific Conditions: In order to be considered for continued employment, you must maintain a valid North Carolina Teaching License in your area of assignment. If the license issued is provisional you must complete the yearly requirements as outlined in writing by the State Licensure Section. Written documentation that you have met these requirements must be presented to your Human Resources Administrator by May 1 of each year.

Your base contract is for ten months. If you have been assigned additional months of employment, such additional months are considered a special duty, pursuant to N.C. General Statute 115C-325(a)(4), and any return to a ten-month position is not a demotion as defined by law. Accordingly, the school system reserves the authority to limit your assignment to ten months at any time.

WAKE COUNTY BOARD OF EDUCATION

_____
(Employee Signature)

By _____
(Board Chair)

_____
(Employee Address)

_____
(Employee Address)

_____
(Date)

_____
(Date)

Revised: 4/27/10

«SchoolDepartment»

APPENDIX B: WAKE COUNTY PUBLIC SCHOOLS TEACHER JOB DESCRIPTION

3219 JOB DESCRIPTION TEACHER 3219

Reports to: Principal

Supervises: May coordinate and direct the activities of teacher assistants

Purpose: To plan, instruct, and organize instructional environments which help students learn subject matter and skills that will contribute to their educational and social development.

DUTIES AND RESPONSIBILITIES

The Teacher is required to follow all Board Policies and procedures, and to comply with the instructions and or directives from his/her supervisor(s). In addition, the Teacher shall perform in a manner that meets the following Standards and Elements.

A. Standard I: Teachers demonstrate leadership

1. Element A: Teachers lead in their classrooms
   a. The Teacher takes responsibility for all students' learning
   b. The Teacher communicates vision to students
   c. The Teacher uses data to organize, plan, and set goals
   d. The Teacher uses a variety of assessment data throughout the year to evaluate progress
   e. The Teacher establishes a safe and orderly environment
   f. The Teacher Empowers students
2. Element B: Teachers demonstrate leadership in the school
   a. The Teacher works collaboratively with all school personnel to create a professional learning community
   b. The Teacher analyzes data
   c. The Teacher develops goals and strategies through the school improvement plan
   d. The Teacher assists in determining school budget and professional development
   e. The Teacher participates in hiring process
   f. The Teacher collaborates with colleagues to mentor and support teachers to improve effectiveness
3. Element C: Teachers lead the teaching profession
   a. The Teacher strives to improve the profession
   b. The Teacher contributes to the establishment of positive working conditions
   c. The Teacher participates in decision-making structures
   d. The Teacher promotes professional growth
4. Element D: Teachers advocate for schools and students
   a. The Teacher advocates for positive change in policies and practices affecting student learning
   b. The Teacher participates in the implementation of initiatives to improve education
5. Element E: Teachers demonstrate high ethical standards
   a. The Teacher demonstrates ethical principles
   b. The Teacher upholds the Code of Ethics and Standards for Professional Conduct

B. Standard II: Teachers establish a respectful environment for a diverse population of students

1. Element A: Teachers provide an environment in which each child has a positive, nurturing relationship with caring adults
   a. The Teacher encourages an environment that is inviting, respectful, supportive, inclusive, and flexible
2. Element B: Teachers embrace diversity in the school community and in the world
   a. The Teacher demonstrates knowledge of diverse cultures
   b. The Teacher selects materials and develop lessons that counteract stereotypes and incorporate contributions
   c. The Teacher recognizes the influences on a child's development, personality, and performance
   d. The Teacher considers and incorporates different points of view
3. Element C: Teachers treat students as individuals
   a. The Teacher maintains high expectations for all students
   b. The Teacher appreciates differences and values contributions by building positive, appropriate relationships
4. Element D: Teachers adapt their teaching for the benefit of students with special needs
   a. The Teacher collaborates with specialists
   b. The Teacher engages students and ensures they meet the needs of their students through inclusion and other models of effective practice
5. Element E: Teachers work collaboratively with the families and significant adults in the lives of their students
   a. The Teacher improves communication and collaboration between the school and the home and community
   b. The Teacher promotes trust and understanding and builds partnerships with school community
   c. The Teacher seeks solutions to overcome obstacles that prevent family and community involvement

