THE RELATIONSHIP BETWEEN TEACHER-ASSIGNED COURSE GRADES AND THE OHIO AIR END OF COURSE 8TH GRADE SCIENCE TEST

by

Mark Robert Potts

Liberty University

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University

2018

THE RELATIONSHIP BETWEEN TEACHER-ASSIGNED COURSE GRADES AND THE OHIO AIR END OF COURSE 8TH GRADE SCIENCE TEST

by

Mark Robert Potts

Liberty University

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University

2018

APPROVED BY:

Michelle J. Barthlow, Ed.D, Committee Chair

David A. Gorman, Ed.D, Committee Member

Gjergj Haxhiu, Ed.D, Committee Member

ABSTRACT

For over a century, teacher-assigned grades have been criticized and maligned as a subjective hodgepodge of academic and non-academic factors that varied from teacher to teacher and generally failed to communicate student academic achievement. Standardized tests focused more objectively on student achievement rather than leveling the playing field by eliminating the subjectivity it seems to have tilted it in favor of white middle-class students. The purpose of this correlational study was to better understand the relationship between standardized tests and teacher assigned grades. This study examined the Ohio End-of-Course exams for 8th grade Science and the teacher assigned final grades for the same sample of students using scores and grades from the 2014-15 school year. The participants in the study were drawn from an accessible sample consisting of 2077 eighth grade students from Northeast Ohio. A bivariate correlation tested the overall strength of the relationship, while Pearson product-moment correlation coefficient analyses determined if the strength of the correlation for students with disabilities and economically disadvantaged students varied. The correlation coefficients for all for groups were positive correlations. The strongest correlations were for the students without disabilities (.537) and the all students group (.536), while students from a low socio-economic background correlated medium strength (.463) and the weakest group was the Students with Disabilities (.248). These results generally fall in line with the literature and call for further research studies in how to better engage and educate students with disabilities and students from low socio-economic backgrounds.

Keyword: high stakes testing, teacher assigned grades, standardized tests, validity, reliability.

Dedication

This dissertation is dedicated to my parents, Robert J. Potts and Linda B. Potts. My father passed away unexpectedly on January 1, 2015. I wish he could have been around to counsel me and further share in my doctoral, career, and life journey. I hope that I make him proud. My mother has always been supportive and loving, for this I am ever grateful.

I was blessed to grow up in a Christian family and had parents who were teachers and passed on the importance of education, took me to church on Sunday, as well as instilled values about life, people, and God that continue to guide me today. I am thankful every day for growing up surrounded by family, friends, and in the supportive Ashtabula, Ohio community of the 1970s and 1980s that I did.

I would also like to dedicate this effort to my sons Nolan Potts and Reagan Potts. The love I have for them was unimaginable before becoming a father. Although they have moved away and are forging ahead with their own education and career journeys, there is not a day that I do not think of them, pray for them, and miss them.

Acknowledgements

There are so many people that have shaped and influenced my life in getting me to this point thus far, I could not do justice to everyone, so even though they may not be mentioned here I know the value they have added by investing in my life. From my teachers, friends, parents, sisters, aunts and uncles, cousins, church family, coaches, and colleagues – thank you for touching my life and for what I have learned from you.

During this dissertation process, an answer to prayer and an undeserved gift from God came into my life in the form of a wonderful woman named Jessica. I only hope that I will be a portion of the blessing to your life that you are to mine. I am so blessed to get to walk beside you for the remainder of my days on earth.

I would like to thank my dissertation chair, Dr. Michelle Barthlow, for her patience, encouragement, and positivity throughout this process. I felt like quitting many times, but you are one of the people I did not want to disappoint. Thank you so much for believing in me. I will always be grateful God put you in my path.

Additionally, I want to acknowledge my committee members, Dr. David Gorman and Dr. Gjergj Haxhiu for their willingness to serve on the committee and to offer honest feedback to help me along this process. Also thank you to Dr. Melissa Watson and Dr. Donna Jones for their feedback and input.

Lastly, I would like to thank God for His grace and mercy. I have been brought through serious health issues and blessed in more ways than I could ever start to explain or could ever deserve. It is my prayer that my life, in spite of so many faults and mistakes, can be remembered as showing the love of Jesus through my actions and words as I move forward.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

American College Test (ACT)

American Institute of Research (AIR)

Americans with Disabilities Act (ADA)

Individuals with Disabilities Education Act (IDEA)

Individualized Education Program (IEP)

Institutional Review Board (IRB)

National Assessment of Educational Progress (NAEP)

No Child Left Behind Act (NCLB)

National Defense Education Act of 1958 (NDEA)

Scholastic Aptitude Test (SAT)

Socio-economic status (SES)

Students with Disabilities (SWD)

Students without Disabilities (SWoD)

CHAPTER ONE: INTRODUCTION

Overview

Chapter one includes the background history of teacher-assigned grades and standardized testing along with the major issues associated with each. The chapter also contains descriptions of the problem statement, purpose statement, and significance of the study as well as the research questions and definitions. The final part of the chapter includes the assumptions and limitations of the study are discussed.

Background

The two most common measures of student achievement are the teacher-assigned grade and standardized testing. Teacher-assigned grades are familiar to most as the mark seen on the report card that traditionally varies from A to F. These grades frequently include non-academic measures such as effort and participation in their calculations. It is because of this reason that teacher-assigned grades have long been criticized for their subjectivity and lack of validity and reliability (Brookhart et al., 2016; O'Connor, 2017).

Gall, Gall, and Borg (2007) stated that standardized tests are "a test that has procedures to ensure consistency in administration and scoring across all testing situations" (p. 195). Standardized tests are tests such as the ACT, SAT, and many other assessments meant to measure the achievement objectively for every student being tested and as such, were intended to increase the equity for all students. Students have been inundated with these assessments as part of the accountability movement. From preschool placement tests to graduation, students are taking high-stakes standardized tests. Ideally, there would be a high correlation between standardized tests and teacher-assigned grades as they are the two most common measures (O'Connor, 2017; Reeves, 2016). Historically, however, this is not the case, especially for

certain populations of students, in particular students with disabilities and students from families with a low socioeconomic status (Kling, Noftle, & Robins, 2012; Toldson & McGee, 2014, Turnbull, 2016).

Since their inception, teacher-assigned grades have historically been denounced for their subjectivity, and have been criticized as being unreliable and invalid (Bowers, 2009; Brookhart et al., 2016; Marzano & Heflebower, 2011). Traditionally, teacher-assigned grades include factors such as effort, participation, attendance, and compliance (Brookhart, 2017; Brookhart et al., 2016; Cornue, 2018; Marzano & Heflebower, 2011; O'Connor, 2017). These non-academic factors tend to obscure the measure of academic achievement, and because each teacher may grade differently under this system, the grades for the same student and same level of academic achievement could vary greatly from teacher to teacher (Reeves, 2016).

At the center of the teacher-assigned grade debate is the message of what the grade should convey. Depending on the individual teacher assigning the grade, the student could be being compared to the rest of the class, mastery of the standard being assessed, or the extent to which they improve during the course. The criterion from which the teacher is assessing achievement can make a significant difference in the grade a teacher assigns. If other elements of behavior are factored in, the individual teacher may be deciding how heavily they are weighed in the final grade (Cornue, 2018; O'Connor, 2017; Reeves, 2016). Without grades being standardized, there is neither validity to their accuracy nor is there any reliability as to their consistency (Brookhart et al., 2016; Marzano & Heflebower, 2011; O'Connor, 2017).

Standardized tests are considered the best measure of academic achievement and are intended to promote an objective measure of knowledge without the subjective factors (Higgins, 2009). Even so, there are well-founded allegations of gender (Duckworth, Tsukayama, & Quinn,

2012; Duckworth & Seligman, 2006; Kling et al., 2012;), racial (Hiss & Franks, 2014; Houser & An, 2014; Thompson & Allen, 2012), socioeconomic (Higdem, Kostal, Kuncel, Sackett, Shen, Beatty, & Kiger, 2016; Houser & An, 2014; Sackett, Kuncel, Beatty, Rigdon, Shen, & Kiger, 2012; Zwick & Himelfarb, 2011), and cultural bias (Garcia, 2011; Rodriguez & Arellano, 2016; Toldson & McGee, 2014), as well as concern over whether or not the tests are fair for students with disabilities (Lane & Leventhal, 2015; Tanis, 2014; Wormeli, 2006; Yell, Katsiyannis, Collins, & Losinski, 2012).

In the current social climate of high stakes accountability, students are inundated with assessments. No Child Left Behind (NCLB), Race to the Top, the Common Core initiative, and new standards with new waves of testing are leaving many teachers frustrated and students overwhelmed (King & Rohmer-Hirt, 2011; Nichols, Glass, & Berliner, 2012; Turnbull, 2016). Instead of the accountability debate taking place among educators, much of the debate is centered in the state legislatures. Although there is a seemingly universal agreement on the need for accountability, there are varying opinions when it comes to how much is enough and how much is too much. The debate between those in favor of greater accountability (King & Rohmer-Hirt, 2011; Supovitz, 2009) and those seeking less accountability (Huddleston, 2014; Minarechova, 2012; Tanis, 2014; Turnbull, 2016) is intense and politically charged.

Amid this push for higher standards and accountability, there is concern over what the effect of testing is on students in general, but especially students with disabilities, economically disadvantaged students, and minorities (Tanis, 2014; Thompson & Allen, 2012; Turnbull, 2016; Yell et al., 2012). It is within this conceptual context of high-stakes testing, the quest for higher standards, and accountability in public schools that the researcher grounded this study. States are pouring billions of dollars into end-of-course tests and want results (King & Rohmer-Hirt, 2011;

Nichols et al., 2012; Supovitz, 2009). In turn, these states need to show a return on investment to their constituents–the tax paying (and voting) public. Additionally, the availability of powerful computer technologies has made it possible to garner vast amounts of data to analyze and assess school effectiveness in ways that would not have seemed possible a generation ago (King & Rohmer-Hirt, 2011). The push for Common Core State Standards, or close derivations thereof, are changing the landscape of public education. The assessments attempting to measure student mastery of these standards should correlate with classroom assignments, and thus final teacher-assigned grades (Marzano & Heflebower, 2011; O'Connor, 2011; Reeves, 2016), but this is often not the case.

Problem Statement

Standardized end-of-course assessments provide valuable data to school systems. The data can be used to drive systemic change, guide where resources will be allocated, and evaluate teachers and administrators, as well as in rating or ranking the school. Teacher assigned grades are a more salient, if not accurate, portrayal of how a student is performing in school. Being an "A" student implies mastery of one's studies in a way far more universally understood than scoring a 738 on the American Institute of Research (AIR) Ohio End-of-Course Science assessment. However, as the literature shows, grades are widely criticized by many esteemed educational experts as not accurately conveying student knowledge of the standards (Guskey, 2015; Marzano, 2000; Marzano & Heflebower, 2011; O'Connor, 2017; Swan, Guskey, & Jung, 2014).

Although similar studies have been done on high school students, Ellison (2008) called for lower grade levels to be studied. Studies completed in other states such as Missouri (Pratt, 2013), Georgia (Philipp, 2014), and California (Steffan, 2010) called for further study with other state standardized assessments in order to compare the results.

The spring of 2014 also marked the first time students in Ohio took this new end-of-course assessment so there have not yet been any studies done on the AIR test in Ohio that correlates teacher-assigned grades and the AIR Science assessment for eighth grade students. Ohio's State Tests, n.d.). The problem is that although teacher-assigned grades and standardized tests are both widely used to measure student achievement, they do not always have a strong correlation (Reeves, 2016). This study will examine the correlation of the two measurements. It will also look at groups identified as having underperformed in other studies to see if these assessments in Ohio statistically correlate with teacher-assigned grades for those particular groups of students in the study.

Purpose Statement

The purpose of this study was to examine the results of Ohio End-of-Course Exams for eighth grade students in science and to measure the degree of their relationship with teacher-assigned final grades in Science for those same students. Further, this study examined the degree of that relationship with regard to students with disabilities and students from economically disadvantaged backgrounds. The study had two variables. The teacher-assigned grades were the predictor variable while the scores on the end of course Science exams were the criterion variable. Not only did this study look to establish the correlation between the grades and standardized test scores, but it also looked for achievement gaps in each of the aforementioned groups. Nationally, there are achievement gaps for each of these identified groups. The researcher identified where those gaps existed and the degree to which they existed.

All of the students in the study were 8th grade public school students in a five county area of Northeast Ohio. Data were solicited from schools in Atwater, Shore, Hilltop, Southward, and Broadmoor counties in Ohio. Including students from multiple schools increased the scope of the

study and allowed for a broad range of students and a significant sample size representing all of the groups to be studied. It included affluent suburban schools, urban schools, and smaller rural districts.

The study allows districts, schools, and educators from the included schools to exam their curriculum, timelines, and methodologies to see how to better align instruction with the standards that the assessment is testing, as well as looking for ways to improve instruction to any groups of students who may be underperforming on this end of course assessment. In the end, the researcher's findings should be a catalyst for conversation about grades and how to use them to improve teaching and learning in the educational system, and perhaps to better align grades with the results of the summative assessments.

Significance of the Study

The study is significant on several levels. Ohio changed their end of course exams in 2014 for Science and Social Studies to AIR (Ohio's State Tests, n.d.). The test is tied to the Common Core Standards, which Ohio adopted a derivation called the Ohio Learning Standards after the idea of Common Core became a political hotbed (Layton, 2014). The test results studied are the first results released of the new test which assessed the new standards.

Teachers should now be teaching the new Ohio Learning Standards after transitioning from the old curriculum as the end-of-course tests are designed to assess the new standards. Therefore, there should be a strong correlation between teacher-assigned grades and end of course AIR test scores (O'Connor, 2017). Also it was important to identify to what degree the new assessments correlated with the grades for students with disabilities, and students who are economically disadvantaged as other states have shown a weak relationship (Huddleston, 2014;

Thompson & Allen, 2012; Toldson & McGee, 2014; Yell et al., 2012). The researcher sought to determine if the newly designed tests helped close the gap or continued the national trends.

The study was also of importance because it shed light on current deficits in educational practice and caused re-examination of classroom teaching methods, strategies, pacing, and other factors that contribute to improved instruction. It could also be a spark for schools to more closely align classroom instruction with the new Ohio Learning Standards the teachers and students are being held accountable for. Ultimately, the results could lead administrators and school district personnel to use the findings to improve pedagogical practices and academic achievement.

Research Questions

RQ 1: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR end of course 8th grade Science test for the 2014-2015 school year?

RQ 2: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR end of course 8th grade Science test scores for the 2014-2015 school year for students with disabilities?

RQ 3: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR end of course 8th grade Science test scores for the 2014-2015 school year for students receiving free or reduced lunches (economically disadvantaged)?

Definitions

1. *High stakes testing* – Nichols et al. (2012) define high stakes testing as "the process of attaching significant consequences to standardized test performance with the goal of incentivizing teacher effectiveness and student achievement" (p. 3).
2. *Reliability* – according to Gall, Gall, & Borg (2007) reliability "refers to the consistency, stability, and precision of test scores" (p. 150).
3. *Standardized tests* – Gall et al. (2007) defines standardized tests as "a test that has procedures to ensure consistency in administration and scoring across all testing situations" (p. 195).
4. *Teacher-assigned grades* - O'Connor (2011) defines a grade as "the symbol (number or letter) reported at the end of a period of time as a summary statement of student performance" (p. 7).
5. *Validity* – defined by Gall et al. (2007) as "refers to the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores" (p. 150).
6. *Ohio's Learning Standards* – an educational framework by the State of Ohio to identify the skills and knowledge a student needs at each grade level ("Learning in Ohio", n.d.).

CHAPTER TWO: REVIEW OF THE LITERATURE

Overview

Chapter two begins with an introduction to the literature review of the study and discusses the conceptual framework on which the study is developed. The empirical evidence section looks at issues with teacher-assigned grades, starting with the history and types of grading systems, and further delves into the purpose of grades, validity, reliability, grade inflation, and gaming the system. Standardized tests are compared with high-stakes testing and how students are impacted in the areas of narrowing the curriculum, pressure, anxiety, and stress issues, and the impact on minority populations. The related literature section looks at the impact on students with disabilities and students who are economically disadvantaged. Issues such as accommodations, test format, effects of poverty, retention rates, and gaps in achievement are discussed prior to the chapter conclusion.

Introduction

Grading practices in American public schools have changed very little since their inception over a century ago (Brookhart et al., 2016, Guskey, 2015; O'Connor, 2017). Teacher-assigned grades in particular have always been subject to the focus of criticism and debate due to their subjective nature and a lack of clarity as to what grades should represent (O'Connor, 2017; Reeves, 2016; Swan et al., 2014). Grade inflation, as it is now, was an issue at Harvard University as early as 1894 as concern was raised over the ease students were receiving high marks (Laurie, 2009).

Standardized tests have also come under fire as favoring white middle-class students and being unfair to others and have been linked to racial (Borich & Peck, n.d; Hiss & Franks, 2014; Huddleston, 2014; Thompson & Allen, 2012; Toldson & McGee, 2014) and gender bias

(Duckworth et al., 2012; Duckworth & Seligman, 2006; Kling et al., 2012), as well as placing socioeconomically disadvantaged students at a further deficit (Higdem et al., 2016; Houser & An, 2014; Sackett et al., 2012; Zwick & Grief-Green, 2007; Zwick & Himelfarb, 2011). Standardized tests are designed to be objective and to eliminate the subjectivity that maligns the perception of teacher-assigned grades. Instead of leveling the playing field, however, there is significant evidence that they may further tilt it to the advantage of the majority (Duckworth et al., 2012; Hiss & Franks, 2014; Huddleston, 2014; King et al., 2012; Sackett et al., 2012; Thompson & Allen, 2012; Toldson & McGee, 2014; Turnbull, 2016; Zwick & Himelfarb, 2011). Regardless of the inherent faults with both, the two most significant measures of student achievement in high school, as predictors for future academic success, are teacher-assigned grades and standardized test scores (Ricketts, 2010; Zwick, 2013).

Conceptual Framework

Researchers have identified many theoretical frameworks in discussing standardized tests and teacher-assigned grades but there does not seem to be any overwhelming consensus as to a theoretical framework. Historically speaking, functionalist theory conjectures that grades and standardized tests originated to separate, sort, and classify students for the purpose of employment (Gallagher, 2003). Horace Mann began standardized testing as a way to monitor schools in Boston from building to building and teacher to teacher to objectively compare instructional quality in the 19th century (Gallagher, 2003). Functionalism as a theory was first proposed by sociologist Emile Durkheim (Burrowes, 1993). At its roots, functionalism supports the idea that the purpose of education is to pass values and shared knowledge from one generation to the next (Edwards, 2006). Functionalism being a meritocracy, used grades and test scores to sort the future blue collar worker from the future white collar worker, the foot soldier

from the officer (Edwards, 2006). Although functionalism was clearly the catalyst for the advent of the traditional grading system and standardized assessments, it is not the reason behind the high-stakes assessments that students and educators alike are inundated with in our current school political climate (Edwards, 2006).

At the theoretical roots of the high-stakes testing accountability movement, schools are currently immersed in is the psychological theory of behaviorism (Supovitz, 2009). Knight (2006) points out the premise of behaviorism is that "people are programmed to act in certain ways by their environment. They are rewarded for acting in some ways and are punished for acting in other ways" (p. 136). Supovitz (2009) believes the predominant underlying theory behind high-stakes testing as an accountability measure is the behaviorist based motivational theory. Supovitz states that motivational theory "holds that the extrinsic rewards and sanctions associated with the high stakes test can be used to motivate school faculty members to improve performance" (p. 214). An obvious problem with this theory is it is a deficit model assuming teachers and students are not intrinsically motivated, but rather need the extrinsic threat of sanctions to motivate them (Supovitz, 2009).

High-stakes testing does not just assess student achievement, but also has an effect on teaching and learning due to the nature of the consequences attached to failure (Pan, 2014). Pan (2014) discusses the effect of washback variability which he defines as "influence tests exert on teaching and learning" (p. 2). In essence, the Pan study looked at the degree to which high-stakes testing influences the motivation, strategies, and performance of learners as well as the manner in which the teacher chooses to cover the material.

Assessments based on the Ohio Learning Standards, which are strongly steeped in the Common Core Standards, have evolved into a more constructivist assessment where students

must use critical thinking to construct responses through Problem-Based Learning where they use higher level thinking skills such as analysis, synthesis, and evaluation (Bostic, Vostal, & Ruffer, 2014; Sutinen, 2008). Yilmaz (2008) states that "constructivism postulates that knowledge cannot exist outside our minds; trust is not absolute, and knowledge is not discovered but constructed by individuals based on experiences" (p. 62). Basically, process is more important than product. One's critical thinking skills are of greater value in a constructivist approach than regurgitating a memorized response (Yilmaz, 2008).