C. Standard III: Teachers know the content they teach

1. Element A: Teachers align their instruction with the North Carolina Standard Course of Study
   a. The Teacher teaches the North Carolina Standard Course of Study
   b. The Teacher develops and applies strategies to make the curriculum rigorous and relevant
   c. The Teacher develops literacy skills appropriate to specialty area
2. Element B: Teachers know the content appropriate to their teaching specialty
   a. The Teacher knows subject beyond the content they teach
   b. The Teacher directs students' curiosity into an interest in learning
3. Element C: Teachers recognize the interconnectedness of content areas/disciplines
   a. The Teacher knows links between grade/subject and the North Carolina Standard Course of Study
   b. The Teacher relates content to other disciplines
   c. The Teacher promotes global awareness and its relevance
4. Element D: Teachers make instruction relevant to students
   a. The Teacher incorporates life skills: leadership, ethics, accountability, adaptability, personal productivity, personal responsibility, people skills, self-direction, and social responsibility
   b. The Teacher demonstrates the relationship between the core content and 21st Century content, including global awareness; financial, economic, business and entrepreneurial literacy; civic literacy; and health and wellness awareness

D. Standard IV: Teachers facilitate learning for their students

1. Element A: Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students
   a. The Teacher knows how students think and learn
   b. The Teacher understands the influences on student learning and differentiate instruction
   c. The Teacher keeps abreast of evolving research
   d. The Teacher adapts resources to address the strengths and weaknesses of students
2. Element B: Teachers plan instruction appropriate for their students
   a. The Teacher collaborates with colleagues
   b. The Teacher uses data for short and long range planning
   c. The Teacher engages students in the learning process
   d. The Teacher monitors and modifies plans to enhance student learning
   e. The Teacher responds to cultural diversity and learning needs of students
3. Element C: Teachers use a variety of instructional methods
   a. The Teacher chooses methods and materials as they strive to eliminate achievement gaps
   b. The Teacher employs a wide range of techniques using information and communication technology, learning styles, and differentiated instruction
4. Element D: Teachers integrate and utilize technology in their instruction
   a. The Teacher knows appropriate use of technology to maximize student learning
   b. The Teacher helps students use technology to learn content, think critically, solve problems, discern reliability, use information, communicate, innovate and collaborate
5. Element E: Teachers help students develop critical thinking and problem-solving skills
   a. The Teacher encourages students to ask questions, think creatively, develop and test innovative ideas, synthesize knowledge and draw conclusions
   b. The Teacher helps students exercise and communicate sound reasoning; understand connections; make complex choices; and frame, analyze, and solve problems
6. Element F: Teachers help students work in teams and develop leadership qualities
   a. The Teacher teaches the importance of cooperation and collaboration
   b. The Teacher organizes learning teams in order to help students define roles, strengthen social ties, improve communication and collaborative skills, interact with people from different cultures and backgrounds, and develop leadership qualities
7. Element G: Teachers communicate effectively
   a. The Teacher communicates clearly with students in a variety of ways
   b. The Teacher assists students in articulating thoughts and ideas clearly and effectively
8. Element H: Teachers use a variety of methods to assess what each student has learned
   a. The Teacher uses multiple indicators, both formative and summative, to evaluate student progress
   b. The Teacher uses assessment systems to inform instruction and demonstrate evidence of students' 21st Century knowledge, skills, performance, and dispositions
   c. The Teacher provides opportunities for self-assessment

E. Standard V: Teachers reflect on their practice

1. Element A: Teachers analyze student learning
   a. The Teacher thinks systematically and critically about learning in their

classroom: why learning happens and what can be done to improve student achievement

b. The Teacher collects and analyzes student performance data to improve effectiveness

2. Element B: Teachers link professional growth to their professional goals
   a. The Teacher participates in continued, high quality professional development
3. Element C: Teachers function effectively in a complex, dynamic environment
   a. The Teacher actively investigates and consider new ideas that improve teaching and learning
   b. The Teacher adapts practice based on data