The current school climate was not necessarily designed by best practices, by professional educators, or by educational researchers, but largely by politicians' impetuous reactions to their constituents' frustrations with a system admittedly in need of reform (Groen, 2012). Although politicians acted with expedience, they did so with little foresight as to the effects on children. Researchers have shown these decisions have had a profound effect on education in the United States, but certainly much of it has not been positive (Groen, 2012; Huddleston, 2014; Minarechova, 2012). Nichols and Berliner (2008) report that students are "exposed to an unprecedented number of tests" (p. 14). Lower graduation rates (Reardon, Arshan, Atteberry, & Kurlaender, 2008; Jakee & Keller, 2017), mandated retentions (Huddleston, 2014; Jimerson & Renshaw, 2012; Nagel, 2015), reduced educational opportunities (Jakee & Keller, 2017; Zellmer, Frontier & Pheifer, 2006), and an achievement gap that shows no signs of closing are a distinct part of the accountability movement's legacy (Erskine, 2014; Hiss & Franks, 2014; Thompson & Allen, 2012; Toldson & McGee, 2014). One could argue how the No Child Left Behind Act (NCLB) has further isolated those it was designed to help the most and have inadvertently stacked the deck even further in favor of white middle class males (Erskine, 2014; Ricketts, 2010).

Erskine's (2014) research supports the premise behind conflict theory. Conflict theory purports those in power game the system to maintain a firm grip on the power (Edwards, 2006), whereas functionalist theory sees the playing field as level, conflict theory sees a system tilting the field to favor the ruling class (Edwards, 2006). Based on standardized tests scores (Ricketts, 2010), questions are asked on assessments that seem to favor the middle-class (Toldson & McGee, 2014), and affect graduation rates (Jimerson & Renshaw, 2012), retention rates (Huddleston, 2014), school discipline data (Jimerson & Renshaw, 2012), and certainly funding systems in a manner that strongly favors wealthier districts (Thompson & Allen, 2012). It is difficult to dismiss aspects of this theory devised by Karl Marx. American schools have done such a despicable job in closing the achievement gaps for segments of our population a Marxist theory actually is able to gain some traction (Erskine, 2014). However, Groen (2012) asserts our current penchant for testing is not part of a government scheme to keep the powerful in power, but rather it is the result of a strong school accountability political movement (Groen, 2012).

It is within this conceptual framework that American public schools have arrived in their current high-stakes climate. It is neither to innocently pass on values and sort students into areas that they may be most successful, nor is it a sinister plot to keep the powerful in power, but rather a political response to a public clamoring for school accountability that was manifested in NCLB (Groen, 2012). Groen (2012) points out that "the accountability movement may have begun as an educational ideology, but as advocates sought to operationalize their agenda, it soon transformed into a political movement driven by administrative concerns rather than educational considerations" (pp. 6-7).

Duckworth, Tsukayama, and Quinn (2012) identified a valid concern when comparing standardized assessments and teacher-assigned grades, positing they are measuring two different

things. Duckworth et al. (2012) shows how teacher-assigned grades measure compliance and factors of behavior which may predict school success whereas standardized tests measure intelligence. In fact, Duckworth and Seligman (2017) found that "self-control predicted rank order gains in report card grades, IQ did not" (p. 715). Although both of these factors may be important in predicting student success, there may be a lack of correlation between teacher-assigned grades and standardized tests due to the fact they may be measuring distinctly different phenomena which is akin to comparing apples to oranges (Duckworth et al., 2012).

Related Literature

Teacher Assigned Grades

Since their beginning, teacher-assigned grades have been maligned as subjective and unfair (Brookhart et al., 2016; O'Connor, 2017; Swan et al., 2014). Researchers in the area of educational measurement are vocal about the unreliable measurement and lack of validity of grades (Brookhart et al., 2016; Swan et al., 2014). Because there is so much variance in what an individual teacher uses to determine a grade, a teacher across the hall would have difficulty replicating it even with the same data (Brookhart et al., 2016; Froman et al., 2010; O'Connor, 2011). Grades can be a sum of total points or weighted by the importance of the assignment. Homework may be a very large factor, or it may just be viewed as practice. Some teachers may give points for participation, give the benefit of the doubt for effort, or may give special treatment to a student for reasons other than academic achievement. The student may be compared against himself (self-referencing), compared against his peers in the class (norm referencing), or compared to standards (criterion referencing) in order to arrive at a grade (Brookhart, 2009). Schneider and Hutt (2014) point out how historically grades are a reflection of "interplay between the work of the classroom and society at large" (p. 202), but even in

schools where there is a grading policy in place, the subjective nature of a teacher-assigned grade and what goes into its final determination is somewhat less than scientific (Brookhart et al., 2016; Swan et al., 2014).

Teacher-assigned grades may take on one or more of several different systems or schemes. Some common grading systems use percentages, letter grading, norm-referenced, mastery, pass/fail, standards based, narrative grading, and contract grading or combinations of these various schemes (Brookhart et al., 2016) All of the systems have benefits and shortcomings. Unfortunately, the most common methods of grading are often those seeming to be the most fraught with problems (Brookhart et al., 2016).

Among the most frequently used grading schemes is the percentage grading system. Percentage grading involves assigning grades based on the percentage of correct answers or the percentage of total points earned during a grading period and was the most used method of reporting grades in both high schools and colleges from about 1890-1910 (Brookhart et al., 2016). There seems to have been a resurgence in the popularity of percentage grades with the advent of grading software and what Guskey (2013) refers to as "the partialities of computer technicians, not from the desire of educators for alternative grading scales or from research about better grading practice" (p. 69). Guskey (2015) says that "percentage grades give the illusion of precision to imprecise and often highly subjective judgments of student performance" (p. 23) and surmises that percentage grades are a deterrent to finding better grading solutions.

Percentages are assumed more accurate as they produce an exact score, but there are many factors included in the percentage grade (Reeves, Jung, & O'Connor, 2017). In the early 20th Century, 50% was considered average with very few outliers going lower than 25% or higher than 75%. Currently, percentage grades are based on 75%, or a C, being average and 60%

or 65% or below as failing (Guskey, 2013). Percentage grades in this scenario negatively skew the distribution (Guskey, 2013). Carey and Carifio (2012) concur that such a distribution is disadvantageous to the students. Using this common grading scale, students have a possible 60-65 levels of failure and only 35-40 levels of relative success which, Guskey (2013) says allows a "greater influence of subjectivity, more error, and diminished reliability" (p. 70).

Another prominent argument against percentage grades is the effect of outliers on the overall percentage average. O'Connor (2011) found grades to be faulty when the mean is used because outliers or even one extreme score can significantly alter the percentage. For example one zero would take nine perfect scores to recover from in a percentage grading system (Guskey, 2015). O'Connor (2011) points out "zeros give a numerical value to something that has never been assessed, and therefore has no basis in reality" (p. 96). Zeros are detrimental in motivating students and they represent inappropriate and faulty math (O'Connor, 2017, Reeves, 2016; Reeves et al., 2017).

Guskey (2013) gives an example of the extreme consequence of a zero grade in stating that moving from a B to an A means moving 10 percentage points or less when in fact moving from a zero to a D would necessitate a much greater effort to recover and states "a single zero can doom a student to failure, regardless of what dedicated effort or level of performance might follow" (p. 71). This perceived accuracy of percentage grades, according to Guskey (2015), is more illusion than reality.

Letter grades became popular in the 1920s more than 100 years after the practice was introduced at Yale (Brookhart et al., 2016). Letter grades use a series of letters to symbolize student achievement. Grades are usually A, B, C, D, and F. Traditionally A means excellent, B means good, C represents average, D is poor, and F indicates failure. Some letter grading

systems add plus or minus to the grade in an attempt to show more precision (Reeves, 2016). A - F grading was emerging as the preeminent grading system by the 1940s along with the 4.0 scale and the percentage grade and although not standardized at the time, all three schemes would eventually become fused together (Schneider & Hutt, 2014). Brookhart (2009) points out the benefit of the letter-grade system is that letter grades seem to have a "shared meaning" (p. 26) as everyone has a concept of what a "straight A" student is or of getting an F in Algebra, but one cannot tell by looking at a letter grade whether one is comparing student performance to a standard or to the student's classmates. O'Connor (2011) gives the illustration of a brilliant writer who always turns in work late and thus receives a teacher-assigned grade of a D, which neither communicates the student's brilliance nor that the student is chronically late turning in work.

Letter grades today are usually merged with a percentage system (Schneider & Hutt, 2014) where 90-100% equates to an A, 80-89% a B, 70-79% equaling a C, 60-69% a D, and below 60% is failing or an F. Many schools, in order to show more rigor, have amped up grading scales where 94% is required for an A 93-86% is a B with below 70% being an F. Letter grades are ingrained in the American psyche but often come up short in communicating a student's strengths and shortcomings (Reeves et al., 2017).

Norm-referenced grades base the student grade on how he or she stands in relation to the rest of the class. Brookhart (2009) notes how norm-referenced grades were used to compare students with one another and was a popular format in the early 1900s. Although norm-referencing works well for ranking purposes, it fell out of favor as a grading system in schools by the 1930s (Brookhart et al., 2016). Stiggins (2014) acknowledges that historically students were ranked by their achievement to separate them by ability and to highlight the differences in

student learning, but that in doing so we have created a culture that has distinct winners and losers. Norm-referencing was used primarily to spark competition among students to motivate them to perform at their highest level (Brookhart, 2009).

The practice of ranking students based on their grades came out of the European tradition and was popularized in the United States largely by Horace Mann was a reformer in the mid-1800s who was particularly interested in the Prussian school model and was influenced by the Prussian organizational innovations (Schneider & Hutt, 2014). Norm-referencing is used more today to determine class rank or as a cut-off score than as a grading scheme to determine a student's progress (O'Connor, 2017).

Mastery grading is defined as a system that permits a student to work on a specified learning objective until they have attained a certain level of assessed competence (Brookhart, 2009). Although mastery grading has been around since at least the 1920s (Brookhart, 2009), it is heavily associated with the work of Benjamin Bloom, creator of Bloom's Taxonomy, from the 1970s and 1980s (Beck & Santoyo, 2017; Guskey, 2010). Mastery learning requires students to receive high quality instruction, then be assessed using a formative assessment to determine exactly what the students have grasped and where additional learning needs to take place (Beck & Santoyo, 2017). According to Guskey (2010) in mastery learning "assessments are not a one-shot, do-or-die experience; instead they are part of an ongoing effort to help students learn" (p. 56). Students take on learning targeted at their needs, referred to as correctives by Bloom and after completion take a second similar formative assessment focusing on the same goals, but with different questions or problems posed (Beck & Santoyo, 2017; Guskey, 2010).

The correctives require high quality instruction, unlike what Guskey (2010) states "often consists simply of restating the original explanations louder and more slowly" (p. 55). This

second assessment reveals the success of the correctives and shows what content still needs to be learned (Beck & Santoyo, 2017). Theoretically, students work through this process until they have mastered the material and the learning goals and move on only after they have attained a mastery level of achievement (Beck & Santoyo, 2017). Students who master the material sooner are given extension or enrichment activities allowing greater depth of learning with respect to the learning goals (Beck & Santoyo, 2017).

Guskey (2010) endorses mastery learning as "one of the most powerful research – supported strategies" (p. 53) that has been studied in the past 40 years. To clarify, mastery learning is the method of teaching advocated by Bloom, while mastery grading is how the teacher would record these grades. In mastery grading, the highest level of achievement would be the grade the student has earned. Much like the SAT or a driver's license exam, once the student has reached the level of mastery, the mastery grade would be the only grade to count (Brookhart, 2009; Guskey 2010). Practice or the number of times taken to reach mastery would not be scored against them. Beck and Santoyo (2017) point out that getting extra chances to get it right is a powerful motivator to many students and helps instill confidence.

However, a mastery grading system is labor intensive for the teacher as students are working individually with corrective exercises, multiple assessments, and extension work along with the high quality instruction which needs to take place. For many teachers in large classrooms, putting a mastery system into practice may seem overwhelming and simply too daunting. Therefore, despite its well-documented success, mastery grading is not the primary grading method in most schools (Guskey, 2010).

According to Brookhart (2009), pass/fail grading was researched at the University of Michigan as early as 1851. This method is still used on a limited basis in many high schools and

colleges, often for non-credit classes, such as freshman orientation. Although similar to mastery grading, the student either reaches the standard or does not. research has revealed that pass/fail courses encourage minimal effort and a just-enough-to-pass attitude (Brookhart, 2009).

Standards grading has had multiple meanings although they are somewhat related. In the early 20th century, standards grading and criterion referencing were used interchangeably. The idea was that grading was based on the students' knowledge of a specific standard (Brookhart et al., 2016). In the past few decades, standards-based grading is more reflective of the state standards that are required. Standards-based grading or standards grading refers to "communicating student progress in relation to grade-level standards" (Brookhart et al., 2016, p. 828). A broader definition of standards-based grading, of which vast amounts have been written, would describe grading that is devoid of non-academic factors and is a pure representation of student academic achievement. The idea is that all standards based grades come purely from students assessments (Guskey, 2015). In other words, the grade is summative based on neither on homework, behavior, attendance nor any other factors, but solely on what the student has learned as proven by an assessment.

There is a very strong movement towards standards-based grades by some of the biggest names in the education assessment field. Marzano (2000), Reeves (2016), Guskey (2015), and O'Connor (2017) endorse improving the validity and reliability of traditional grades in order to cause grades to communicate student achievement of given standards as purely and as replicable as possible. Future references to standards-based grading will be describing this definition.

Narrative grading is simply written feedback about a student's work (Finefter-Rosenbluh & Levinson, 2015). It involves written commentary as feedback and may or may not include the use of letter or numerical grades (Brookhart, 2009). Although this practice may be the most

successful in communicating student progress, it is rarely implemented in schools largely because of the demand of effectively writing a detailed essay about the progress of each child in the classroom and for each subject area. Although documented to be effective, in schools where teachers instruct 150 or more students daily, Finefter-Rosenbluh and Levinson (2015) cite that lack of time presents a major obstacle in using a narrative grading scheme that is so labor intensive and time consuming.

The importance of this detailed feedback should not be underestimated. Hattie (2009) conducted a meta-analysis of more than 800 meta-analyses and evaluated the impact of many factors on student achievement. Hattie found that some practices had a greater impact on student achievement than many of the societal factors thought by many to be insurmountable. These practices included teacher-student relationships, teacher clarity, and feedback. In response to the Hattie study, Reeves (2016) called feedback the strongest of all tools for educators, and exclaims that feedback not only outweighs "most other instructional interventions, it is also more important than socioeconomic status, drug use, nutrition, exercise, anxiety, family structure, and a host of other factors that many people have claimed are overwhelming" (p. 13).

The narrative format does a much better job of summing up what a student has accomplished over a grading period as opposed to just a letter grade and is also formative in that it speaks as to what the student needs to continue to improve on (Finefter-Rosenbluh & Levinson, 2015). However, it lacks practicality in over-crowded classrooms where teachers are often charged with so many duties for so many students (Finefter-Rosenbluh & Levinson, 2015). Under the right circumstances, narrative grading has plenty of merit to be an effective practice in student assessment and communication of student progress; however, the trends at the high

school level are shrinking to almost non-existence as there may be one or two high schools still using narrative grading exclusively (Finefter-Rosenbluh & Levinson, 2015).

Contract grading involves a written contract between the student and teacher that specifies what the student must accomplish to earn an A, B, or C (Brookhart, 2009). The amount of effort and depth of the work become a matter of choice for the student (Gencel & Saracaloglu, 2018). This would be the grading system used with a layered curriculum, for example or in project based learning where students can use their own creativity and take ownership of the product or project that they produce and have a choice as to what grade they would like to earn (Gencel & Saracaloglu, 2018). Contract grading often involves rubrics where the grade the student earns gets higher with the amount of difficulty, creativity, or effort that is put into the final product (Brookhart, 2009).

In order to make sense of teacher-assigned grades, it is important to determine what the teacher is trying to convey with the mark. Froman et al. (2010) pose some of the central issues when they ask

> does the student's grade represent the level of achievement, the gain in achievement, or some combination of the two? Is the student's effort included in the grade, or are high achievers given good marks regardless of effort? Are pupils marked according to their own potential learning ability or in relation to their classmates' achievement? (p. 1)

O'Connor (2011) points out that school communities must develop a "shared vision of the primary purpose of grades" (p. 8). In other words, schools must agree upon what grades mean and what it is they are to communicate. They also have to decide what, if anything, non-academic goes into the calculation.

Many prominent educational researchers are also clearly on the side that grades should reflect student academic knowledge as purely as possible, and the ultimate goal in grading should be a teacher-assigned grade strongly correlating with the standardized test scores and thus communicating what the student has mastered academically (Guskey, 2015; Marzano, 2000; O'Connor, 2017; Reeves, 2016). Those who favor standards-based grading that compliance, behavior, attendance, and participation, and attitude should be left out of any grade calculations, but rather should simply be a clear measure of what the student is able to do academically (Cornue, 2017; Guskey, 2015; O'Connor, 2011; Reeves, 2016).

The concept of validity is one of the areas where teacher-assigned grades are most criticized. There are researchers who feel that are grades are only valid when student achievement is effectively and accurately communicated, leaving grades that are contain non-academic factors as meaningless (Cornue, 2017; Guskey, 2015, Marzano, 2000; O'Connor, 2011; Reeves, 2016). In close relation to validity is the concept of reliability. Gall, Gall, and Borg (2007) define reliability as referring "to the consistency, stability, and precision of test scores" (p. 150). Guskey and Jung (2016) describe the concept of inter-rater reliability which they define as the ability of trained, competent teachers to consistently grade student work and formulate the same summary grade based on the same evidence. They point out that if teachers can grade a piece of work and consistently rate it at the same grade, it would suggest that the grade was a reliable measure of student achievement (Guskey & Jung, 2016).

When teachers create their own criterion for grades, the reliability is compromised. In a 1912 study by Stark and Elliott on reliability, 147 English teachers were given two identical papers to grade. The first paper was graded as low as a 64 and as high as a 98 by different

teachers. The second paper grades ranged from 50 to 97 (Stark & Elliott, 1912). The same paper could have been graded as an A or an F depending on which teacher graded it.

Assuming the lack of reliability could stem from the subjective nature of what constitutes good writing, Stark and Elliott repeated the study using geometry the following year. Mathematics papers given to 128 teachers varied from 28 to 95, an even greater disparity than the English papers (Guskey, 2015). Guskey (2015) explained some of the factors that led to the large discrepancy (a range of 67 points) as "some teachers deducted points only for a wrong answer, many others gave students varying amounts of partial credit for their work. Still others considered neatness, form, and spelling in the grades they assigned" (p. 25). So even in a subject area that is thought to be more objective than writing, the grades varied at an even greater range as teachers had styles which were clearly not replicable nor reliable making them invalid as a measure of student achievement.

Brimi (2011) replicated the Stark and Elliott study nearly a century later with 90 English teachers. Even after all the teachers in the study had received a mandatory training in grading writing assignments for 16 hours by their district, and some had received an additional three hours, the teacher assigned grades on the papers ranged from 50 to 96 on a 100 point scale with a mean score of 81.1599, nearly identical to the findings in the Stark study in 1912 (Brimi, 2011). As for the letter grades assigned, 10 scored the paper as an A, 18 marked it as a B, 30 rated the paper a C, nine saw it as a D, while six gave it a grade of F. If reliability is the ability to replicate the same score, Stark and Elliott (1912) and Brimi (2011) demonstrated an extreme lack of reliability.

Stark and Elliott (1912) and Brimi (2011) showed reliability can be extremely weak even among highly trained professional teachers who were specifically trained in assessing student

writing. The results of the Brimi (2011) study may be an example of what Stiggins (2014) calls lacking assessment literacy, which Stiggins says may be attributed to teacher ignorance or indifference to research on assessment. In discussing the aforementioned research, Guskey (2013) points out that even if an assignment did have 100 specific levels of student performance, that even the best trained instructors are not likely to be able to distinguish those levels with any consistency.

According to Reeves (2016) grades must be accurate, fair, specific, and timely. Similarly, O'Connor (2013) calls for grades to be supportive of learning, accurate, meaningful, and consistent. As much as educators often claim to be striving for these essentials and breaking away from antiquated ranking systems, they often fall into using systems that compare students. A grade of C is still described as average. In referring to grades, Wormeli (2006) states "average is a norm-referenced idea; we're comparing students against each other. We're hypocrites when we do this. We claim to be about standards and what individual students learn, but fall right back into comparing students with one another" (p. 99).

Another fault of teacher-assigned grades is their penchant for inflation. Grade inflation is described by Nagel (2015) as an increase in grades without a corresponding rise in academic achievement. Likewise, Guskey (2015) describes grade inflation as an increase in student grades without a comparable increase in student achievement or performance. Wongsurawat (2009) notes that grade point averages have been steadily increasing in the United States for many years. Laurie (2009) points out that grade inflation still exists, but he also highlights it is easier to score well on a test covering a chapter than a standardized summative test on the entire course.

Grade inflation is apparent in a 1994 study by the U.S. Department of Education's Office of Educational Research and Improvement where students who received mostly A's in English

and attended a school that had a poverty rate of 75% or more, had equivalent standardized reading scores with students who received C's or D's in schools with a more affluent student body (Marzano, 2000). Godfrey (2011), in a College Board study, concluded similarly that in drawing comparisons to different schools "that despite seemingly equal grades, scores on achievement tests show great differences in the student populations (p. 15).

Perhaps part of the grade inflation issue begins when students are taught from an early age that grades are a commodity to be exchanged for good behavior, cooperation, and effort. One would be hard-pressed to find a kindergarten student who asks how many points is this worth? (Vatterott, 2015). Vatterott (2015) points out students are taught that grades are a commodity in an orchestrated barter system where points are traded for compliant behaviors–make it easy on the teacher and you will be rewarded (Vatterott, 2015). The system students are taught unfortunately puts product ahead of process (Vatterott, 2015). This barter system economy focuses the students on the grade and not what is learned. Learning becomes secondary to those points. Grades become a means of control for the teacher as opposed to a measure of learning.

Vatterott (2015) describes this emphasis on grades as having "inadvertently created an addiction to form over substance" (p. 6) and perpetuating a culture that often that values compliance over learning. This system that places the importance of grades in higher esteem than learning is rife with gaming and cheating to gain the highest grade with the least amount of effort. Savvy students game the system by dropping difficult classes, taking courses reputed to be easier and teachers who are easy A's to get the reward of the grade with as little effort as possible as opposed to taking the more challenging classes or teachers because they may learn more (Schneider & Hutt, 2014). Schneider and Hutt (2014) cited a University of Illinois report from 1968 about grades that found that focusing on grades encouraged cheating, gaming the

system, and adopting views that aligned with the instructors. Clearly, grades being viewed as a commodity is not a recent phenomenon.

Motivation and economics may also play into the grade inflation quotient. Godfrey (2011) cites the pressure and high stakes nature of grades can force teachers to "choose between adhering to grading guidelines and using grades as motivators to reward good effort rather than achievement" (p. 15). Using grades as motivation, besides compromising the grade itself, can often backfire with unintended consequences. Dueck (2014) points out that feedback has been shown to be far more effective than using grading penalties. Although he acknowledges that students should receive consequences for their actions or inactions with regard to school behavior and assigned work, Guskey (2015) notes that some teachers use grades as weapons to hold over students.

Behaviors should be addressed and consequences should be imposed for misdeeds, but they should not reflect on the grade the student has earned through academic achievement. Dueck (2014) states that some students are unaffected by grading penalties and would prefer to take the penalty as opposed to making the effort required by the teacher. Guskey (2015) concurs that giving a student a failing grade for not doing the work often accommodates their wishes by allowing them to not do the work. Vatterott (2015) recognizes "most teachers have come to realize that good grades motivate students who value good grades and bad grades don't motivate anyone" (pp. 16-17). Wormeli (2006) concurs that low grades fail to motivate students and instead push them farther away.

Economically, especially at colleges and universities, Wongsurawat (2009) surmises that a system that is dependent on attracting and retaining students may inadvertently encourage grade inflation. Additionally, often student evaluation of their instructors becomes a part of the

instructor's evaluation of their job performance for the school or college. Human nature may play a factor in giving the benefit of the doubt to someone who is going to be evaluating one's performance in a rather subjective manner. This phenomenon where an instructor's leniency in grading affects student evaluation scores has been dubbed the leniency hypothesis (Patrick, 2011). The results of a study by Patrick (2011) on student evaluations of teachers, however indicated that the students' opinion of the teacher was not significantly impacted by the grade they expected to receive.

When well-intentioned teachers inflate grades to help one of their students get an advantage, there can be unintended consequences. For instance, when a teacher adapts grades to help a struggling learner, they do so believing that these adjustments and adaptations will help make the higher grade more attainable to the student and provide the impetus for motivation to work harder (Jung & Guskey, 2012). Jung and Guskey (2012) state that "in reality, these adapted grades can lead students to believe that their grades are a result of who they are, not what they do" (p. 10), and negatively affect motivation which is the opposite effect it was intended to have. Likewise, in the conclusion of his study on grade inflation affecting law school admissions, Wongsurawat (2009) found evidence that in the 1990s, grade inflation may have had a negative impact on minority enrollment in higher education as grade inflation is thought to have caused a greater emphasis on standardized scores.

An ACT study by Woodruff and Ziomek (2004) that researched grade point averages as compared to ACT scores from 1991 to 2003 found significant grade inflation among U.S. high schools during that time period, but a more recent ACT commissioned study by Zhang and Sanchez (2013) did not find significant evidence of overall grade inflation nationally from 2004-2011. Although it was found to be the trend during the time span, the study points out there are

still significant variance from school to school (Zhang & Sanchez, 2013). Likewise, Godfrey (2011), in studying grade inflation trends and SAT scores concurs in a longitudinal study stating that "despite the general rise in grades, standardized scores on the SAT remained relatively unchanged" (p. 6). Studies on grade inflation have mixed results; however, as research by Pattison, Grodsky, and Muller (2013) indicates an increase in grade inflation since the 1970s in high schools but a decrease at four-year colleges. They further state the drop in grades is especially pronounced in selective universities. Although a primary purpose of grades is to communicate progress to both the parent and the student, O'Connor (2017) says a grading system is ineffective when teacher-assigned grades fail to communicate the student's academic achievement level in an effective manner. He further warns not to use behaviors as part of the grade as it should reflect only the academic achievement of the student (O'Connor, 2017). Grades are often not accurate predictors of knowledge as it is not uncommon to see a B or C student excel on a standardized test or an A student come away with mediocre results (Vatterott, 2015).

Although there is no shortage of criticism for teacher-assigned grades, Bowers (2011), highlights that emerging research shows value in grades as a blended assessment of both student academic content knowledge and a student's habits and behaviors that add to their ability to succeed in the school setting. Although grades are known to be a statistically weak indicator of student academic knowledge, they strongly indicate overall outcomes for student success at school (Brookhart et al., 2016).

Toldson and McGee (2014) questioned the predictive effectiveness of the SAT and ACT as research has shown an inability to predict college success for minorities. Even going back over two decades, the Moffatt (1993) study showed that the SAT was not valid in predicting

academic success for African-American students at any age which he partially attributed to a "racial bias" (p. 5). Hiss and Franks (2014) studied a sample in excess of 123,000 college students and found that students with low grade point averages and high SAT or ACT scores generally do not succeed in college, whereas students with high grade point averages and low SAT or ACT scores tended to do well. So, even with limited validity, teacher-assigned grades may have some value because clearly they mean something.

The same non-academic behaviors and habits that are rewarded as part of a good teacher-assigned grade, are behaviors which may serve a student well in his or her future academic endeavors. Laurie (2009) showed while assessing students whether through standardized testing or using teacher-assigned grades how both were useful as they are most likely measuring different criterion, so their results should be expected to differ.

Ellison (2008) was also not so quick to render teacher-assigned grades as worthless citing how grades can supplement standardized test scores in looking at the student as a whole in helping predict future success. Zwick (2013) found that the best measure of future academic success may be a combination of the grade and standardized test score. Ellison (2008) further surmises "because grades act as a proxy for behavior and achievement orientation and also act as feedback to shape students' academic self-concepts, they are valuable predictors of outcomes that are influenced by those characteristics of students" (p. 2). Bowers (2009) holds that grades thus represent not just what the student knows, but "represent an assessment of how well students play the game of being schooled" (p. 10). Even so, some researchers feel there is untapped potential for data-driven decisions in teacher-assigned grades (Bowers, 2011).

Putting the value of teacher-assigned grades into perspective, Rick Wormeli (2006) says grades are the teacher's personal opinion, "not infallible truths about students' mastery" (p. 95).

He points out how it is a mistake to allow grades to get out of perspective regarding student self-worth as they are inherently subjective and inaccurate as a tool to measure student achievement Wormeli, 2006).

Standardized Tests

Standardized tests were created with the intension of counteracting the inherent problems with teacher-assigned grades. Standardized tests by design allow the comparison of apples with apples. They are designed to strictly measure academic knowledge and are the same test for all students regardless of where they live, what school they attended, or their background. Although this may seem fair on the surface, standardized tests which are used as a measuring stick for judging schools, promoting students, and as a predictor for college success, and whose scores serve as gatekeeper for most colleges, has had its fair share of controversy and criticism as well. Higgins (2009) comments that "of the many hot-button issues in education, standardized tests is one of the most contentious" (p. 3).

When speaking of standardized tests and high stakes testing, one tends to use them synonymously, however, there is a difference. Standardized testing can be high stakes testing, but not necessarily (Nichols et al., 2012). Nichols et al. (2012) describe high stakes testing as a process where the goal is to encourage teacher performance and stimulate student achievement by attaching meaningful consequences to the standardized test results. Standardized testing is defined by Gall et al. (2007) as "a test that has procedures to ensure consistency in administration and scoring across all testing situations" (p. 195). Therefore, a standardized test is a specific type of test, whereas high stakes tests depend on how the results are applied. Standardized tests were born out of distrust and inconsistencies in teacher-assigned grades. High stakes testing evolved from the call for accountability.

Although high stakes testing has been in American schools on a limited basis dating back to the 1870s, it was first made a prominent concern by the federal government when the National Defense Education Act (NDEA) of 1958 was enacted in response to the Soviet Union's Sputnik launch the previous year (Munro, 2008). Later, concern with decreasing SAT scores in the late 1970s and publication of *A Nation at Risk* in 1983 which was a clarion cry for school reform and increased student achievement (Huddleston, 2014). During the 1990s, President Bill Clinton called for an end to social promotion, urging schools to make certain students could pass tests to be promoted to the next grade (Huddleston, 2014).

The No Child Left Behind Act (NCLB) pushed for even more high stakes testing as the school accountability movement gained strength. NCLB was met with bi-partisan support as target groups were addressed, but sanctions were awaiting schools that did not make adequate progress as verified by test scores (Groen, 2012). NCLB has changed the landscape of American public education with accountability measures that are far reaching into every aspect of public education today, but nothing as pervasive as the high stakes accountability measures that have engulfed our teachers and students in the form of assessments. Thompson and Allen (2012) state that NCLB ultimately became equated with producing better test scores as their primary objective.

Zellmer et al. (2006) looked into the instructional ramifications to students of NCLB. They found innumerable opportunity costs, time lost to logistical preparation, test administration, and loss of services the students did not receive due to the heavy toll of high stakes testing in terms of instructional time, and time spent by teachers, administrators, guidance counselors, specialists, and support staff in preparing for and administering the tests (Zellmer et al., 2006). Mora (2011) points out "the importance ascribed to standardized tests within public education

has significant impact on pedagogical practices" (p. 1). Curriculum is often narrowed increasing instructional time for tested subjects and cutting out curriculum or activities that are not (Huddleston, 2014; Thompson & Allen, 2012). Narrowing the scope of content to only what may appear on the test and teacher-driven lessons to ensure a pace that will get the class through the material are common practices that are deleterious to the learning process (Mora, 2011). Mora (2011) contends this unintended consequence of high stakes testing probably robs students of opportunities to participate in learning activities that would promote and encourage skill development in problem-solving and other higher order thinking. Cutting so many of the aspects of schooling students find interesting and enjoyable has led many students to become apathetic about their schooling (Thompson & Allen, 2012).

One of the byproducts of high stakes testing is the pressure, anxiety, and stress associated with student test scores (Watson, Johnson, Loder, & Dankiw (2014). Watson et al. (2014) point out that holding educators accountable for the outcomes of their students is reasonable and appropriate, there are other factors that "call into question the fairness, accuracy, and usefulness of this structure of assessment and the resulting accountability" (p. 2). Watson et al. (2014) report that due to the accountability through high stakes testing, teacher morale was lower and they reported feelings of anxiety, guilt, and embarrassment. Negative experiences with the culture of testing has caused many educators to leave the profession (Watson et al., 2014).

Hazel (2010) found that as teacher stress associated with high-stakes testing increased so did incidents of bullying. Hazel (2010) surmised that the testing pressure "led teachers to teach in ways that contradicted their ideas of good educational practice and the students' social-emotional development was ignored" (p. 351) Emotional stress runs high for everyone involved, and teachers feel pressured to teach to the test. Research by Turnbull (2016) cites that forty percent

of parents reporting that stress due to testing negatively affected their children. Ullman (2005) found no evidence linking student achievement with high stakes testing pressure, and in fact found it led to an increase in dropout rates.

Huddleston (2014) points out other unintended consequences of high pressure, high-stakes testing such as many documented incidents of cheating on behalf of not just students, but teachers and administrators as well. There are cases where teachers focus on the students who are close to passing to get them over the hump, but do not want to waste time on the students who do not have a chance to pass the test. To top it off, most of the negative unintended consequences take place in lower income schools where students are most in need of a good education (Huddleston, 2014).

In studying high stakes testing and student outcomes, Nichols et al. (2012) have found that the pressure associated with accountability has not increased reading scores and has decreased graduation rates. Likewise long term studies have found that high consequence testing has little or no impact on student achievement and likely contributes to higher dropout rates (Reardon et al., 2008). Breiner (2015) points out that "the US education system is spending billions per year on math and science testing, with very little or mixed research supporting their effectiveness" (p. 103) In fact, NAEP data reveals students progressed faster in math before NCLB and the onslaught of testing occurred, while reading scores have been stable during the same time period (Breiner, 2015). Watson et al. (2014) found studies reporting the negative effects of high stakes testing outnumbering similar studies that reported positive outcomes at a rate of nine to one. Nichols and Valenzuela (2013) found that fear of failing or job loss did motivate teachers and students. Supovitz (2009) did find high stakes testing motivated both teachers and administrators to examine their instructional practices in attempt to align their

curriculum and classroom instruction to the high stakes exams their students would be held to. Huddleston (2014) also mentions teacher incentive and motivation as well as interventions for at-risk students which have sprung up to try to hold back the sanctions that come with failure. Supovitz (2009), however, concludes although there are positives to high stakes testing, it is not a true reform as it exposes shortcomings and problems within the classroom, school, or district, but does not provide solutions or guidance as to how to fix the problems.

Although Reardon et al. (2008) concur with the argument for high stakes testing for graduation was to create incentives for teachers and students to work harder towards the requirements, they found that for others adding more barriers to graduation discouraged students from persevering and raised the rate of students who chose to drop out of school. This finding was even more pronounced in socially and academically disadvantaged students (Reardon et al., 2008).

Ricketts (2010) determined "gender, ethnicity, disability status, Limited English Proficiency status, and socioeconomic status all significantly impact" (p. 140) standardized test scores. Although some of this can be deduced intuitively such as suburban schools will generally outperform urban schools, students with learning disabilities will usually average a lower score than their non-disabled peers on a test to measuring learning, and students with Limited English Proficiency may not score as well on a test given in English as native speakers. Huddleston (2014) notes that teachers have long recognized that students from some backgrounds tend to flourish more than others which researchers have often attributed to opportunity gaps that follow class lines. The literature can shed some insight as to why these groups are so effected.

Duckworth and Seligman (2006) looked at gender differences and showed how girls get better grades when assigned by teachers, however boys get higher scores on standardized tests

(p. 205). Clark (2013) concurs standardized tests prove the existence of a gender achievement gap. Duckworth and Seligman (2006) concluded that girls are more self-disciplined which gave them an advantage when grades took behavior measures into account, whereas the standardized tests filtered out factors other than the student's academic achievement. Harvey, Suizzo, and Jackson, (2015) concurred that girls were found to attract the most positive teacher bias most likely stemming from their behavior. Similarly, Duckworth et al. (2012) proposed "that standardized achievement test scores assess competencies determined more by intelligence than by self-control, whereas report card grades assess competencies determined more by self-control than by intelligence" (p. 11).

Kling et al. (2012) confirmed the findings above as their study found women to earn higher grades than men, score lower on the SAT, and score higher on a conscientious scale which measured "the tendency to be careful, deliberate, and thorough" (p. 600). This study showed how students who were more conscientious earned higher than predicted grades based on their SAT scores, in most cases this meant women (Kling et al, 2012). Perhaps this explains why girls consistently get better teacher-assigned grades even though they lag slightly behind boys on standardized tests (Borich & Peck, n.d.).

Students with Disabilities

Another group of paramount concern is students with disabilities. Protected by the Individuals with Disabilities Education Act (IDEA) or Section 504 of the Rehabilitation Act of 1973, students with disabilities are put in a precarious situation with some of the high stakes testing. According to IDEA, students with disabilities must participate in the same high stakes testing as their non-disabled peers in one of three fashions; they either take the test without modifications or accommodations, take the test with modifications or accommodations, or take

an alternate assessment (Yell et al., 2012). Historically students with disabilities were excused from taking standardized tests and thereby schools were not being held accountable for their progress, however that changed after the 1997 Amendments to IDEA which required the inclusion of students with disabilities in state testing (Lai & Berkeley, 2012). In 2004, with its reauthorization, IDEA required the provision of testing accommodations allowing students to be appropriately included in the testing (Lai & Berkeley, 2012).

Although students with disabilities are often allowed accommodations by the states, frequently the research regarding those accommodations is limited if existent at all (Lai & Berkeley, 2012). It is difficult for schools to interpret data when results from studies are mixed (Lai & Berkeley, 2012). Prior studies have shown how students with disabilities do, in fact, benefit from test accommodations, specifically from having test items read aloud and extended time (Powell, 2012). Even so, there are few accommodations which have been proven to give a differential boost (Jamgochian & Ketterlin-Geller, 2015). Jamgochian and Ketterlin-Geller (2015) explain that to observe a "differential boost, two outcomes are expected: the test scores for students with disabilities are higher when they receive an accommodation than when they do not, and the increase in test scores is not observed in their peers without disabilities" (p. 29).

Powell (2012) points out accommodations should not alter the construct of the test, but rather allows the student with disabilities to perform in an environment that lessens the barriers that are present due to their disability. Jamgochian and Ketterlin-Geller (2015) further state "test accommodations should not change the meaning or interpretation of the test scores, but instead, reduce the impact of the features of the tests that are irrelevant to the construct and act as access barriers" (p. 29). Some examples of access barriers given by Jamgochian and Ketterlin-Geller (2015) involved insufficient audio or visual support, unnecessarily difficult vocabulary, lack of

physical supports, as well as anxiety and fatigue due to the testing. Powell (2012) cites that accommodations often include small group or individual testing, reading tests aloud, allowing extended time to finish tests, using large print, or providing scribes to assist in writing.

One accommodation with promise which has been studied involves the test response format (Powell, 2012). Some research shows performance differences between students answering multiple choice questions versus answering constructed-response questions (Powell, 2012). It would appear that students answering in the multiple choice format enjoy a significant advantage over those answering constructed-response questions even though the questions stems were identical (Powell, 2012). In a study involving math multiple choice questions and constructed-response questions, Powell surmised "the advantage (of multiple choice questions) may occur because students are given the opportunity to discriminate between reasonable and unreasonable responses. With this discrimination, students show mathematical competence even if they are capable of deriving an answer independently" (p. 7).

Making an accommodation should mediate the student's personal abilities and characteristics and the barrier that may skew the measurement of the student's knowledge and skills (Jamgochian & Ketterlin-Geller, 2015). Lai and Berkeley (2012) found teachers did a poor job when it comes to correlating which students needed accommodations and students who actually benefited from them. It was also found teachers often over accommodate or assign accommodations that the students do not profit from, and in some cases negatively impacted students with learning disabilities (Lai & Berkeley, 2012). Ysseldyke, Nelson, Christenson, Johnson, Dennison, Triezenberg, Sharpe, & Hawes (2004) cites two matters of importance with test performance; student opportunity to learn the material they are tested on, and ensuring access to the test through proper accommodations.

Parents who feel requiring their disabled student to pass an exit exam in order to graduate is contrary to the principles and ideals of IDEA have taken their complaints to court (Yell et al., 2012). Some parents feel high stakes testing discriminates against children with disabilities which is a violation of Section 504, while others question if it is a violation of a child's right to a free, appropriate public education which is guaranteed by IDEA (Yell et al., 2012).

If students are being educated in self-contained classes or are learning a modified curriculum, it is simply unfair, not to mention illegal, to hold them to the consequences of failing a high stakes test that their education did not prepare them for. Jung and Guskey (2012), point out that the *Board of Education v. Rowley* case in 1982 specified "legal provisions require that IEPs written for children with disabilities enable them to achieve passing marks and advance from grade to grade" (p. 6). Furthermore, Jung and Guskey (2012) warn that a failing grade may prove legally that the student was not provided with appropriate services. On the other hand, the opposite action of assigning passing grades to students who have not succeeded in making the prescribed performance standards, would also be inappropriate as it would not accurately reflect the student's academic achievement (Jung & Guskey, 2012). Yell et al. (2012) estimate that in spite of already high failure rates for students with disabilities, high stakes exit exams used to determine graduation eligibility will continue to increase.

LaSalle, Roach, and McGrath (2013) cite the 1989 *Daniel R.R. v. State BOE* as an early case that established a student with disabilities should "be educated in the least restrictive environment to the maximum extent possible" (p. 136). This ruling came with a two-pronged test; that a student with disabilities must be educated in a general education classroom with necessary supplementary aids and services necessary to succeed, and if that was not possible the student had to be mainstreamed to the maximum extent possible (LaSalle et al., 2013). It also

stated students could not be denied services because of the cost to the school district (LaSalle et al., 2013). This ruling was reinforced in *Oberti v. BOE of Clementon School District* in 1993 with the statement "inclusion is a right, not a special privilege for a select few" (LaSalle et al., 2013, p.136). It will be of the utmost importance for schools to follow the court cases and adjust accordingly.

Despite the claims of unfairness, students with disabilities have been making noticeable progress in academic achievement and other outcomes (Ysseldyke et al., 2004). Since the inception of NCLB, students with disabilities have increased graduation rates, decreased dropout rates (Ysseldyke et al., 2004), are receiving better instruction, have increased participation and performance levels in high stakes testing, are receiving the skills necessary to succeed at the higher grades, and are the subgroup with the biggest gains in passing mandatory tests (Katsiyannis, Zhang, Ryan, & Jones, 2007).