Legal Reference: G.S. 115C-326; State Board of Education Policy TCP-C-004
Adopted: October 4, 1982
Revised: June 17, 1991
Revised: May 21, 2002
Revised: July 20, 2010

APPENDIX C: ALTERNATE GRADING SYSTEMS FROM *WAD-JA-GET?*

(KIRSCHENBAUM, NAPIER, AND SIMON, 1971)

Alternate System	Advantages	Disadvantages
Written Evaluations	• More helpful to students, parents, and admission officers than letter or number grades • Encourage the teacher to think more about each student as an individual • Encourages on-going attention to student needs, better school-community relations and parental responses	• Allows even more subjectivity from teachers • Not all teachers know how to write meaningful evaluations • More time-consuming • Create extra work for the school's records office
Self-Evaluation	• Important learning experience for students • Teachers using this type of system have found student to be very fair and objective and often harder on themselves than the teacher would be • Tends to encourage students to take more responsibility in setting educational goals	• Novelty for students tends to wear off and they take the exercise less seriously • In situations where students do not respect their teacher they tend to abuse the opportunity of grading themselves
Give Grades but Don't Tell the Students	• Once students get use to the idea, tension over grades decreases • Students stop comparing themselves to one another and begin to shift their focus towards learning	• Initially, might increase tension • Many of the problems of traditional grading remain

Contract System	• Student anxiety is eliminated in that they know from the beginning what he/she must do to obtain the grade he/she wants • If a teacher provides explicit details, some of the subjectivity is eliminated • Encourages differentiated instruction when each child is considered	• The quality of work is over-emphasized in contracts and tends to be the sole basis for a grade • Difficult t find creative ways to measure the quality of the different types of work students may contract to do • Danger that teachers will be too ambiguous in attempting to state the qualitative distinctions between grades
Mastery Approach or Performance Curriculum (Five-Point System)	• A student's grade becomes more meaningful to him because it is tied to a performance level • Much of the teacher's subjectivity in grading is eliminated • When students know where they are headed they are likely to get there faster • The focus of this system is on success • The students has freedom to pursue his own path of mastering the course content • The teacher is held accountable for stating his objectives, providing resources and helping students obtain mastery • In the performance curriculum, the cooperation among teachers can generate better morale and the sharing of resources	• Requires considerable skill on the part of teachers and administrators • Limits a teacher's freedom to run his classes in just his own way • Possible for teachers to use the mastery approach without allowing student to pursue their own ways of achieving proficiency • Mastery approach can discourage students from setting and working towards their own goals • Faculty must be involved in setting up a performance curriculum

Pass/Fail Grading	• Students are more relaxed, less anxious, and less competitive • Better learning environment • No point to cheating • Freed from the pressures of traditional grading, some students do more work than is required	• Some teachers will you pass/fail grading as an excuse to avoid all evaluations • Does not distinguish between students of different abilities • Freed from the pressures of traditional grading, some students do less work than is required • Difficult for teachers to establish the level of mastery needed to earn the passing grade • No help to students who perform poorly – same pressures exist
Credit/No Credit Grading	• Same as those for pass/fail with one additional advantage – "No credit" does not connote failure	• Same as those for pass/fail with the exception of the last bullet
Blanket Grading	• Same as those for pass/fail	• Same as those for pass/fail • This system would violate most school's written or unwritten grading policies

APPENDIX D: SAMPLE REPORT CARD (MARZANO, 2000)

FIGURE 7.5
A Report Card with Academic and Nonacademic Grades

Name:	Al Einstein
Address:	1111 E. McSquare Dr.
City:	Relativity, CO 80000
Grade Level:	11
Quarter:	Fourth

Course Title	Overall	Academic	Nonacademic
Algebra II & Trigonometry	B+	C	A
Advanced Placement Physics	A	A	B
U.S. History	C+	B	C
American Literature	A	A	A
Physical Education	A–	A	B+
Chorus	B–	C	A
Geography	B	B	B
Current GPA:	**3.33**	**3.24**	**3.43**