On the negative side, it has been noted how the lower grades are experiencing overage students due to mandated retentions, insufficient funding to adequately remediate those who fail the tests, and the fact that no matter how well a student does throughout the school year, they can be retained based on the results of a single assessment (Ysseldyke et al., 2004). It has also been found how students with language-learning disabilities are at high risk for academic failure (Koutsoftas, Gray, Nippold, & Schneider, 2012), that student with disabilities in general were most vulnerable to failing and falling victim to the associated consequences, and also students with disabilities make the school appear to be less effective (Katsiyannis et al., 2007; Tanis, 2014), and the implications of many failing to earn their diploma because of not passing a state exam (Ysseldyke et al., 2004).

Economically Disadvantaged Students

Vernaza (2012) asserts "poverty is the single best explanation research has found for why children differ in ways that affect school performance both before they enter school and after they are enrolled" (p. 2). Socioeconomic status (SES) factors into multiple facets of a child's life (Clark, 2013). A strong positive correlation has been found between a student's SES and standardized test performance (Vernaza, 2012). Clark (2013) highlights that SES "has an impact on human development, especially on health, education, and human well-being" (p. 109). It also impacts resources, educational opportunities, and family income has great baring on where a family lives, and the school the child will attend, as well as having a definite effect on how a student achieves in the classroom (Clark, 2013).

Children from economically disadvantaged backgrounds enter school less prepared for success in the classroom than their middle class peers (Vernaza, 2012), and Watson et al. (2014) point out low SES "can have enormous effects on each individual student's performance and consequently, their stress levels" (p. 3). Eric Jensen (2009) states common behaviors of children in poverty are chronic tardiness, lack of motivation, inappropriate behavior, use of profanity, and disrespect to others. Attendance problems may be attributed to parents having negative attitudes with respect to school due to their past experiences (Jensen, 2009). Jensen also states how a lack of hope and optimism contribute to a lack of motivation, and children in poverty often see the future as more negative than positive. Students are simply not likely to try if they feel that failure is imminent or that they are not smart enough to succeed (Jensen, 2009).

Heppt, Haag, Bohme, and Stanat (2014) point out a considerable difference in language mastery from an early age between families with low versus high socioeconomic status. Jensen (2009) states children who grow up in poverty typically have a substantially smaller vocabulary

than their middle-class peers. Much of the vocabulary deficit can be explained by a lack of opportunity. Children from low-income families on average hear about 13 million words by the time they are four years old, whereas middle class children hear twice that many words, and upper class are exposed to three times as many words as their low-income peers (Jensen, 2009). There is also a large disparity with regards to the number of books and picture books between homes of differing SES (Heppt et al., 2014). Mothers from middle class families have been shown to read more to their children and use a richer vocabulary and more complex sentences when engaging their children than their less privileged counterparts (Heppt et al., 2014). Heppt et al. (2014) also note that "there is strong evidence that these socioeconomic related differences in language exposure and home literacy environment predict preschoolers' vocabulary knowledge, use of complex syntactical structures, and oral language skills" (p. 63).

Nichols and Valenzuela (2013) assert "test related pressures are greatest among high-poverty schools because they have the most difficulty passing test benchmarks and are, therefore, disproportionately susceptible to sanction-based consequences" (p. 155). Data also shows much more negative effects that high stakes testing has on economically disadvantaged students. When students have to pass a test to graduate, poor students are at a greater risk of dropping out of school (Nichols & Valenzuela, 2013). Vernaza (2012) shows that the pressures are greater for teachers as well as the students in stating "educators teaching in Title 1 schools (i.e., schools receiving federal financial assistance due to the high enrollment of students from low socioeconomic backgrounds) face additional pressures of preparing economically and educationally disadvantaged children" (p. 1). It has also been documented how teachers feel frustrated over being compared to teachers from middle class schools even though they are responsible for test scores of economically disadvantaged students (Vernaza, 2012).

One of the unintended consequences of high-stakes testing is increased retention rates (Huddleston, 2014). Retention is holding a student in the same grade for an additional year because they have failed to make a pre-determined cut-off score on a standardized test or because the teacher feels that they do not possess the necessary skills needed for the next grade. Besides being used to determine who is eligible to graduate, high stakes tests are used to make determinations with regards to promotion or retention. Huddleston (2014) observes "low-income, ethnic minority students are most often targeted for retention" (p. 11). Bali, Anagnostopoulos, and Roberts (2005) likewise found "racial minority students and students living in poverty constitute the majority of those who are retained" (p. 134). Those in favor of retention argue that students should be left back until they are able to master the academic skills necessary for the next grade, while opponents to retention claim that the most vulnerable students are targeted, rarely show any appreciable academic improvement, and increase their risk of dropping out of school (Huddleston, 2014).

Nagel (2015) purports "holding back students to repeat a year of school is one of the most deleterious actions on student achievement. Retention is one of the few areas where it is difficult to find any studies with a positive effect" (p. 64). Jimerson & Renshaw (2012) found that students who are retained are 5-10 times more likely to drop out. Jimerson and Renshaw (2012) further found although some modest academic gains can be found in students who were retained, research suggests that often no gains were made and even if they were, they were not maintained in future years following being retained. Citing a longitudinal study, Jimerson and Renshaw (2012) also state that "retention is predictive of emotional distress, low self-esteem, poor peer relations, tobacco use, alcohol and other drug abuse, early sexual activity, suicidal intentions, and violent behaviors during adolescence" (p. 12).

Although at this juncture in history, every student in the United States of America is subjected to high stakes testing and standardized testing throughout their elementary, middle, and high school careers, the two biggest standardized tests for college admission are the Scholastic Aptitude Test (SAT) and the American College Test, now known as the ACT. Various studies have shown the SAT and high school grades have a strong correlation for middle class, traditional Caucasian students and was a strong predictor for success among that population of students (Borich & Peck, n.d.; Hiss & Franks, 2014). The SATs, however, were invalid when predicting success for African-American students at any age (Borich & Peck, n.d.; Hiss & Franks, 2014; Thompson & Allen, 2012; Toldson & McGee, 2014).

A deeper look must be taken at standardized tests scores which prove invalid for minority populations (Muir, 2005) especially if they are being used as a gatekeeper for future academic opportunities. Studies show since the inception of the SAT and ACT tests, African-American students have attained the lowest scores of any of the racial groups measured (Toldson & McGee, 2014). One has to wonder about the cultural fairness of the test. One of the primary goals of the accountability movement and NCLB was to close the achievement gap between African-American students and their white peers. According to Thompson and Allen (2012), not only has that goal failed but it has "been harmful to teachers and countless African-American students, as indicated by high school dropout rates, NAEP, reading and math scores, student apathy, under-representation in gifted classes, and school suspension and expulsion rates" (p. 224). Much of the literature below aimed at low socioeconomic students applies to many African-American students as they are overrepresented in the poor neighborhoods and low socioeconomic schools, especially urban districts (Thompson & Allen, 2012).

Along these lines, according to Zwick and Himelfarb (2011), Latino students score a standard deviation lower than their White and Asian-American peers. In an Arizona study, Garcia (2011) found only 29% of students from predominantly Latino high schools passed the graduation test as sophomores, whereas 47% of students from predominantly white schools passed as sophomores. Garcia (2011) warns if states continue to narrow the curriculum down to fit a graduation test, the results will severely limit the prospects of postsecondary education for Latino students, even from schools that meet accountability criteria. Latinos, too, are overrepresented in low socioeconomic neighborhoods and schools and face disadvantages not encountered by many of their more affluent peers (Garcia, 2011; Rodriguez & Arellano, 2016).

Zwick and Grief-Green (2007) showed a higher correlation between SAT scores and socioeconomic status (SES) than between SAT scores and grades. In other words, high income equals high SAT score, low income equals low SAT score which had a higher correlation than high teacher-assigned grades equals high SAT scores, low grades equals low SAT scores. Zwick and Grief-Green (2007) point out the association between test scores and SES is common knowledge in the measurement community even though it often grabs headlines as news. That study boldly concludes it is "indisputable that SAT scores and SES are positively correlated" (Zwick & Grief-Green, 2007, p. 21). Likewise, Borich and Peck (n.d.) note that standardized test scores predict better for middle class students than for disadvantaged students.

Sackett et al. (2012) affirms the SAT-SES correlation and adds that because schools are relying on these tests for admission decisions, that access to higher education is effectively being limited. The study adds how aside from that fact, lower socioeconomic students are less likely to apply to college in the first place (Sackett et al., 2012). Although Sackett et al. (2012) do not find a sinister plot to keep low socioeconomic students down, he attributes the correlation to "some

combination of educational opportunity, school quality, peer effects, and other social factors" (p. 1006). Zwick and Grief-Green (2007) point out the impact that school resources, or lack thereof, have on student achievement, and ultimately test scores. Nonetheless, students in the lower socioeconomic realm have a decided disadvantage to their more affluent peers when it comes to scoring well on the SAT.

Where SES is involved, Zwick and Grief-Green (2007) state "if we recognize that adequate resources are needed to foster student achievement, and that many students do not have the kinds of home and school environments that are conducive to learning, how can we expect test scores to be uncorrelated with socioeconomic factors?" (p. 43). It has been estimated students in poor neighborhoods are three years behind their wealthier peers academically (Higgins, 2009). As pointed out by Froman et al. (2010) despite their careful design, tests are inherently flawed and only measure a small sample in a specific time. How do we come up with a test that is fair to all the groups of children? The Zwick and Grief Green (2007) study summarizes that there will not be a socioeconomically equitable test until society is socioeconomically equitable.

Despite all of the issues standardized testing presents, Higgins (2009) suggests that although by no means the perfect tool, standardized tests are currently the best option to quantify student achievement. He surmises even though standardized tests are highly accurate, curriculum must first be aligned so that everyone is teaching the same material and using the same achievement tests. This would identify schools who were excelling and bring about discussions of best practices, and in this way testing could become a more effective tool for student achievement (Higgins, 2009).

Summary

It appears from the literature there is no perfect assessment currently in practice. Standardized tests have their faults, and although they are important in their use as predictors, they are clearly not infallible. Toldson and McGee (2014) cite that “black students' scores on the SAT and ACT have been relatively flat for the last 20 years, although significant gains have been made in Black students' graduation rates and college-degree attainment” (p. 1). Therefore there must be other or better means to predict success. Kettler, Elliott, Kurz, Zigmond, Lemons, Kloo, Shrago, Beddow, Williams, Bruen, Lupp, Farmer, & Mosiman (2013) looked extensively at assessment measures and found the best predictor on a standardized achievement test was the previous year’s standardized achievement test.

Teacher assigned grades, although scoffed at in much of the literature, seem to be able to contribute something of value to predicting future success. Although most of the literature pushes for grades which only reflect academic achievement, others have shown there is value in grades as they may take into account softer skills and factors such as self-discipline, cooperation, effort, and attendance. Are these not important complements to pure intelligence in predicting future success? It seems that teacher-assigned grades should not be discounted to the point of worthlessness.

One possible solution which has gained notoriety is the standards based progress report. Welsh, D’Agostino, and Kaniskan (2013) describe standards based progress reports as differing from traditional letter grades as they require “teachers to report student performance levels on specific educational goals instead of broad content area” (p. 26). Guskey and Jung (2009) believe standards based progress report help improve communication with parents. It is also thought to help teachers focus on their instruction on individual students as well as making the teachers

more familiar with the standards they are charged with teaching (Welsh et al., 2013). Another benefit Guskey (2001) highlights is in "helping teacher to differentiate between process, progress, and the quality of student work products" (p. 21).

Standards based progress reports are also believed to have positive implications for students with special needs as they are easily adaptable to IEP goals and modified standards (Guskey & Jung, 2009). On this type of report, students are graded in each of the standards from the curriculum. This gives much more specific information than a letter grade as parents and students can see which particular standards they have mastered and which still need work. Guskey and Jung (2009) surmise "as a result, students with disabilities and their families can have information that they are able to interpret accurately and use effectively" (p. 60). A follow up study by Swan et al. (2014) found most teacher and parents who used the standards based report card found it beneficial in communicating student achievement with regards to the standards.

O'Connor (2017) and others have pushed for standards based grading that purifies grades from influences outside of academic achievement, but relatively few schools are using such a system. Teachers and parents alike tend to resist changing formats as each has difficulty moving away from traditional grades, regardless of the arguments against their validity. Human nature dictates how people have an aversion to change and old habits are hard to break as Allen (2005) acknowledges with "it is hard to change the invalid grading schema that has become embedded in our minds" (p. 218).

Summarizing the research above, it is necessary to move towards a system that separates student achievement from characteristics such as effort, participation, and attendance (yet somehow still acknowledges them), and is acceptable to teachers, parents and students. In the

meantime, it appears the status quo will continue for the foreseeable future in using both teacher assigned grades and standardized test scores, as imperfect as they may be.

CHAPTER THREE: METHODS

Overview

This chapter includes a rationale for the design provided. The research questions, null hypotheses, participants and setting, instrumentation, procedures, and data analysis will be described. This quantitative correlational study is designed to investigate the relationship between standardized tests and teacher-assigned grades. This study examined the Ohio End-of-Course exams for 8th grade Science and the numerical teacher-assigned final grades for the same sample of students using scores and grades from the 2014-15 school year. The strength of the relationship overall will be tested using a bivariate correlation, while three Pearson product-moment correlation coefficient analyses will determine if the strength of the correlation for students with disabilities and low socioeconomic students varies.

Design

This quantitative correlational study determined the strength of the relationship between AIR end-of-course test scores in 8th grade Science and teacher-assigned final grades for students in 8th grade Science. Gall, Gall, and Borg (2007) define correlational research as "studies in which the purpose is to discover relationships between variables through the use of correlational statistics" (p. 332). This study was quantitative by definition because it performed statistical analyses on numerical data to look for trends (Gall et al., 2007). The correlation study determined to what degree the teacher-assigned end of course final numerical grade in 8th grade Science is predictive of the AIR end of course test for 8th grade Science. One would assume that the end-of-course grade showing how the student has performed in class should correlate closely to an end-of-course exam that is designed to measure academic achievement of the standards in

8th grade Science. There were two variables in the correlation as the teacher-assigned grade is the Predictor Variable and the AIR end-of-course Science test score is the Criterion Variable.

Research Questions

RQ1: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year?

RQ2: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test scores for the 2014-2015 school year for students with disabilities?

RQ3: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test scores for the 2014-2015 school year for students receiving free or reduced lunches (economically disadvantaged)?

Null Hypotheses

$H_0$1: There is no statistically significant correlation between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year.

$H_0$2: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students with disabilities.

$H_0$3: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students without disabilities.

H_04: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students receiving free or reduced lunch (economically disadvantaged).

Participants and Setting

The participants in the study were former 8th grade students from the 2014-2015 school year with regard to teacher-assigned final grades for 8th grade science students and the AIR end of course test scores for 8th grade Science. The sample size of 2077 students ensured viable numbers in all the subgroups. All of the students were public school students in a five county area of Northeast Ohio. Data were solicited from schools in Atwater, Shore, Hilltop, Southward, and Broadmoor counties in Ohio. These five counties make up the northeastern tip of the state.

Demographically, Atwater County is 93.4% white, 3.8% African-American, 3.8% Hispanic or Latino with a median household income of $40,516 and 18.7% of the 99,175 residents living below the poverty level (Ohio Quick Facts from the US Census Bureau, 2015). Shore County has 229,203 residents of which 93.0% are white, 3.9% African-American, and 4.0% Hispanic or Latino. The median income in Shore County is $56,018 and 9.2% of the residents live below the poverty level (Ohio Quick Facts from the US Census Bureau, 2015). Hilltop County has 94,295 residents earning a median income of $67, 563. White residents make up 96.9% of the population, while 1.3% identify themselves as African-American and 1.4% are Hispanic or Latino. Only 8.1% live below the poverty level (Ohio Quick Facts from the US Census Bureau, 2015). Southward County has a median income of $42,880 with 17.4% of its 205,175 residents below the poverty level. Southward County has a racial makeup of 88.8%

white, 8.6% African-American, and 1.6% Hispanic or Latino (Ohio Quick Facts from the US Census Bureau, 2015). Broadmoor County has 293,204 residents of which 80.9% are white, 15.9% African-American, and 5.4% Hispanic or Latino. Broadmoor County has a median income of $41, 058 with 17.6% of residents below the poverty level (Ohio Quick Facts from the US Census Bureau, 2015).

Even these Census Bureau statistics only show averages when the fact is there are tremendous discrepancies even within the counties. Broadmoor County, for example, is home to one of the poorest and lowest performing urban districts in the state of Ohio. Just outside the city are three wealthy suburban districts that continually have some of the top scores in the state and whose incomes bolster the median. The sample was an accessible sample which Gall et al. (2007) define as "all of the individuals who realistically could be included in the sample" (p. 167). It is thus effectively a convenience sample. The demographics of the sample was dependent on the positive responses received and involved urban, rural, and suburban school districts and students.

Instrumentation

The instrument used for this quantitative correlation was the AIR end of course 8th grade Science test scores for the 2014-2015 from the same sample of 8th grade students. The purpose of this instrument was to measure student achievement in 8th grade Science for all Ohio students with the exception of 1% of students who are alternatively assessed.

In 2010, Ohio approved and adopted rigorous new standards of academics for Science (Science State Test Specifications and Resources, 2015). A new model curriculum for Grade 8 was developed he following year (Science State Test Specifications and Resources, 2015). Ohio law called for an assessment that aligned to both the model curriculum and the new standards.

According to Ohio Revised Code 3301.079 the assessment is to be administered online and will measure student achievement with regards to the standards. It also specifies that the assessment will be given in two parts and that the data from the test results will be made available to the school no later than June 30th of each year (Science State Test Specifications and Resources, 2015).

The test blueprint is constructed to have three reporting categories: Earth Science with 21-23 questions accounting for 38%-43% of the test, Life Science with 16-18 questions for 29%-33% of the test, and Physical Science with 16-18 questions and making up 29%-33% of the assessment (Science State Test Specifications and Resources, 2015). The questions are divided into four item types: multiple-choice, enhanced selected-response, a machine-scored constructed-response, and a human-scored constructed response (Science State Test Specifications and Resources, 2015). The grading scale will produce a raw score that will be converted to a scaled score. This scaled score is used to determine if the student has passed or failed the test and at which level. The scaled score ranges and categories are: 575-674 Limited; 674-699 Basic; 700-724 Proficient; 725-765 Accelerated; and 766-868 Advanced (Scaled Score Ranges for Ohio State Tests, n.d.). A score over 700 or Proficient is considered passing for the school report card, with anything under 700 failing. (Ohio's State Tests, n.d.) The scaled scores will be used as the data for this study.

According to the Ohio Department of Education, the reliability level on AIR end of course Science tests is .78. Although 2014-2015 is the first year for the Air End-of-course tests for Science, AIR is the same company that was hired to handle the Ohio Graduation Tests which had been in existence since 2005 (Ohio's State Tests, n.d.), but are being phased out as the New Ohio Standards were implemented. The New Ohio Standards closely emulate the Common Core

Standards however the name was changed when a grassroots resistance to the Common Core began to gain momentum (Layton, 2014).

The final grade in 8th grade Science that was assigned by the teacher and used in the school transcripts for the 2014-2015 school year from each student in the sample is archival data and served as the predictor variable. Both the validity and reliability of teacher-assigned tests has been called into question as there is a well-documented lack of accuracy and consistency in grades from teacher to teacher and school to school (Allen, 2005, O'Connor, 2011).

Procedures

After submitting an IRB packet and gaining approval from the Liberty University Institutional Review Board, the researcher executed the research. The researcher first solicited approval to collect and use data from individual school districts in Atwater, Shore, Hilltop, Southward, and Broadmoor Counties in Ohio to gather a viable sample. In order to collect the data, the researcher sought permission to use the scores and corresponding teacher-assigned grades without data from which any student might be identified. All schools in the five county were contacted. As much data was gathered as could be from the cooperating schools in order to garner a viable sample. After gathering the data, data analyses was performed to test the relationships studied as well as each group to be examined including students with disabilities and students with low SES.

Data Analysis

The data was entered into SPSS and a Bivariate Correlation was performed using a Pearson *r* for each of the three hypotheses. Gall et al. (2007) calls this test the "product-moment correlation coefficient (*r*)" and says it is "computed when both variables that we wish to

correlate are expressed as continuous scores" (p. 347). The Pearson *r* was the appropriate test for determining the strength of correlations on the variables in this study (Gall et al., 2007).

The Assumption of Bivariate Outliers test used a scatter plot between the predictor variables (x) and criterion variable (y) to look for extreme outliers. The Assumption of Bivariate Normal Distribution used a scatter plot between the predictor variables (x) and the criterion variable (y) and looked for the classic "cigar shape." If there was the appearance of an abnormal distribution, the Kolmogorov-Smirnov test for Assumption of Normality was employed (Gall et al., 2007). Each of the three hypotheses reported descriptive statistics (*M*, *SD*), number (*N*), degrees of freedom (*df*), observed *r* value (*r*), significance level (*p*) and power.

CHAPTER FOUR: FINDINGS

Overview

The purpose of this study was to examine the relationship between teacher-assigned grades and scores on a standardized, high stakes, end of course test. A bivariate correlation, the Pearson product-moment correlation, was used to create a correlation coefficient to quantify the relationship between the criterion variable which was the AIR End of Course Test score and teacher-assigned grades, the predictor variable, for those same students. Also factored was whether or not the participant was a student with or without a disability and/or if the student was disadvantaged in being from a low socio-economic situation as defined by whether or not they received a free or reduced lunch. On two occasions when the normality of the distribution was not tenable, a Spearman's rho test was conducted.

Research Questions

RQ1: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year?

RQ2: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test scores for the 2014-2015 school year for students with disabilities?

RQ3: Is there a relationship between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the Ohio AIR End-of-Course 8th grade Science test scores for the 2014-2015 school year for students receiving free or reduced lunches (economically disadvantaged)?

Null Hypotheses

$H_0$1: There is no statistically significant correlation between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year.

$H_0$2: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students with disabilities.

$H_0$3: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students without disabilities.

$H_0$4: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students receiving free or reduced lunch (economically disadvantaged).

Descriptive Statistics

The sample consisted of a total of 2077 participants from 9 school districts in a five county area of northeast Ohio. The students were all in the 8^{th} grade during the spring 2015 AIR end of course Science assessment and also received a final teacher-assigned grade for 8^{th} grade science that same year. Of the total number of participants, 1318 or 63.4% were classified as disadvantaged based on qualifying for free or reduced lunch. Students with disabilities comprise 283 students or 13.6% of the sample. Three of the nine districts were urban districts. Those students make up 966 or 46.5% of the participants. Two districts were rural districts and make up

186 or 8.95% of the participants. The four remaining districts are combinations of suburban and rural represent 925 students or 44.5% of the participants in the study.

The data for teacher-assigned grades were scaled as the grade of A was quantified as a 4, a B was a 3, a C was a 2, a D was signified with a 1, and an F was given a zero. The test scores ranged from a minimum score of 575 and a maximum score of 843. In the categories of disadvantaged and students with disabilities, a 1 indicates yes and a 2 indicates no. For example, under the category of disadvantaged if the student is marked with a 1, it signifies that the student has been classified as disadvantaged. By definition of this study, disadvantaged means the student receives a free or reduced lunch. Likewise a 1 under SWD would indicate that the student has been identified with a disability as defined in this study as having an IEP.

Student data that was incomplete was discarded from the study. For example, if students had been issued a final teacher-assigned grade but did not have an end-of-course test score, they were eliminated. There were also a few instances where an end of course assessment score was entered but the grade was missing or recorded as I for incomplete. Those students were also eliminated. Finally, data was run for outliers and each outlier was checked to make certain that it was truly part of the data distribution and not a data entry error.

One school had two final grades listed (one for each semester). In that case the grades were averaged for a final grade. For this study, if the student received a B for each semester the grade of B was used for the study. If the student received an A first semester and a B second semester, the study used a final grade of B as that was the direction of the trend. Likewise if the student received a B first semester and an A second semester, an A was recorded as the final grade because it trended that direction. If a student received a B one semester and a D the other semester then it was averaged and recorded as a C.

Table 1

Data for AIR End-of-Course Science Test Scores

AIR Test Scores	N	M	SD
All	2077	700.77	43.59
SWD	283	663.56	35.96
SWoD	1794	706.64	41.76
Disadvantaged	1318	686.03	38.61

Table 2

Data for all Teacher-Assigned Grades

Teacher Assigned Grades	N	M	SD
All	2077	2.75	1.076
SWD	283	2.1696	1.00329
SWoD	1794	2.845	1.0585
Disadvantaged	1318	2.5152	1.08668

Results

Null Hypothesis One

$H_0 1$

The first null hypothesis was tested as follows:

$H_0 1$: There is no statistically significant correlation between teacher-assigned course grades in 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year.

A box and whisker plot was used to screen for outliers. Outliers were found, however, each was double checked to ensure their validity as part of the study and were included in the data. The next step was to test for assumptions. Based on the histogram as shown in *Figure 1* and normal Q to Q plot of the AIR test results in *Figure 2*, the distribution appears to be nearly normal.

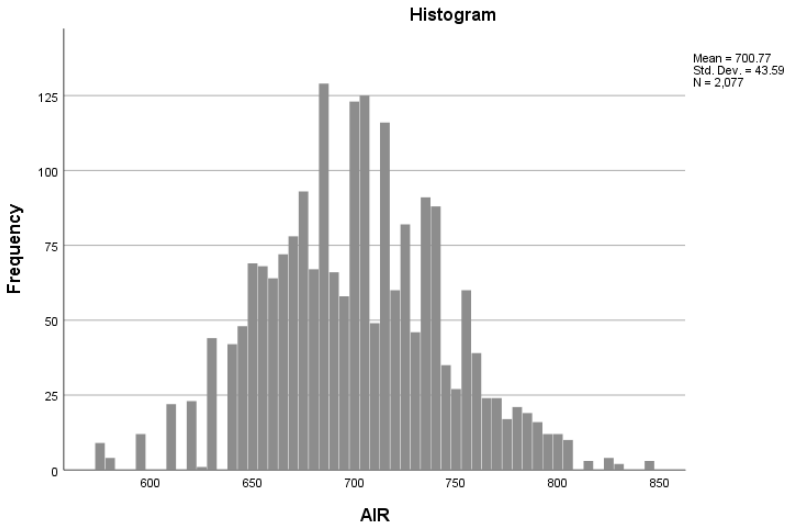

Figure 1: Histogram of all AIR Test Scores

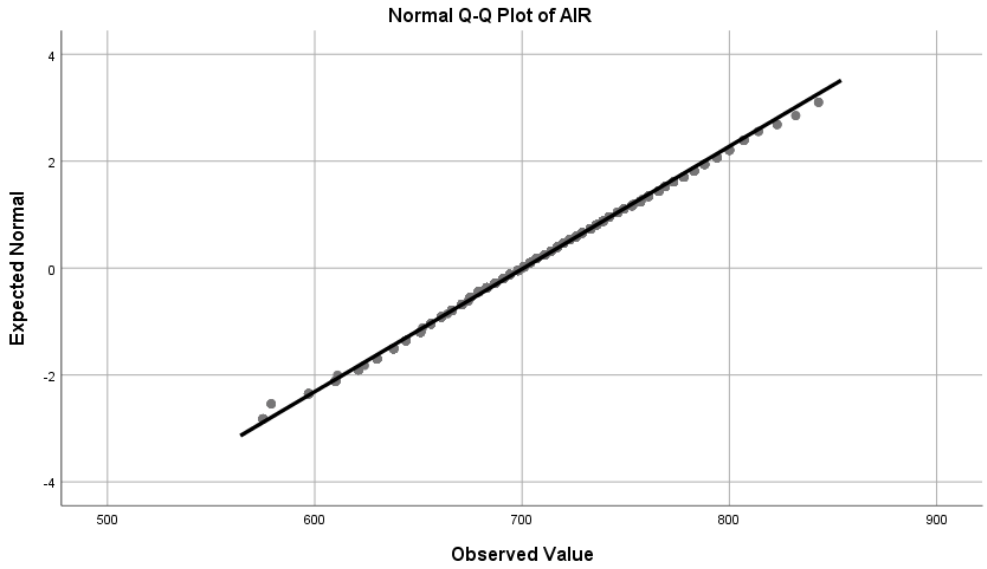

Figure 2: Normal Q to Q Plot of all AIR Test Scores

The histogram in *Figure 3* and Normal Q to Q Plot of the Teacher Assigned Grades as shown in *Figure 4*, however, do not appear to be normally distributed and thus the assumption of normality is not tenable. Therefore a Spearman's Rho test was implemented with the results shown in Table 5. Each of the two variables, the teacher-assigned grades and the AIR test results, were observed independently of the other.

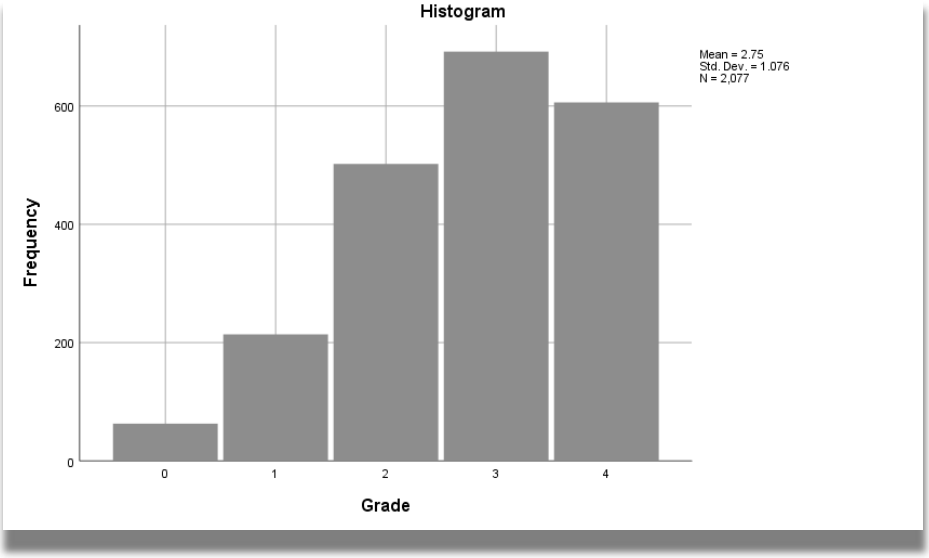

Figure 3: Histogram of all Teacher-Assigned Grades.

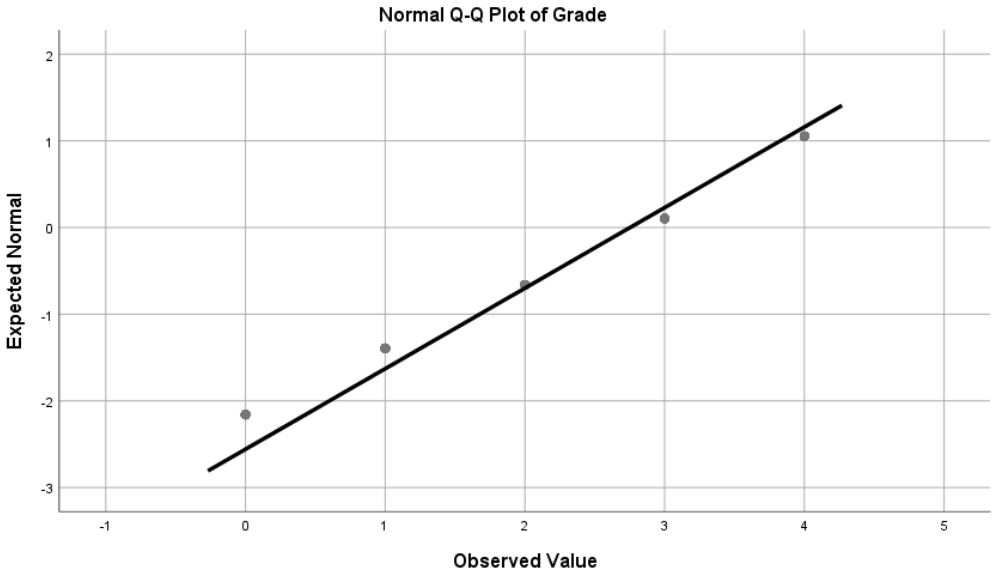

Figure 4: Normal Q to Q Plot of all Teacher-Assigned Grades.

The descriptive data in Table 3 also shows the variables are normally distributed as the skewness of the AIR test scores (.104 with a standard error of .054) falls the range of 1 and -1, and the Kurtosis (.016 with a standard error of .107) within the parameters of 2 and -2. The distribution of the teacher-assigned grades shows a Skewness of -.579 and a Kurtosis of -.386. Both of these fall within the realm of a normal distribution.

Table 3

Skewness and Kurtosis Chart for all AIR Test Scores and all Teacher-Assigned Grades

Category	Skewness	Std. Error	Kurtosis	Std. Error
AIR test Scores	.104	.054	.016	.107
Teacher Grades	-.579	.054	-.386	.107

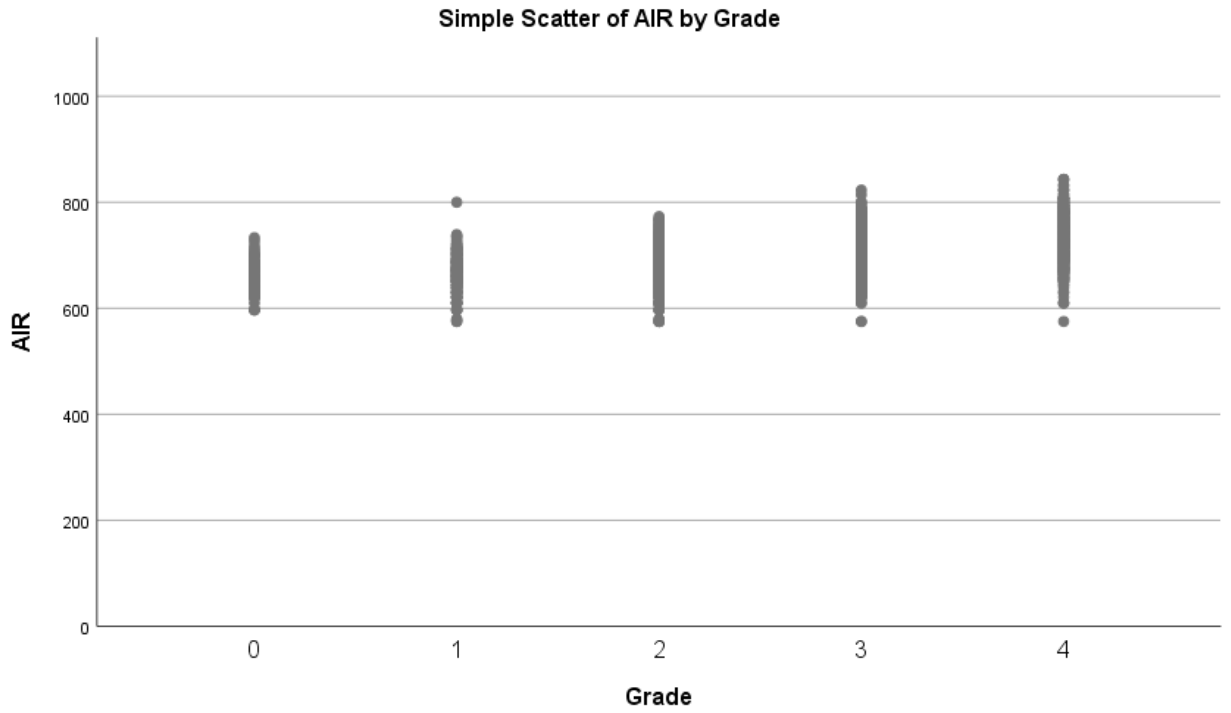

Figure 5: Simple Scatterplot of all AIR Test Scores by all Teacher-Assigned Grades.

The final step was to conduct a linear regression as displayed in *Figure 6* which was scatter plotted showing a cigar shape, making the assumption of homoscedasticity tenable.

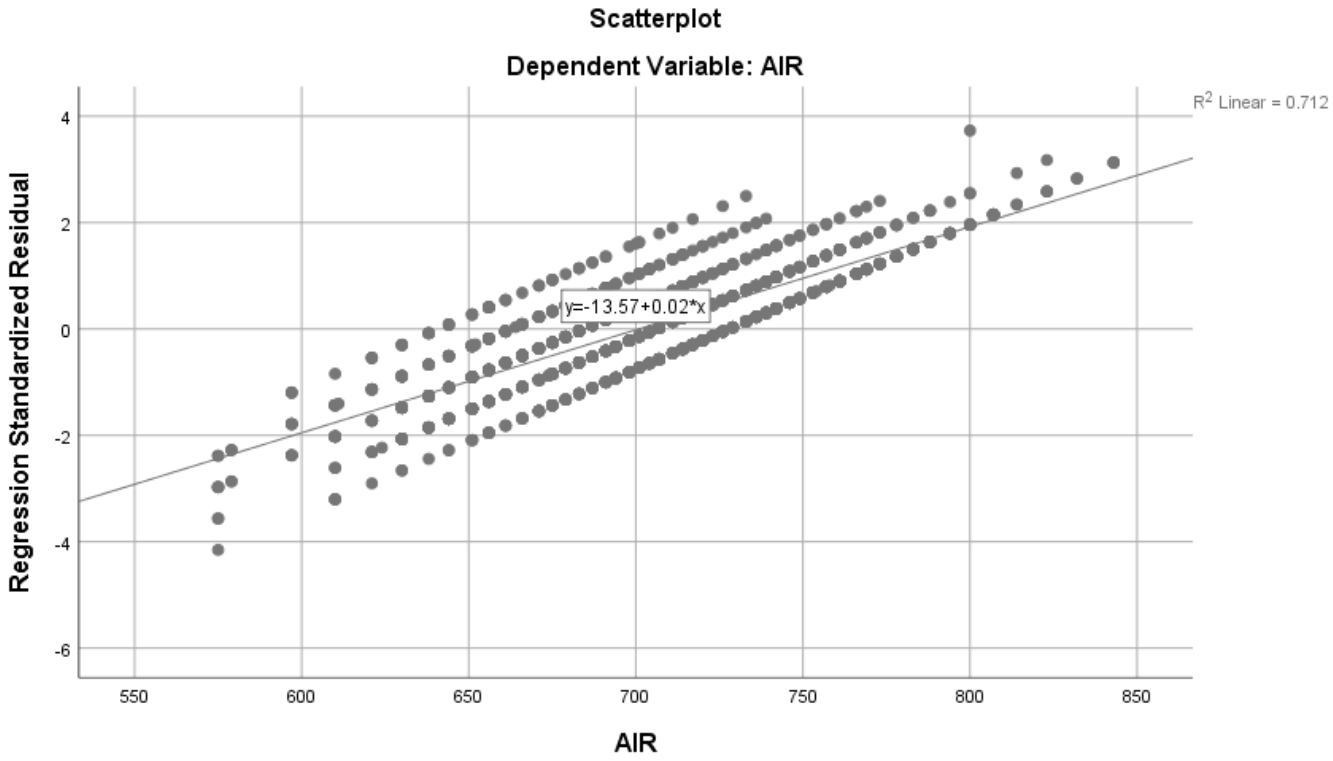

Figure 6: Linear Regression Scatterplot with all AIR Test Scores as Dependent Variable

The normal distribution, the assumption of homoscedasticity, and the skewness and kurtosis numbers of the Pearson Product moment correlation would indicate a statistically significant result that the data is normally distributed and positively correlated so the null hypothesis was rejected. A Pearson-product moment correlation was run with the alpha level set at $p < .01$. As shown in Table 4, the two variables are strongly related, $r = .536$.

Due to the histogram resulting in an assumption of normality that is not tenable, a Spearman's rho test was also conducted. The results displayed in Table 5 confirmed a strong correlation at $r = .557$ with an alpha level set at $p < .01$. The Spearman's rho test was similar to and confirmed the bivariate regression results.

Table 4

Pearson Product-Moment Correlation results between all AIR Test Scores and all Teacher-Assigned Grades

		AIR	Grade
AIR	Pearson Correlation	1	.536**
	Sig. (2-tailed)		.000
	N	2077	2077
Grade	Pearson Correlation	.536**	1
	Sig. (2-tailed)	.000	
	N	2077	2077

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5

Spearman's Rho Correlation results between all AIR Test Scores and all Teacher-Assigned Grades

			AIR	Grade
Spearman's rho	AIR	Correlation Coefficient	1.000	.557**
		Sig. (2-tailed)	.	.000
		N	2077	2077
	Grade	Correlation Coefficient	.557**	1.000
		Sig. (2-tailed)	.000	.
		N	2077	2077

**. Correlation is significant at the 0.01 level (2-tailed).

Null Hypothesis Two

$H_0$2

The second null hypothesis was tested as follows:

$H_0$2: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students with disabilities.

Of the 2077 students in the subject group, 283 were identified as students with disabilities. The mean AIR test score for the students with disabilities group was 663.56 as

shown in Table 6, while Table 1 shows the group without disabilities had a mean score of 706.64.

Table 6

Data for AIR Test Scores for Students with Disabilities

AIR Test Scores	N	M	SD
SWD	283	663.56	35.96

The distribution of the AIR test scores in Table 7 shows a Skewness of .421 with a standard error of .145 which falls within the normal distribution range of 1 and -1, and a Kurtosis of .747 with a standard error of .289 which lies within the normal 2 and -2 range.

Table 7

Skewness and Kurtosis Chart for AIR Test Scores for Students with Disabilities

Category	Skewness	Std. Error	Kurtosis	Std. Error
AIR test Scores	.421	.145	.747	.289

The histogram in *Figure 7* and normal Q-Q plot shown in *Figure 8* affirm the variables are nearly normally distributed.