Standards Rating

Algebra II & Trigonometry		(1) (2) (3) (4)
Mathematics Standard 1:	Numeric Problem Solving	2.5
Mathematics Standard 2:	Computation	
Mathematics Standard 3:	Measurement	3.0
Mathematics Standard 4:	Geometry	
Mathematics Standard 5:	Probability	
Mathematics Standard 6:	Functions	3.0
Mathematics Standard 7:	Data Analysis	2.5
Reasoning Standard:	Decision Making	
Communication Standard:	Written	
Communication Standard:	Oral	
Nonachievement Factor:	Effort	3.5
Nonachievement Factor:	Behavior	4.0
Nonachievement Factor:	Attendance	3.5
Academic: 2.75	**Overall: 3.21**	**Nonacademic: 3.67**

Advanced Placement Physics		(1) (2) (3) (4)
Science Standard 1:	Structure/Properties of Matter	3.5
Science Standard 2:	Energy Types	4.0
Science Standard 3:	Motion	4.0
Science Standard 4:	Forces	3.75
Reasoning Standard:	Experimental Inquiry	4.0
Reasoning Standard:	Problem Solving	
Communication Standard:	Audience	
Communication Standard:	Oral	
Nonachievement Factor:	Effort	3.75
Nonachievement Factor:	Behavior	3.0
Nonachievement Factor:	Attendance	3.0
Academic: 3.85	**Overall: 3.55**	**Nonacademic: 3.25**

U.S. History		(1) (2) (3) (4)
History Standard 1:	Civilization and Society	
History Standard 2:	Exploration & Colonization	3.5
History Standard 3:	Revolution and Conflict	1.5
History Standard 4:	Industry and Commerce	
History Standard 5:	Forms of Government	2.5
Reasoning Standard:	Comparing and Contrasting	2.75
Reasoning Standard:	Problem Solving	
Communication Standard:	Written	4.0
Nonachievement Factor:	Effort	3.0
Nonachievement Factor:	Behavior	1.75
Nonachievement Factor:	Attendance	2.25
Academic: 2.85	**Overall: 2.59**	**Nonacademic: 2.33**

APPENDIX E: SAMPLE HIGH SCHOOL REPORT CARD

(BROOKHART, 2009)

FIGURE 7-4 Example of a high school report card

WASHINGTON HIGH SCHOOL

NAME	GRADE LEVEL		SCHOOL YEAR		STUDENT NO.

		1ST			2ND			3RD			4TH			FNL EX GRD	FINAL GRADE
COURSE	TEACHER	GR	COMM	CL ABS	GR	COMM	CL ABS	GR	COMM	CL ABS	GR	COMM	CL ABS		

	CREDITS ATTEMPTED	CREDITS EARNED	QUALITY POINTS	QUALITY POINT AVG.		1ST	2ND	3RD	4TH	CUM
CURRENT MARKING PERIOD					DAYS ABSENT					
CUMULATIVE TO DATE					TIMES TARDY					

SEE REVERSE SIDE FOR KEY TO GRADES AND COMMENTS

GRADING KEY

A - SUPERIOR ACHIEVEMENT
B - ABOVE AVERAGE ACHIEVEMENT
C - AVERAGE ACHIEVEMENT
D - BELOW AVERAGE ACHIEVEMENT
MINIMUM ACCEPTABLE ACHIEVEMENT
F UNSATISFACTORY WORK
NO CREDIT ALLOWED
(CONFERENCE WITH TEACHER RECOMMENDED)
I - WORK INCOMPLETE, NO CREDIT ALLOWED
(CONFERENCE WITH TEACHER RECOMMENDED TO DETERMINE IF WORK CAN BE COMPLETED AND GRADE REMOVED)
P - PASSED
W - WITHDREW

TEACHER COMMENTS

1. PUTS FORTH BEST EFFORT
2. COMMENDABLE EFFORT, GOOD ATTITUDE. COOPERATES IN CLASS.
3. AVERAGE PREPARATION AND PARTICIPATION IN CLASS ROUTINE
4. CARELESS PREPARATION OF ASSIGNMENTS AND WRITTEN WORK.
5. POOR PREPARATION, UNCOOPERATIVE AND INDIFFERENT TO CLASS ROUTINE
6. EXCESSIVE ABSENCE OR TARDINESS
7. BEHAVIOR PROBLEMS LIMIT LEARNING PROGRESS
8. DEFICIENCY REPORT SENT
9. INTERVIEW WITH PARENTS REQUESTED