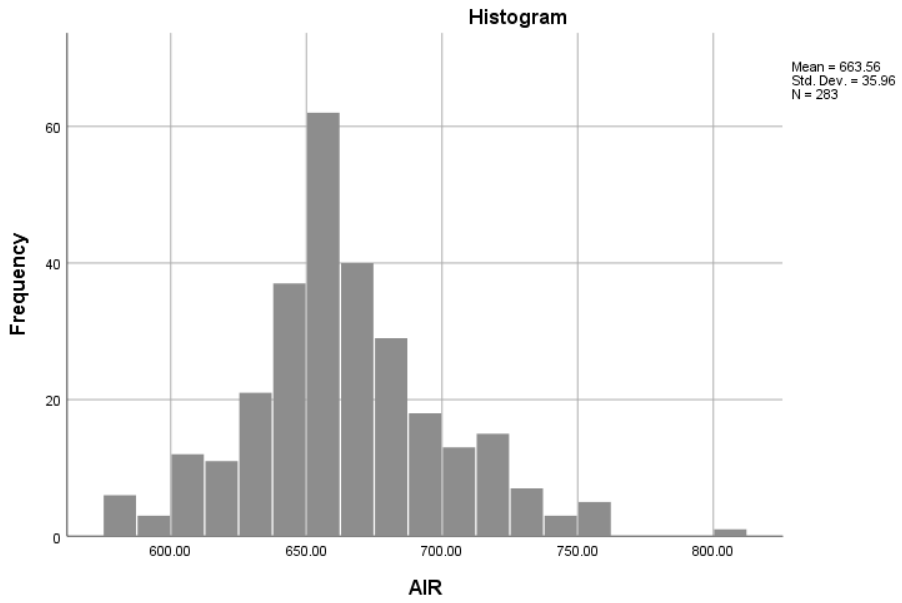

Figure 7: Histogram of AIR Test Scores for Students with Disabilities.

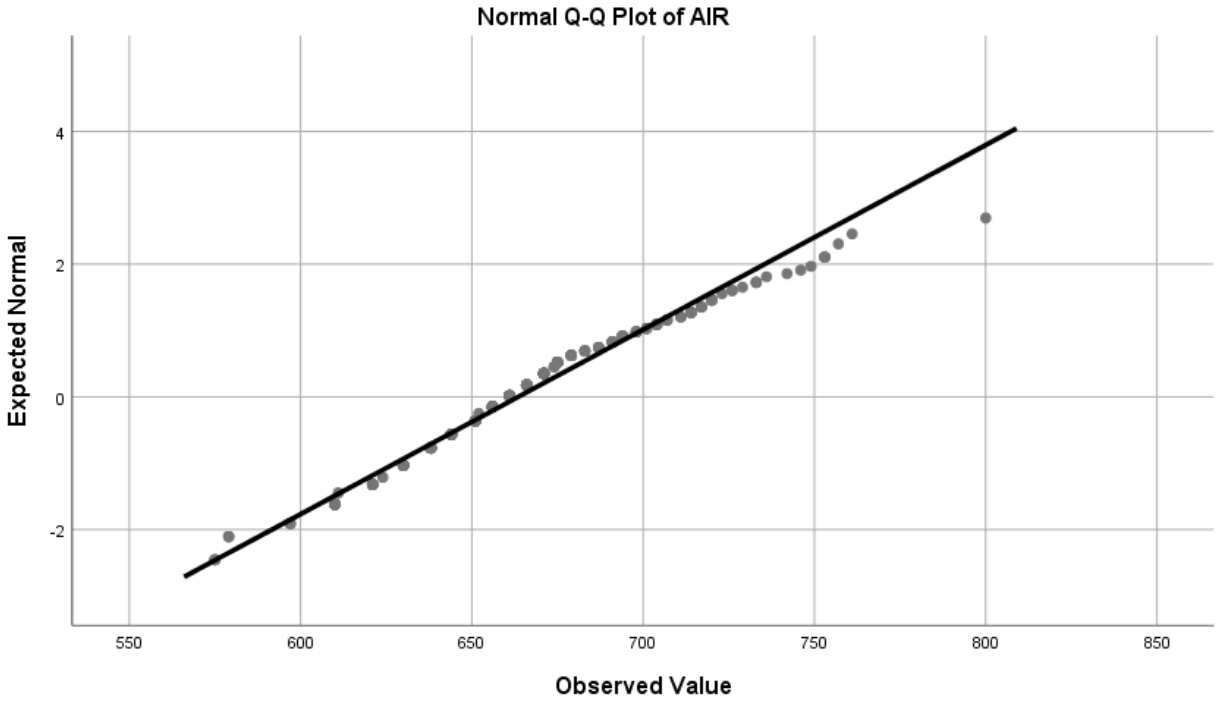

Figure 8: Normal Q to Q Plot of AIR Test Scores for Students with Disabilities.

The teacher assigned grades for the students with disabilities (SWD) group (Table 8) shows a mean grade of 2.1696 as opposed to the 2.75 of the all category and much lower than the students without disabilities (SWoD) group grade mean of 2.845.

Table 8

Data for Teacher-Assigned Grades for Students with Disabilities

Teacher Assigned Grades	N	M	SD
SWD	283	2.1696	1.00329

The distribution of the teacher assigned grades in Table 9 shows a Skewness of -.197 with a standard error of .145 and a Kurtosis of -.270 and a standard error of .289. Both of these fall within the parameters of a normal distribution range of 1and -1 for Skewness and 2 and -2 for Kurtosis.

Table 9

Skewness and Kurtosis Chart of Teacher-Assigned Grades for Students with Disabilities

Category	Skewness	Std. Error	Kurtosis	Std. Error
Teacher Assigned Grades	-.197	.145	-.270	.289

The histogram in *Figure 9* and normal Q-Q plot in *Figure 10* are generally aligned to affirm the variables are normally distributed and the scatterplot in *Figure 11* is generally linear making the assumption of normal distribution tenable.

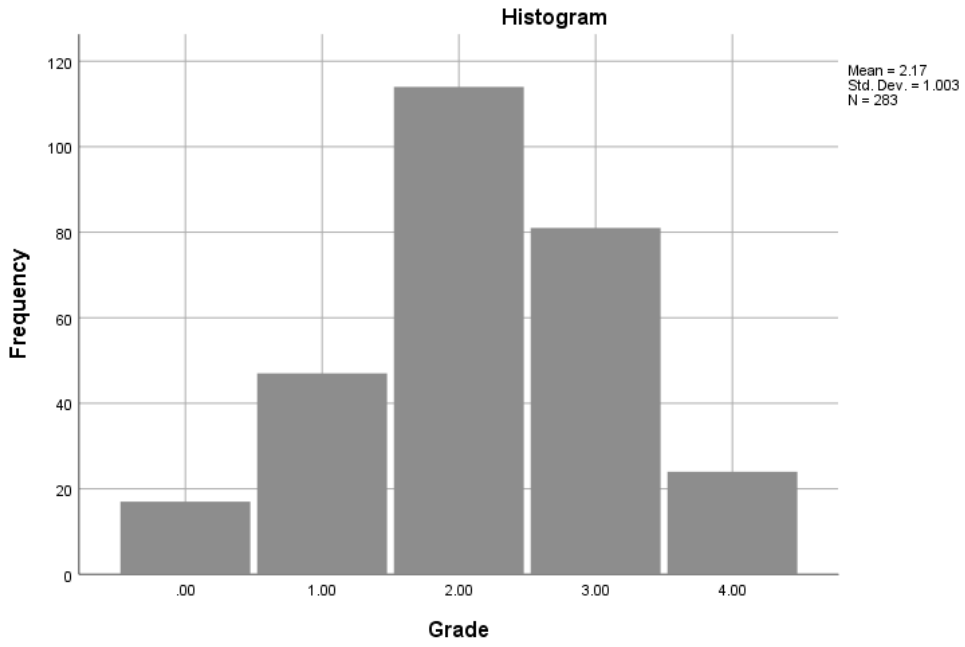

Figure 9: Histogram of Teacher-Assigned Grades for Students with Disabilities.

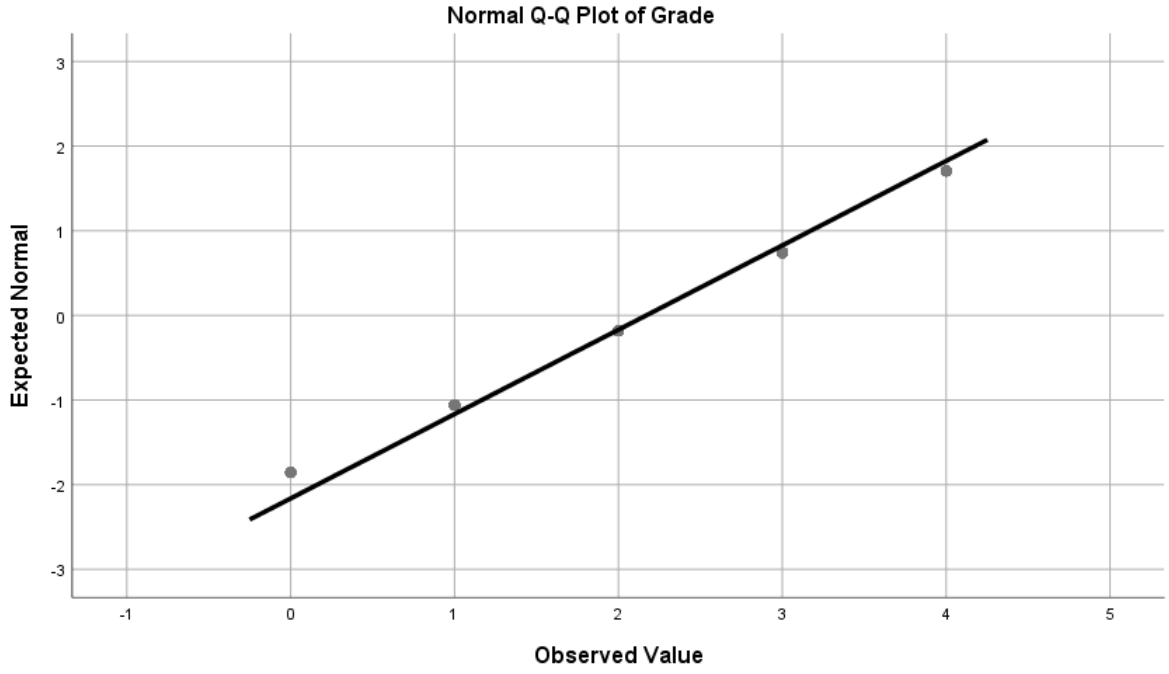

Figure 10: Normal Q to Q Plot of Teacher-Assigned Grades for Students with Disabilities.

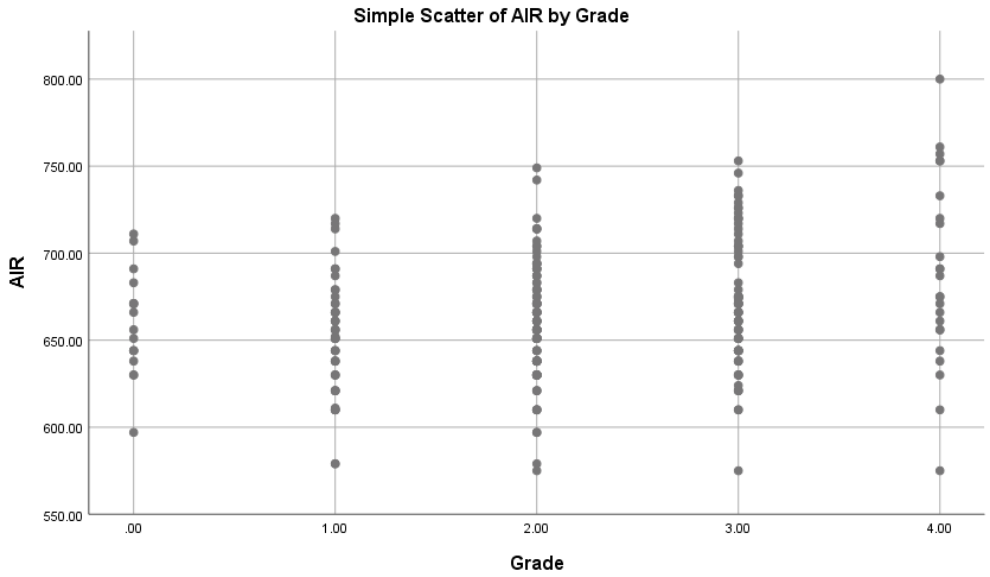

Figure 11: Simple Scatterplot of AIR Test Scores by Teacher-Assigned Grades for Students with Disabilities.

Lastly, a linear regression analysis was conducted and scatter plotted showing that homoscedasticity in tenable with a cigar shaped distribution as shown in *Figure 12*.

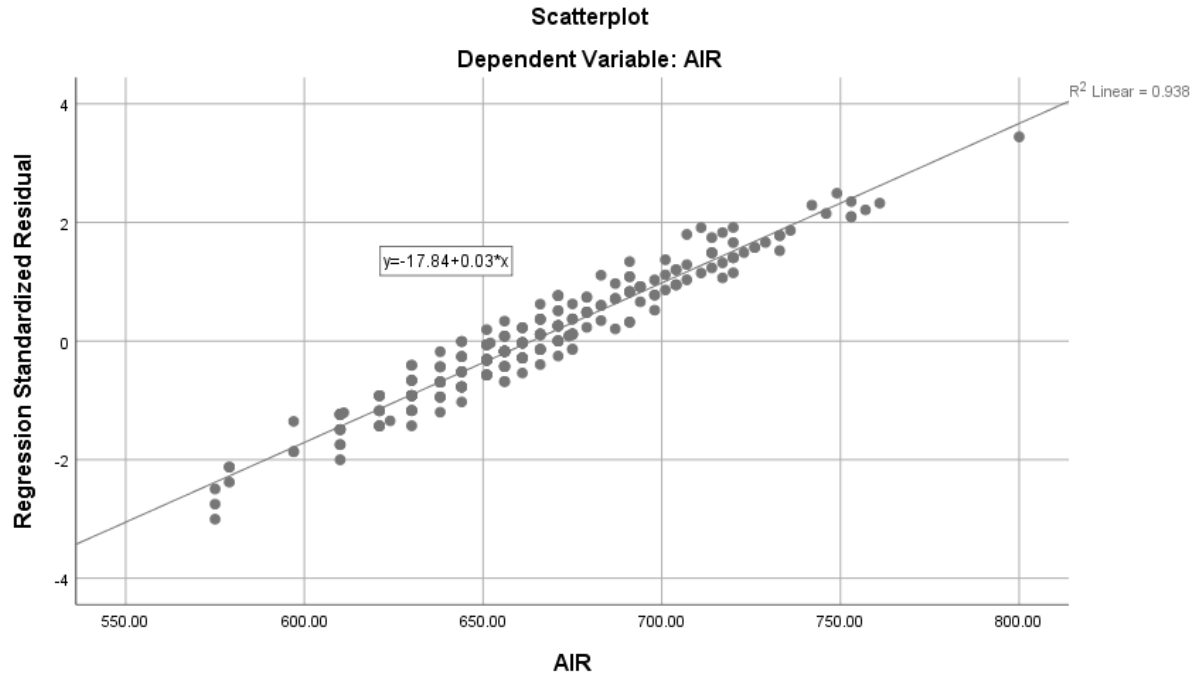

Figure 12: Linear Regression Scatterplot of AIR Test Scores for Students with Disabilities as the Dependent Variable.

Table 10 shows the two variables have a small positive correlation, r (281) = .248, p = .000. As students' classroom grade went up (M = 2.17, SD = 1.00329) their AIR scores also increased (M = 663.5583, SD = 35.96, N = 283). Null hypothesis 2 is rejected, there is a significant, positive correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students with disabilities.

Table 10

Pearson Product-Moment Correlation results between AIR Test Scores and Teacher-Assigned Grades for Students with Disabilities

		AIR	Grade
AIR	Pearson Correlation	1	.248**
	Sig. (2-tailed)		.000
	N	283	283
Grade	Pearson Correlation	.248**	1
	Sig. (2-tailed)	.000	
	N	283	283

**. Correlation is significant at the 0.01 level (2-tailed).

Null Hypothesis Three

$H_0$3

The third null hypothesis was tested as follows:

$H_0$3: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students without disabilities.

Table 11

Data for AIR Test Scores for Students without Disabilities

AIR Test Scores	N	M	SD
Students Without Disabilities	1794	706.64	41.76

Additionally, the Skewness for the students without disabilities for the AIR test score distribution group was .094 with a standard error of .058 and Kurtosis was .150 with a standard error of .116 as shown in Table 12. Since the Skewness was between 1 and -1 there is no perceived problem with skewness, and likewise there is also no problem with kurtosis as it fell between 2 and -2.

Table 12

Skewness and Kurtosis Chart of AIR Test Scores for Students without Disabilities

Category	Skewness	Std. Error	Kurtosis	Std. Error
AIR test Scores	.094	.058	.150	.116

Also, the histogram in *Figure 13* and normal Q-Q plot in *Figure 14* affirm the variables are nearly normally distributed.

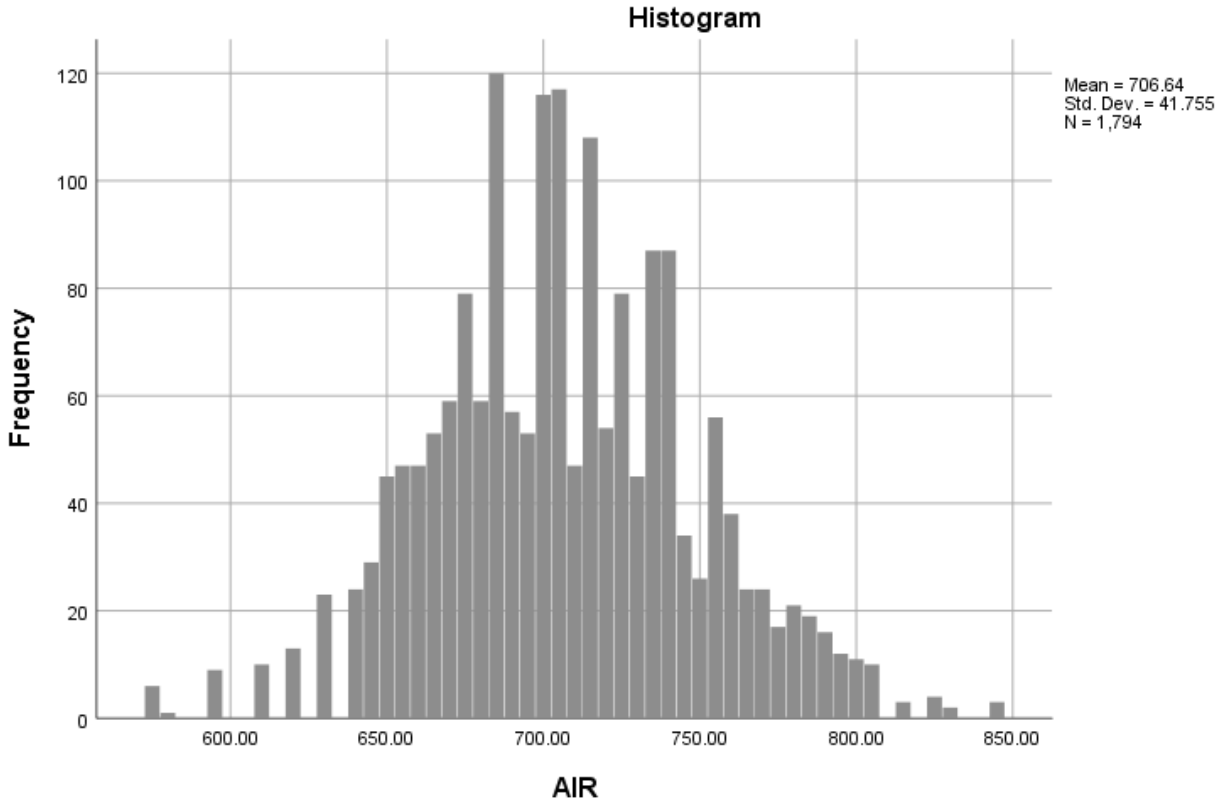

Figure 13: Histogram of AIR Test Scores for Students without Disabilities.

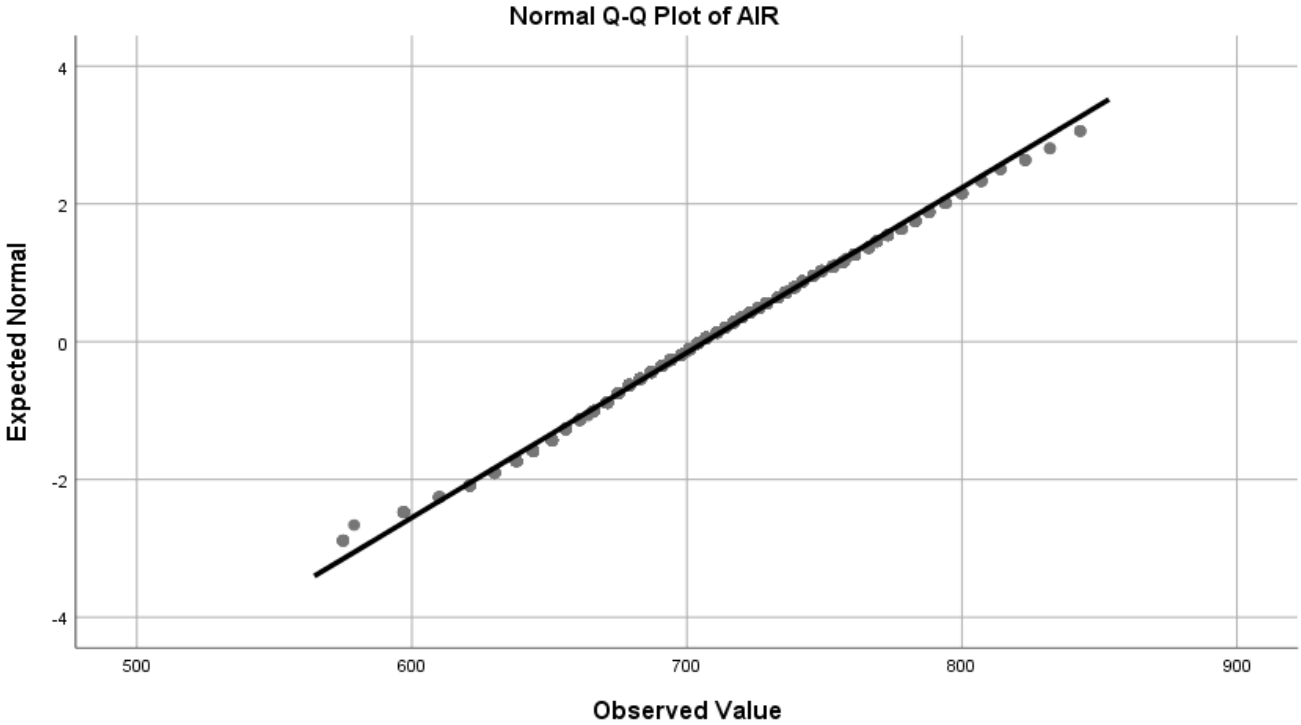

Figure 14: Normal Q to Q Plot of AIR Test Scores for Students without Disabilities.

The Students without Disabilities group (Table 13) had the highest mean teacher assigned grades of the groups tested. They had a mean grade of 2.845 which was higher than the all group with 2.75.

Table 13

Data for Teacher-Assigned Grades for Students without Disabilities

Teacher Assigned Grades	N	M	SD
SWoD	1794	2.845	1.0585

The distribution of the teacher assigned grades in Table 14 shows a Skewness of -.680 with a standard error of .058 and a Kurtosis of -.249 with a standard error of .116. Both of these fall within the range that would indicate a normal distribution between 1and -1 for Skewness and 2 and -2 for Kurtosis.

Table 14

Skewness and Kurtosis Chart of Teacher-Assigned Grades for Students without Disabilities

Category	Skewness	Std. Error	Kurtosis	Std. Error
Teacher Assigned Grades	-.680	.058	-.249	.116

The histogram in *Figure 15* and Normal Q to Q Plot of the Teacher Assigned Grades in *Figure 16*, however, do not appear to be normally distributed and thus the assumption of normality is not tenable. Therefore a Spearman's Rho test was conducted as shown in Table 34.

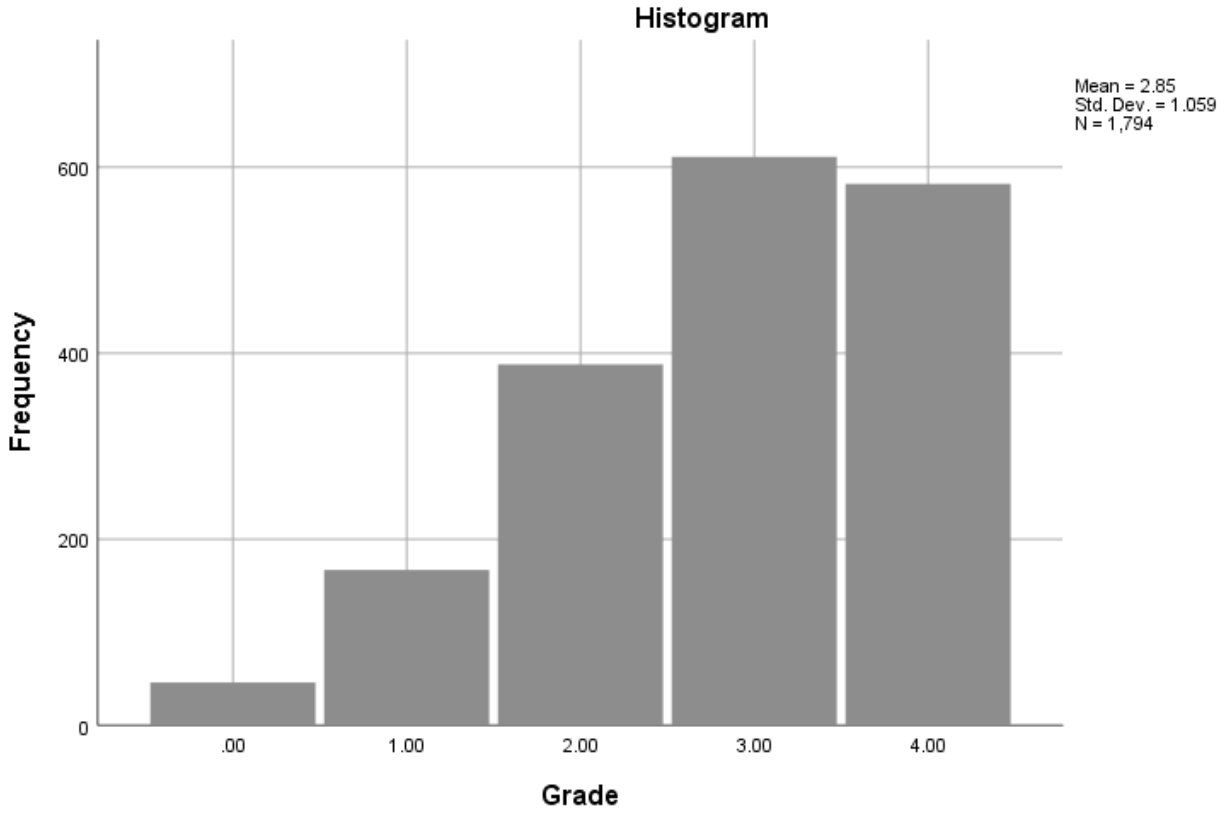

Figure 15: Histogram of Teacher-Assigned Grades for Students without Disabilities.

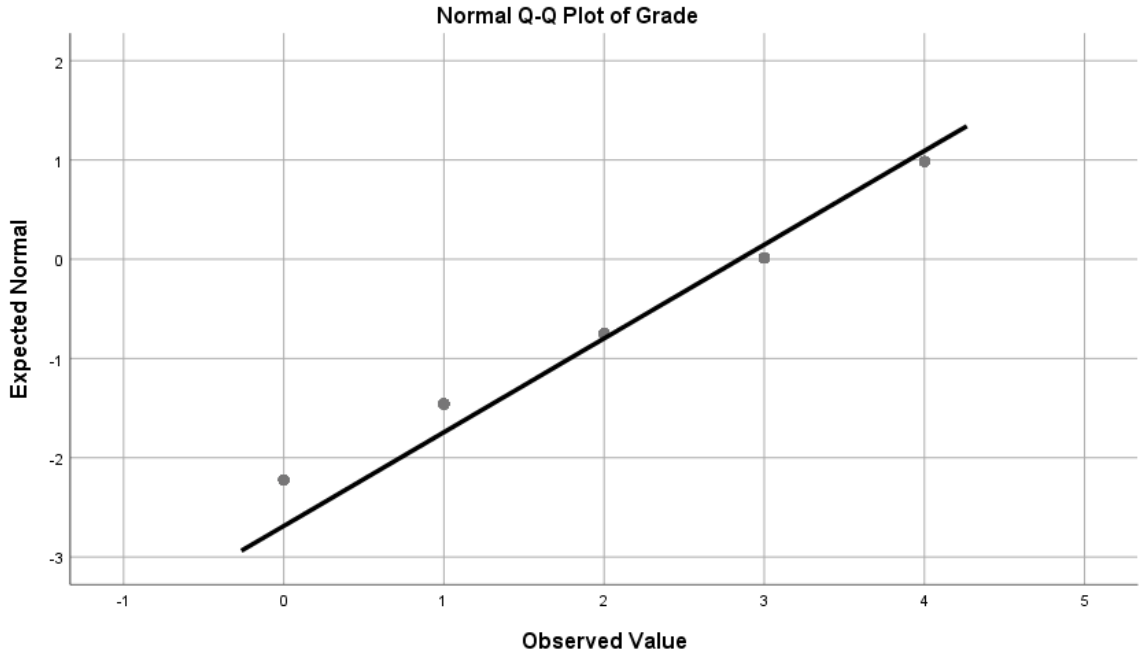

Figure 16: Normal Q to Q Plot of Teacher-Assigned Grades for Students without Disabilities.

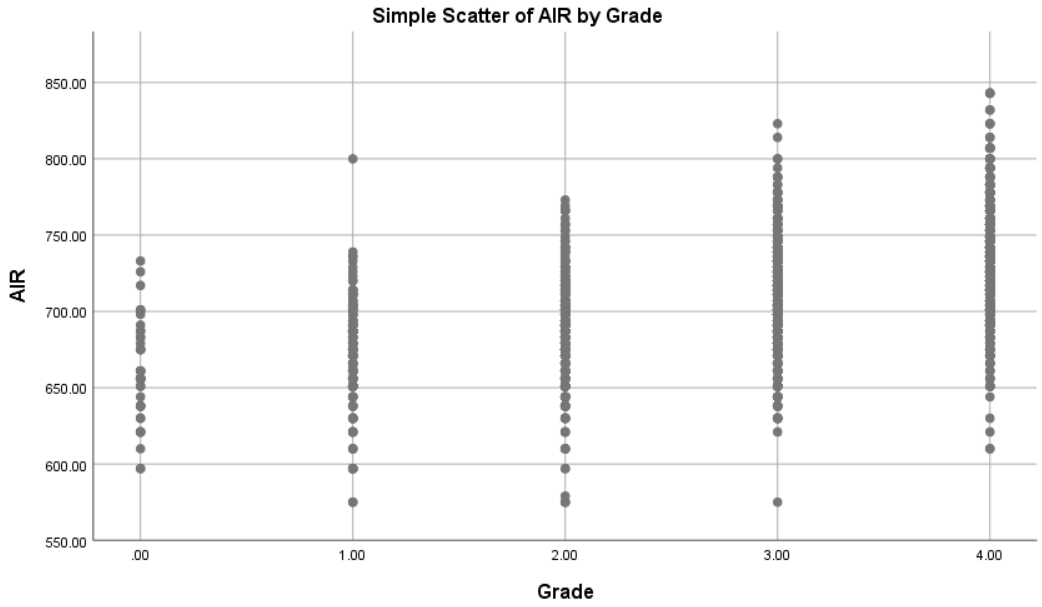

Figure 17: Simple Scatterplot of AIR Test Scores by Teacher-Assigned Grades for Students without Disabilities.

Finally, a linear regression analysis was conducted and plotted in *Figure 18* which aligned in a cigar shape making the assumption of homoscedasticity tenable.

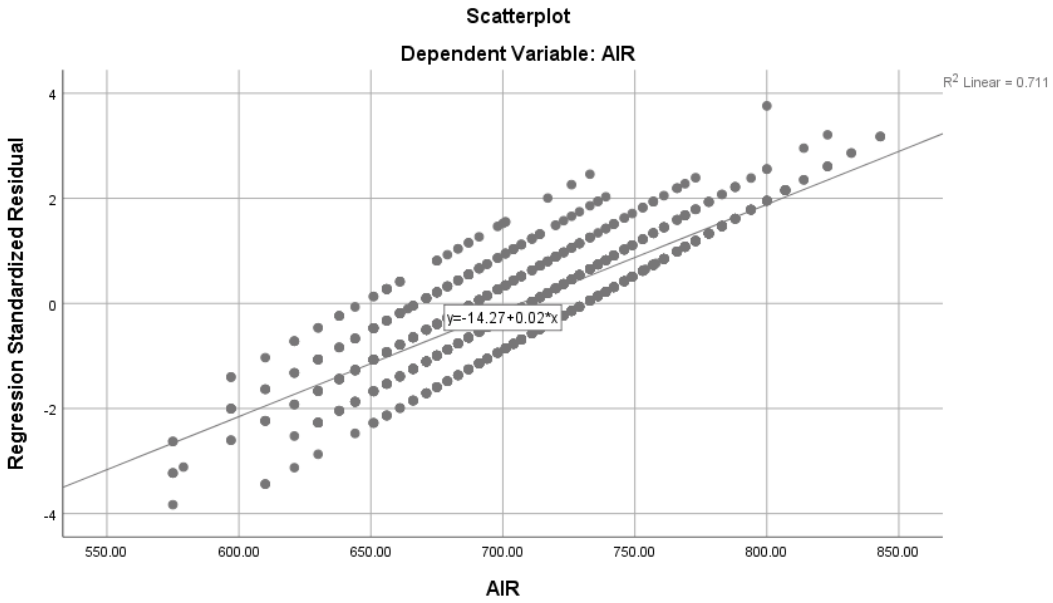

Figure 18: Linear Regression Scatterplot with AIR Test Scores for Students without Disabilities as the Dependent Variable.

Table 15 shows the two variables have a strong, positive correlation, r (1792) = .537, p = .000. As students' classroom grade went up (M = 2.8450 SD = 1.0585) their AIR scores also increased (M = 706.645, SD = 41.75503, N = 1794). Null hypothesis 3 is rejected. There was a significant, positive correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for students without disabilities.

Due to the histogram of Teacher Assigned Grades resulting in an assumption of normality that is not tenable, a Spearman's rho test (Table 16) was conducted. It confirmed a strong correlation at $r = .550$ with an alpha level set at $p < .01$. The Spearman's rho results are similar to and confirm the bivariate correlation results.

Table 15

Pearson Product-Moment Correlation results between AIR Test Scores and Teacher-Assigned Grades for Students without Disabilities

		AIR	Grade
AIR	Pearson Correlation	1	.537**
	Sig. (2-tailed)		.000
	N	1794	1794
Grade	Pearson Correlation	.537**	1
	Sig. (2-tailed)	.000	
	N	1794	1794

**. Correlation is significant at the 0.01 level (2-tailed).

Table 16

Spearman's Rho Correlation results between AIR Test Scores and Teacher-Assigned Grades for Students without Disabilities

			Grade	AIR
Spearman's rho	Grade	Correlation Coefficient	1.000	.550**
		Sig. (2-tailed)	.	.000
		N	1794	1794
	AIR	Correlation Coefficient	.550**	1.000
		Sig. (2-tailed)	.000	.
		N	1794	1794

**. Correlation is significant at the 0.01 level (2-tailed).

Null Hypothesis Four

$H_0$4

The fourth null hypothesis was tested as follows:

$H_0$4: There is no statistically significant correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for disadvantaged students receiving free or reduced lunch (economically disadvantaged).

The disadvantaged group consists of 1318 students who earned a mean score of 686.03 with a standard deviation of 38.61. This compares with the group of all students consisting of 2077 students who earned a mean score of 700.77 with a standard deviation of 43.59 as shown in Table 1.

Table 17

Data for AIR Test Scores for Economically Disadvantaged Students

AIR Test Scores	N	M	SD
Disadvantaged	1318	686.03	38.61

The Skewness for the students from a low socioeconomic background for the AIR test score distribution was .056 with a standard error of .067 and Kurtosis was .201 with a standard error of .135 as shown in Table 18. Since the Skewness was between 1 and -1 there is no apparent problem with Skewness, and likewise there is no perceived problem with Kurtosis as it fell between 2 and -2.

Table 18

Skewness and Kurtosis Chart of AIR Test Scores for Economically Disadvantaged Students

Category	Skewness	Std. Error	Kurtosis	Std. Error
AIR test Scores	.056	.067	.201	.135

Also, the histogram in *Figure 19* and normal Q-Q plot in *Figure 20* are aligned to show the distribution of the variables are nearly normal.

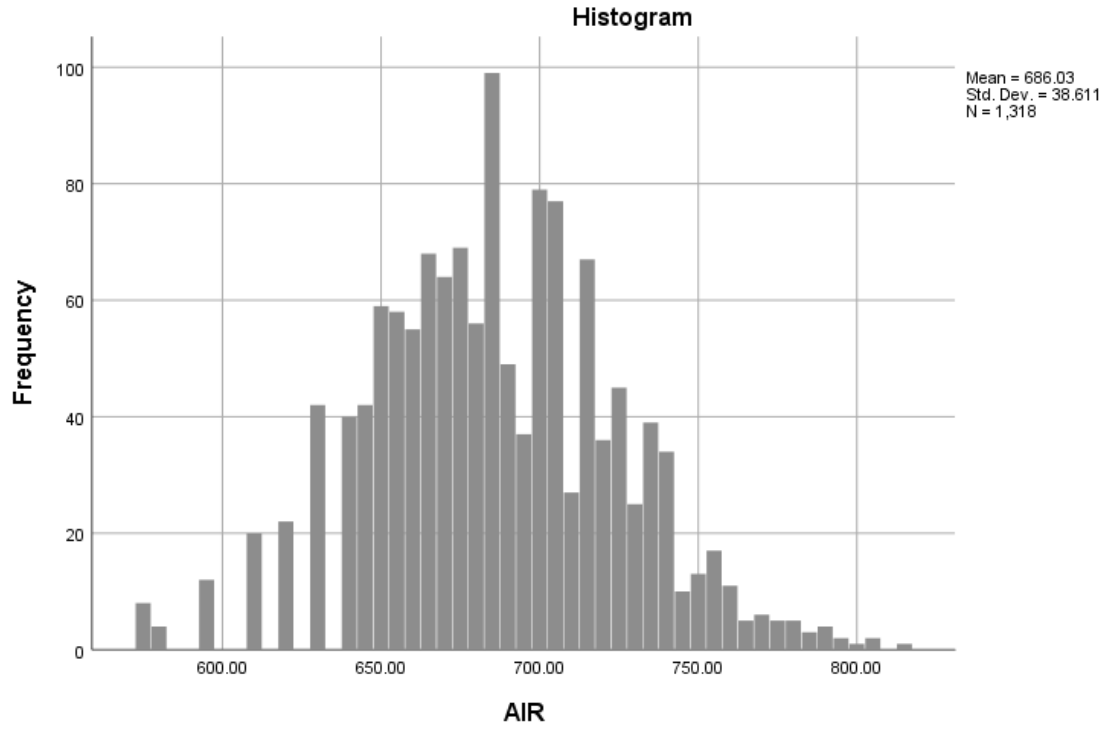

Figure 19: Histogram of AIR Test Scores for Economically Disadvantaged Students.

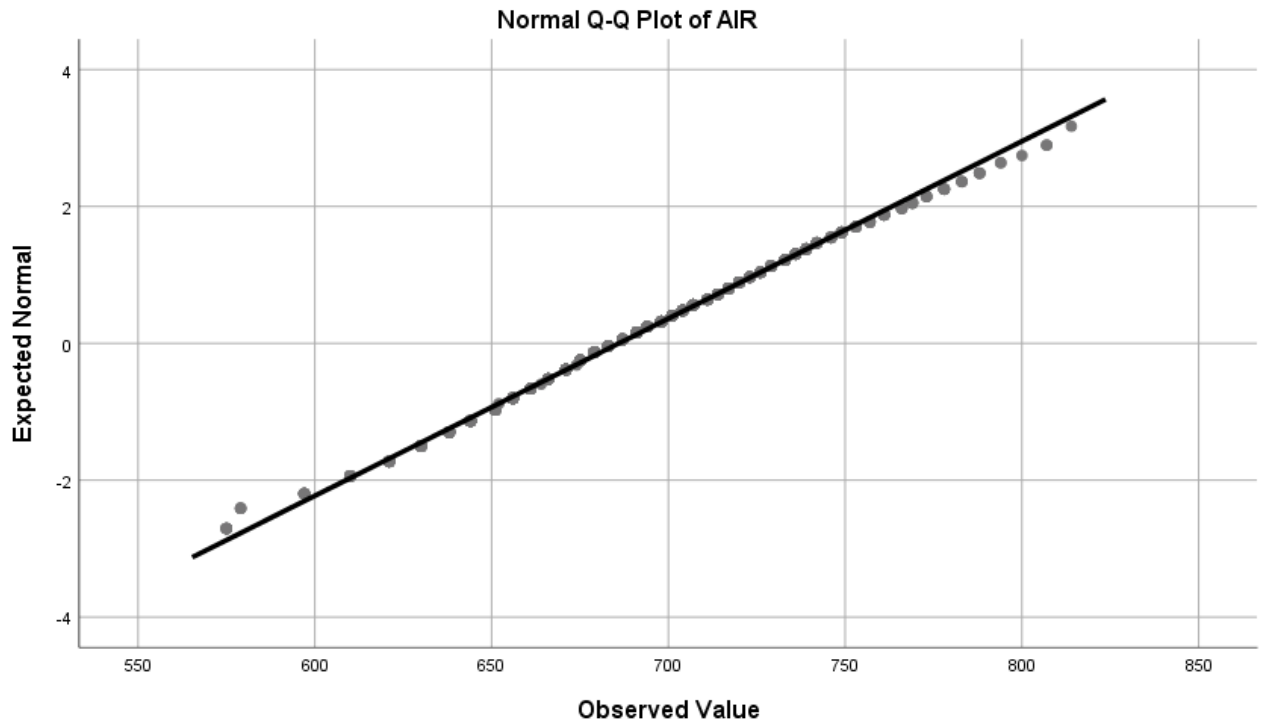

Figure 20: Normal Q to Q Plot for AIR Test Scores for Economically Disadvantaged Students.

The students in the disadvantaged group (Table 19) had a mean grade of 2.5152. Consistent with the literature, this is lower than the all category grade of 2.75 and lower than the students without disabilities who scored a 2.845, but it is higher than the students with disabilities group mean grade of 2.1696.