SCHOOL PHONES

WASHINGTON HIGH SCHOOL OFFICE
223-5080

GUIDANCE OFFICE
223-5079

APPENDIX F: SAMPLE REPORT CARD FOR STUDENT WORK HABITS

(GUSKEY & BAILEY, 2001)

SOCIAL DEVELOPMENT AND WORK HABITS	1st	2nd	3rd	4th
Works and plays cooperatively				
Works well independently				
Accepts responsibility for returning homework, books & school-related materials on time				
Uses time profitably				
Demonstrates self-control				
Demonstrates neatness and organizational skills				
Respects rights, opinions, and property of others				
Follows rules and displays appropriate behavior				
Follows oral directions				
Follows written directions				
Listens during instructional lessons				
Seeks help when needed				
Effectively solves social conflicts				

Key to Skills Grades:

4 = Consistently or Independently
3 = Usually
2 = Sometimes
1 = Seldom
NE = Not Evaluated

APPENDIX G: SAMPLE REPORT CARD FOR GRADES 9-12

(O'CONNOR, 2002)

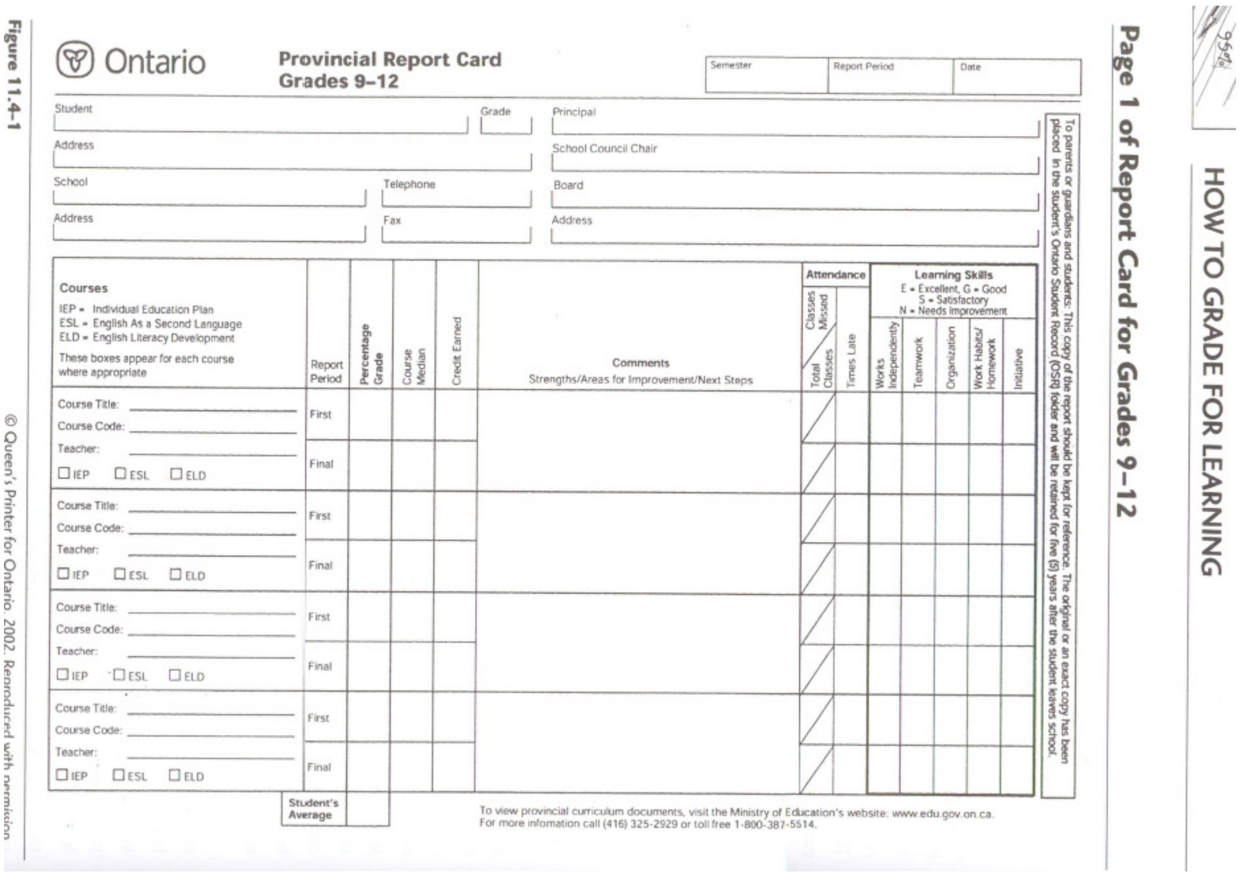

Figure 11.4-1

Page 1 of Report Card for Grades 9–12

HOW TO GRADE FOR LEARNING

95%

Ontario

Provincial Report Card
Grades 9–12

Semester
Report Period
Date

Student
Grade
Address
School
Telephone
Address
Fax

Principal
School Council Chair
Board
Address

Courses IEP = Individual Education Plan ESL = English As a Second Language ELD = English Literacy Development These boxes appear for each course where appropriate	Report Period	Percentage Grade	Course Median	Credit Earned	**Comments** Strengths/Areas for Improvement/Next Steps	**Attendance** Total Classes / Classes Missed	Times Late	**Learning Skills** E = Excellent, G = Good S = Satisfactory N = Needs Improvement Works Independently	Teamwork	Organization	Work Habits/ Homework	Initiative
Course Title: Course Code: Teacher: ☐ IEP ☐ ESL ☐ ELD	First											
	Final											
Course Title: Course Code: Teacher: ☐ IEP ☐ ESL ☐ ELD	First											
	Final											
Course Title: Course Code: Teacher: ☐ IEP ☐ ESL ☐ ELD	First											
	Final											
Course Title: Course Code: Teacher: ☐ IEP ☐ ESL ☐ ELD	First											
	Final											
	Student's Average											

To parents or guardians and students: This copy of the report should be kept for reference. The original or an exact copy has been placed in the student's Ontario Student Record (OSR) folder and will be retained for five (5) years after the student leaves school.

To view provincial curriculum documents, visit the Ministry of Education's website: www.edu.gov.on.ca.
For more infomation call (416) 325-2929 or toll free 1-800-387-5514.

Report Card Response Form

Ontario

Provincial Report Card, Grades 9–12

Report Period	Date

Response Form

Student 's Name _____ Grade _____

Address _____

School _____ Telephone _____

Principal _____ Teacher _____

Student's Comments

To be completed by the student in consultation with the teacher as part of the student's annual education plan.

1. Goals

2. Achievement

3. Action Plan

4. Community Involvement

Student's Signature _____ Date _____ Teacher's Signature _____ Date _____

Parent/Guardian's Comments

To be completed by the parent/guardian, signed, and returned to the school

Parent/Guardian's Signature _____ Date _____

(School information regarding opportunities for parent-teacher communication)

APPENDIX H: TOPICS BASED GRADE BOOK

(MARZANO, 2000)

FIGURE 4.8
A Topics-Based Grade Book

Assessment Key:			
	A. Quiz: Sept. 10 B. Homework: Sept. 10 C. Homework: Sept. 15 D. Homework: Sept. 17 E. Quiz: Sept. 20	F. Unit Test #1: Sept. 22 G. Performance Task: Sept. 24 H. Homework: Sept. 29 I. Quiz: Oct. 1 J. Homework: Oct. 6	K. Quiz: Oct. 8 L. Homework: Oct. 11 M. Homework: Oct. 13 N. Quiz: Oct. 15 O. Unit Test-Performance Task: Oct. 6

		Topics				
Students/Assessments		Precipitation	Ocean Currents	Measurement of Temperature	Reading Tables	Estimation
Bill	A	1.5		1.0		2.0
	B	2.0			1.5	
	C	1.5				2.0
	D	2.0				
	E	1.5		1.5		2.0
	F	2.0		1.5	1.5	
	G	2.5		1.5	1.5	2.0
	H		2.0			
	I		2.0			
	J			2.0	1.5	
	K		2.0		2.0	
	L		2.0			
	M		2.5			
	N		2.5			
	O	2.5	2.5	2.0	2.0	
Final Topic Score		2.25	2.5	1.5	1.75	2.0
Marv	A	3.0		4.0		3.0
	B	4.0			3.5	
	C	3.5				3.5
	D	4.0				
	E	4.0		4.0		3.5
	F	3.5		3.5	3.5	
	G	3.5		3.5	4.0	3.5
	H		4.0			
	I		4.0			
	J			3.5	4.0	
	K		3.5		4.0	
	L		3.5			
	M		4.0			
	N		3.5			
	O	3.5	4.0	4.0	4.0	
Final Topic Score		3.75	3.75	3.75	4.0	3.5

Note: Final topic scores are not necessarily averages of column scores.

APPENDIX I: TEACHER GRADING PRACTICE SURVEY

zoomerang

Teacher Grading Practices (1)

Created: November 05 2010, 9:18 AM
Last Modified: November 05 2010, 10:49 AM
Design Theme: Basic Blue
Language: English
Button Options: Labels
Disable Browser "Back" Button: False

_____

Teacher Grading Practices

Page 1 - Question 1 - Choice - One Answer (Drop Down) [Mandatory]

Use the drop down arrow to select the high school where you are currently assigned :

- ❍ Garner
- ❍ Millbrook

Page 1 - Heading

Sections 2-6 contain statements regarding factors one might consider/utilize when calculating a grade. Please respond to each statement indicating the items you consider/utilize when calculating a grade. There is no right or wrong response. Please be candid with your responses. They are anonymous.

Page 1 - Question 2 - Rating Scale – Matrix [Mandatory]

As a teacher I,

	Always (>95% of the time)	Almost Always (80-95% of the time)	Varies (40-79% of the time)	Some (10-39% of the time)	Little to None (<10% of the time)
give zeros for late work.	☐	☐	☐	☐	☐
give partial credit for work that is turned in late.	☐	☐	☐	☐	☐
accept late work for full credit.	☐	☐	☐	☐	☐
give a zero for cheating.	☐	☐	☐	☐	☐
allow a student to retake a test for full credit when caught cheating.	☐	☐	☐	☐	☐
allow a student to retake a test for partial credit when caught cheating.	☐	☐	☐	☐	☐

Page 1 - Question 3 - Rating Scale - Matrix [Mandatory]

As a teacher I,

	Always (>95% of the time)	Almost Always (80-95% of the time)	Varies (40-79% of the time)	Some (10-39% of the time)	Little to None (<10% of the time)
offer extra credit for academic tasks.	☐	☐	☐	☐	☐
offer extra credit for nonacademic tasks.	☐	☐	☐	☐	☐

Page 1 - Question 4 - Rating Scale - Matrix [Mandatory]

As a teacher when assigning a grade to a group task I,

	Always (>95% of the time)	Almost Always (80-95% of the time)	Varies (40-79% of the time)	Some (10-39% of the time)	Little to None (<10% of the time)
give the same grade to all members of the group.	☐	☐	☐	☐	☐
give each member an individual grade as well as a group grade.	☐	☐	☐	☐	☐
only give individual grades.	☐	☐	☐	☐	☐

Page 1 - Question 5 - Rating Scale - Matrix [Mandatory]

As a teacher I,

	Always (>95% of the time)	Almost Always (80-95% of the time)	Varies (40-79% of the time)	Some (10-39% of the time)	Little to None (<10% of the time)
lower grades for disruptive behavior.	☐	☐	☐	☐	☐
give a class participation grade.	☐	☐	☐	☐	☐
consider neatness when grading assignments.	☐	☐	☐	☐	☐
consider a student's academic growth when assigning a grade.	☐	☐	☐	☐	☐
consider a student's academic ability when assigning a grade.	☐	☐	☐	☐	☐
consider a student's effort when assigning a grade.	☐	☐	☐	☐	☐
use professional judgment to determine grades.	☐	☐	☐	☐	☐

Page 1 - Question 6 - Rating Scale - Matrix [Mandatory]

As a teacher I,

	Always (>95% of the time)	Almost Always (80-95% of the time)	Varies (40-79% of the time)	Some (10-39% of the time)	Little to None (<10% of the time)
do not record grades of less than 60%.	☐	☐	☐	☐	☐
include the scores from common formative assessments in my grading scheme.	☐	☐	☐	☐	☐
provide opportunities for students to retake assessments to improve their grades.	☐	☐	☐	☐	☐
grade on a curve.	☐	☐	☐	☐	☐
weigh graded work completed at the beginning of the term equally with graded work completed at the end of the term.	☐	☐	☐	☐	☐

Page 1 - Heading

The following section contains statements regarding collaboration and communication with regards to your grading practices. Please choose the response that most accurately reflects your practices.

Page 1 - Question 7 - Rating Scale - Matrix [Mandatory]

Collaboration and Communication

	Always (>95%)	Almost Always (80-95%)	Varies (40-79%)	Some (10-39%)	Little or None (<10%)
My PLT has a unified grading system.	☐	☐	☐	☐	☐
I work with my PLT to create common summative assessments.	☐	☐	☐	☐	☐
I use the same grading procedures as others in my PLT on graded work.	☐	☐	☐	☐	☐
I communicate grades to students in a timely manner.	☐	☐	☐	☐	☐
I inform students when there is little/no chance of passing my class before the end of the term.	☐	☐	☐	☐	☐
I communicate grades to parents in a timely manner.	☐	☐	☐	☐	☐

Thank You Page

Thank you for taking the time to complete this survey!

Screen Out Page

(Standard - Zoomerang branding)

Over Quota Page

(Standard - Zoomerang branding)

Survey Closed Page

(Standard - Zoomerang branding)

APPENDIX J: INSTITUTIONAL REVIEW BOARD APPROVAL LETTER

EAST CAROLINA UNIVERSITY
University & Medical Center Institutional Review Board Office
IT-09 Brody Medical Sciences Building• 600 Moye Boulevard • Greenville, NC 27834
Office 252-744-2914 • Fax 252-744-2284 • www.ecu.edu/irb

TO: Ruth Steidinger, MA, 1516 Kilarney Drive, Cary, NC, 27511

FROM: UMCIRB

DATE: February 2, 2011

RE: Expedited Category Research Study

TITLE: "A Study to Determine the Concurrent Validity between Course Grades and Standardized Assessment Results"

UMCIRB #11-036

This research study has undergone review and approval using expedited review on 1/29/11. This research study is eligible for review under an expedited category number 5 & 7. The Chairperson (or designee) deemed this unfunded study **no more than minimal risk** requiring a continuing review **in 12 months.** Changes to this approved research may not be initiated without UMCIRB review except when necessary to eliminate an apparent immediate hazard to the participant. All unanticipated problems involving risks to participants and others must be promptly reported to the UMCIRB. The investigator must submit a continuing review/closure application to the UMCIRB prior to the date of study expiration. The investigator must adhere to all reporting requirements for this study.

The above referenced research study has been given approval for the period of **1/29/11** to **1/28/12.** The approval includes the following items:

- Internal Processing Form (dated 12/12/10, received 1/17/11)
- Informed Consent (revised, received 1/21/11)
- Zoomerang teacher grading practices
- Request for school data
- Data analysis sheet

The Chairperson (or designee) does not have a potential for conflict of interest on this study.

The UMCIRB applies 45 CFR 46, Subparts A-D, to all research reviewed by the UMCIRB regardless of the funding source. 21 CFR 50 and 21 CFR 56 are applied to all research studies under the Food and Drug Administration regulation. The UMCIRB follows applicable International Conference on Harmonisation Good Clinical Practice guidelines.