Table 19

Data for Teacher-Assigned Grades for Economically Disadvantaged Students

Teacher Assigned Grades	**N**	**M**	**SD**
Disadvantaged	1318	2.5152	1.08668

The distribution of the teacher assigned grades in Table 20 shows a Skewness of -.386 with a standard error of .067 and a Kurtosis of -.544 with a standard error of .135. Both of these lie within the range of a normal distribution of 1and -1 for Skewness and 2 and -2 for Kurtosis.

Table 20

Skewness and Kurtosis Chart of Teacher-Assigned Grades for Economically Disadvantaged Students

Category	Skewness	Std. Error	Kurtosis	Std. Error
Teacher Assigned Grades	-.386	.067	-.544	.135

The histogram in *Figure 21* and normal Q-Q plot in *Figure 22* are generally aligned to affirm the variables are nearly normal in distribution and the scatterplot in *Figure 23* shows basic linearity as well.

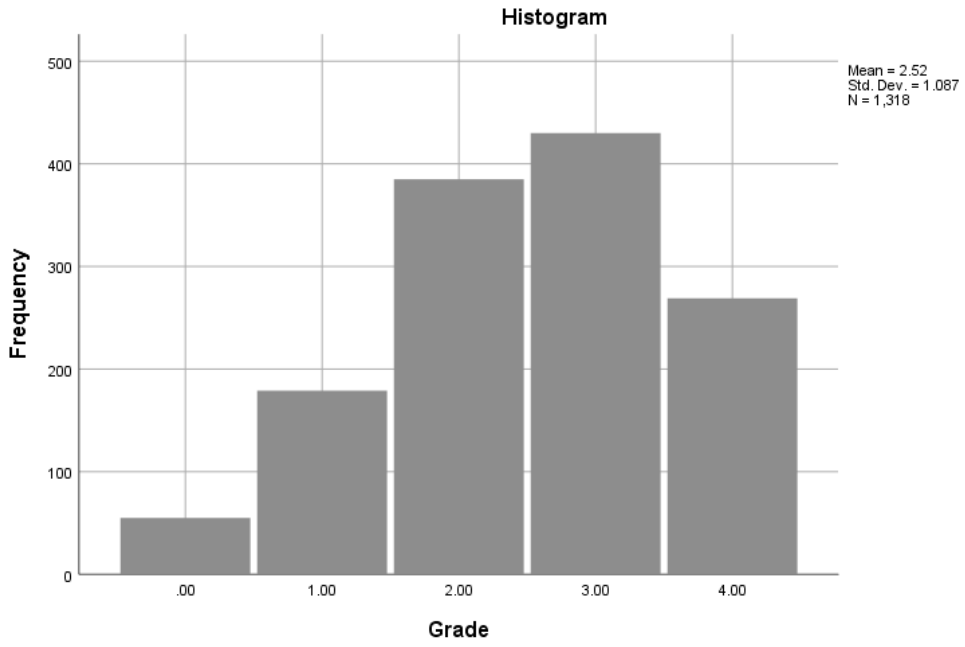

Figure 21: Histogram of Teacher-Assigned Grades for Economically Disadvantaged Students.

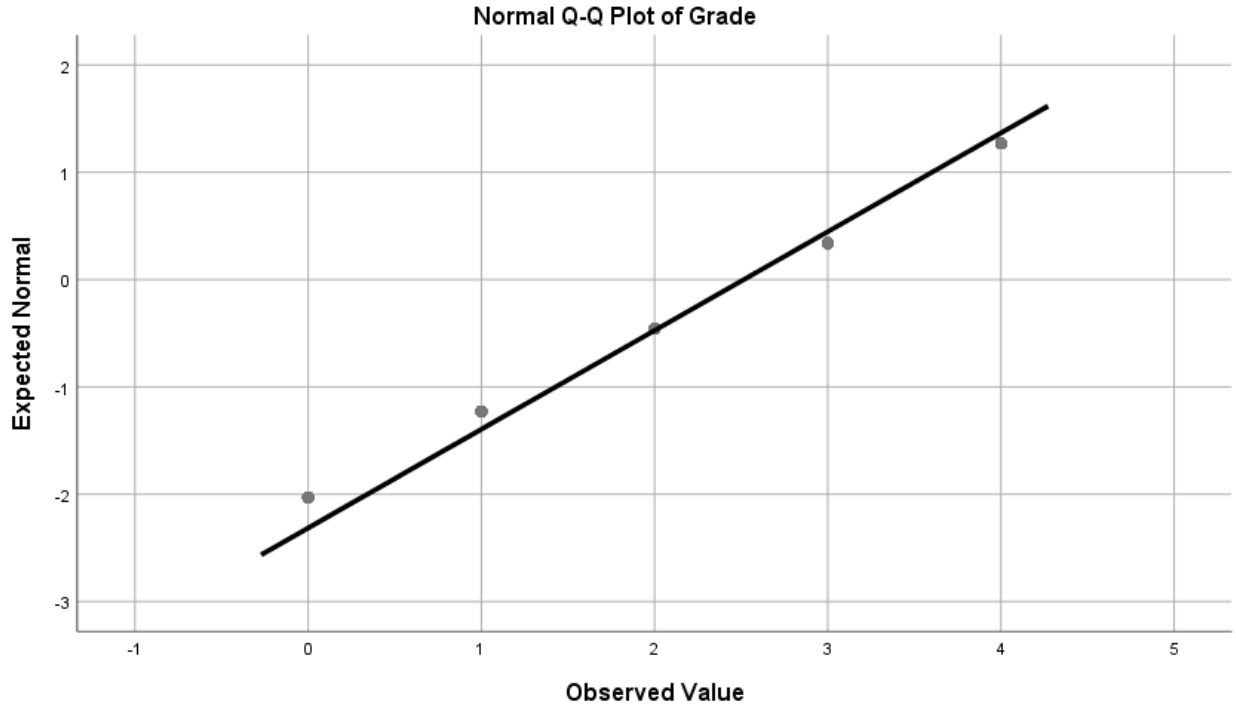

Figure 22: Normal Q to Q Plot of Teacher-Assigned Grades for Economically Disadvantaged Students.

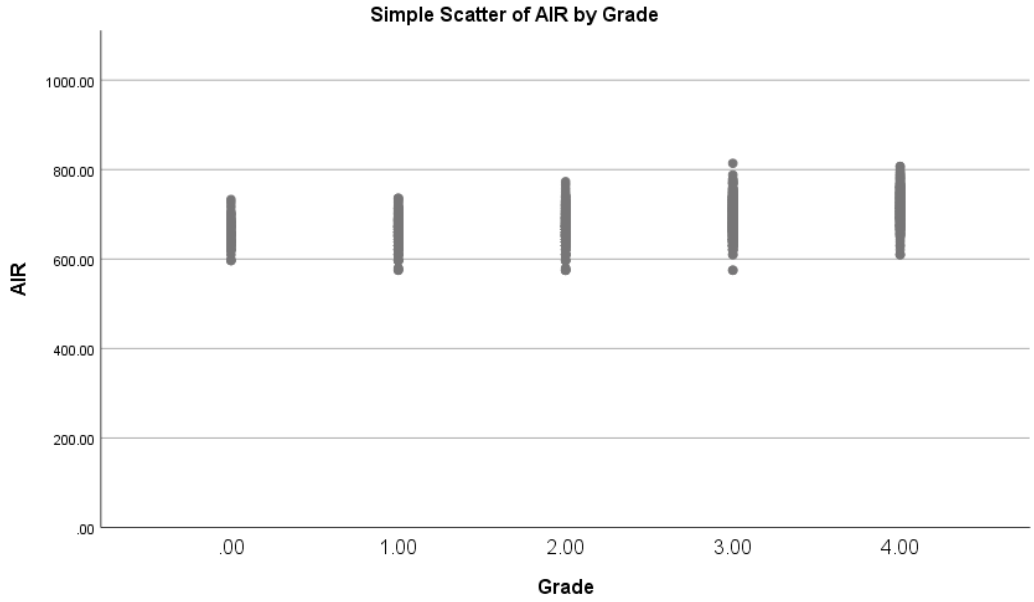

Figure 23: Simple Scatterplot of AIR Test Scores by Teacher-Assigned Grades for Economically Disadvantaged Students.

Finally, a linear regression was run and scatter plotted that did align in a cigar shape again making the assumption of homoscedasticity tenable as shown in *Figure 24*.

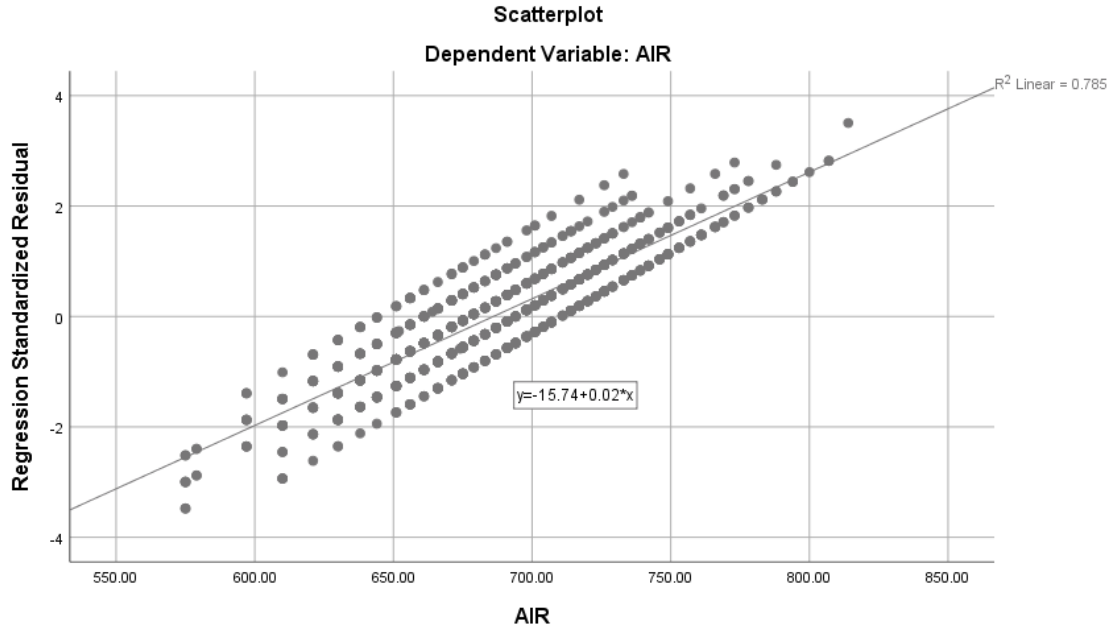

Figure 24: Linear Regression Scatterplot with AIR Test Scores for Economically Disadvantaged Students as the Dependent Variable.

The fourth null hypothesis was tested using the Pearson product-moment correlation and a correlation coefficient was computed to assess the relationship between the teacher assigned grades and AIR end of course science assessment with respect to economically disadvantaged students. Table 21 shows the two variables have a medium strength, positive correlation, r (1318) = .463, p = .000. As students' classroom grade went up (*M* = 2.5152 SD = 1.08668) their AIR scores also increased (*M* = 686.03, SD = 38.61, N = 1318). Null hypothesis 4 is rejected. There is a significant, positive correlation between teacher-assigned course grades for 8th grade Science for the 2014-2015 school year and the scores of the Ohio AIR End-of-Course 8th grade Science test for the 2014-2015 school year for economically disadvantaged students.

Table 21

Pearson Product-Moment Correlation Results between AIR Test Scores and Teacher-Assigned Grades for Economically Disadvantaged Students

		AIR	Grade
AIR	Pearson Correlation	1	.463**
	Sig. (2-tailed)		.000
	N	1318	1318
Grade	Pearson Correlation	.463**	1
	Sig. (2-tailed)	.000	
	N	1318	1318

**. Correlation is significant at the 0.01 level (2-tailed).

Overall, the distributions were checked for normality using histograms and normal Q to Q plots. Separate histograms and normal Q to Q plots were used for the AIR test results and for the teacher assigned final course grades. Each of these two variables were observed independently of the other. In two instances when the histogram portrayed a distribution that did not appear to be tenable, a Spearman's rho test was conducted to confirm the correlation.

Additionally, a scatterplot was used to look at the linearity of the distributions. Finally, a linear regression analysis was conducted to ensure a normal, linear distribution prior to testing with the Pearson product-moment correlation. A separate Pearson test was conducted for each of the groups including all of the students, the students with disabilities, students without disabilities, and disadvantaged students.

Summary

In conclusion, the distributions for the AIR test scores and teacher assigned grades all showed to be normal distributions for two of the four null hypotheses with the other two being confirmed using a Spearman's rho. The correlation coefficients for all for groups showed varying degrees of positive correlations. The strongest correlations according to the Pearson Product Moment Correlation were for the Students without Disabilities (.537) and the all students group (.536), while students from a low socio-economic background correlated medium strength (.463) and the weakest, although positively correlated, group was the Students with Disabilities (.248).

CHAPTER FIVE: CONCLUSIONS

Overview

This study investigated the correlation between teacher-assigned grades and test scores on an end-of-course Science assessment. Previous studies have shown a dubious correlation, especially with students with disabilities (Yell et al., 2012; Wormeli, 2006) and students that come from a low socio-economic background (Heppt et al., 2014; Nichols & Valenzuela, 2013; Sackett et al., 2012; Zwick & Grief-Green, 2007; Zwick & Himelfarb, 2011). The results of this study generally aligned with research that preceded it. The implications of this study will hopefully spark discussion about grading systems, preparation for standardized tests, finding better ways educate students living in poverty, more successfully teaching students with disabilities, and perhaps discussion on the culture of testing in general.

Discussion

The purpose of this quantitative, correlational study was to find the relationship between teacher-assigned grades and end-of-course assessment scores while taking into account the populations of students with low SES and students with disabilities. In analyzing the effects of an accountability movement that saw NCLB bring high stakes testing to the forefront and more rigorous standards within states such as the Common Core and, in Ohio, the new Ohio Learning Standards, there is evidence of elements of behaviorism, motivational theory, and constructivism undergirding the changes that have been made in American education over the past two decades. The driving force behind the reform however was not espousing an educational theory but rather a political movement that was in response to a call for increased accountability for schools by the public.

Within the literature, studies were cited that indicate SES is a stronger factor than grades on a standardized test (Grief Green, 2007), that grades better predict success in college than SAT scores (Hiss & Franks, 2014), and that grades are invalid and unreliable that two studies show large groups of teachers grading the same papers and the grades running the full gamut from A to F (Stark & Elliott, 1912; Brimi, 2011), and that the leaders in educational measurement find teacher-assigned grades to be tremendously flawed (Guskey, 2015; O'Connor, 2017; Reeves, 2016). Kettler et al. (2013) found that the best predictor of a future assessment score to be the previous test score of that particular student. With all that discrepant data, let examine the results of this study.

The first research question asked if there is a relationship between teacher-assigned grades in 8th grade Science and the Ohio AIR End-of-Course 8th grade Science test during the 2014-2015 school year. The results indicated that there was a strong positive relationship between the two sets of scores. Although teacher-assigned grades and AIR test scores had a strong relationship at .536, the strength of the relationship is not perfect. This aligns with the findings that speak to the lack of validity and replicability due to a multitude of factors going into teacher assigned grades as opposed to the test which is more a straight test of academic knowledge or intelligence (Guskey, 2013, 2015; Marzano, 2000; O'Connor, 2011, 2013, 2017; Reeves, 2016). Although it is intuitive that students who get good grades would do well on tests and students who perform poorly in school as measured by grades would do likewise on tests, the strength of the relationship is still less than perfect which lends credence to those who have been critical of the lack of validity and reliability of grades (Brookhart et al., 2016; Cornue, 2018; Marzano & Heflebower, 2011; O'Connor, 2017; Reeves, 2016).

The second research question asked if a relationship exists between teacher assigned course grades in 8th grade Science and the Ohio AIR End-of-Course test scores for students with disabilities. The results showed a small, but positive correlation. Students with disabilities are often held to lesser standards and thus an A grade in a classroom with a modified curriculum would likely yield a lesser score on a standardized end of course exam than a student that completed the full curriculum that is being tested (Jung & Guskey, 2012). Lack of motivation, lack of confidence, and a history of grade inflation or social promotion in spite of unsuccessful academic endeavors is also likely a factor (Guskey, 2015; Vatterott, 2015; Wormeli, 2006). The literature also points to a lack of appropriate accommodations as negatively affecting the results of students with disabilities (Jamgochian & Ketterlin-Geller, 2015; Jung & Guskey, 2012; Lai & Berkeley, 2012; Powell, 2012). Additionally, in Ohio, students with disabilities may be excused from the consequences of the end of course exams if decided by the IEP team. Students who have been made aware of the fact that their scores really do not matter as far as effecting their graduation or promotion may have significantly less motivation to push themselves in test preparation or performance than students who may be motivated by grades or test scores (Dueck, 2014; Guskey, 2015; Vatterott, 2015). Jensen (2009) asserts that students lack motivation if they doubt their ability or feel that failure is imminent. This would likely hold true for students for whom the test material is difficult coupled with the knowledge that their results will not matter. The high-stakes nature of the test is simply not present for many students that fall in this category. It also follows that the grades for Students without Disabilities had the strongest correlation to the AIR test scores since students with disabilities had the lowest correlation.

The final research question asked if a relationship exists between the Ohio AIR End-of-Course Science test and teacher assigned final course grades in Science for students with lower

socioeconomic status. There was again a positive correlation of medium strength at .463. This was a stronger correlation than expected. Zwick and Grief-Green (2007) found a higher correlation between SAT scores and socioeconomic status (SES) than between SAT scores and grades. In summary, high income is associated with high SAT score, low income is associated with low SAT score which had a stronger correlation than high teacher assigned grades being associated with high SAT scores, and low grades being associated with low SAT scores. That particular comparison was not made in this study, but it is a powerful testament to both the lack of strength of the correlation between grades and a standardized test and the impact of low socio-economic status on students. In this study, the score was likely closer to the all-students group because the group of disadvantaged students in this study was exceptionally high. Not only were there three urban schools involved in the study where nearly all of the students qualified for free lunch, the other schools had significant numbers of students qualifying for free lunch as well.

The literature indicates that grades and test scores are measuring two different but important things even though there is a correlation showing that the higher a student's grades are, there is often a corresponding higher test score (Laurie, 2009). Grades were largely disparaged in the literature, but grades seem to have value as a strong predictor of school success (Bowers, 2009) to the extent that students with low grade point averages and high test scores are shown to be less successful in college than students with high grades and lower test scores (Hiss & Franks, 2014). So even though teacher assigned grades may be a less than perfect correlation to standardized test scores, they clearly have a value in predicting student success in school.

Implications

Even with its limitations, the implications of this study are important for current educators in that they show that although there is a positive correlation between the teacher-

assigned grades and the Ohio AIR End-of-Course exams, the correlation becomes weaker for disadvantaged students and even more so for students with disabilities. Typically, the students with higher grades also scored better than their peers with lower grades. The subjectivity of teacher-assigned grades was most evident with the groups with lower correlation coefficients – namely the students from lower socioeconomic backgrounds and students with disabilities. The results of this study should spark conversation about grading systems, and how to better prepare some of our most at-risk students for academic success. Additionally, questions regarding the equity of testing need to be addressed as gender, ethnicity, English proficiency, disability status, and socio-economic status all are shown to significantly impact test scores (Ricketts, 2010), and lower income schools are those most negatively affected by high stakes testing (Huddleston, 2014).

Limitations

There are several limitations to this study that could prompt further study and future research. The first limitation is in the setting. Although a variety of schools were studied from five different counties, all of the participants were in Northeast Ohio. Because of both the area and the school districts studied, there was a higher than normal number of students from lower socio-economic backgrounds that met the criteria for disadvantaged. There also was a large number of students with IEPs that counted towards the Students with Disabilities category. Although the sample was valid, the percentages are higher than the national average and have been shown to significantly affect the outcomes.

The second limitation is that the study was not broken down into more specific groupings. The study looked at students in general, students with disabilities, students without disabilities, and disadvantaged students. The study stopped short of breaking down the

demographics further and looking at African-American students, Hispanic students, gender differences, and students who are English language learners. Also no distinctions were made for potential differences such as students being raised by a single parent, students being raised by a grandparent, students in foster care, or students living in public housing as opposed to a single family home.

The third limitation was that no distinction was made about the teacher that taught the tested class. If the student was being taught by a first year teacher, a veteran master teacher, or if substitute teachers frequented the classroom could significantly affect the academic outcomes of the students. Teacher experience, status, or efficacy however were not considered in this study.

The fourth limitation was in how the grades were recorded. Percentages could show a bit more depth than letter grades although the research also questions the validity of percentage grades, however the data came in letter grade only from eight of the nine districts. Additionally, grading policies of each school district was not considered. For example, one school could have a no-zero policy where a missing assignment count as 50%, another school could have a mastery system where students could correct and re-submit assignments with poor grades, and yet another may punish a late assignment with a zero. All of these factors could affect the letter grade given. Although these schools all graded on a 10 point scale where 90-100% is an A, 80-89% is a B and so forth, grade scales vary from school to school whereas a 94% may be needed for an A at one school making a 93% a B at one school and an A at another. Also, as the research pointed out, some teachers give points for bringing in tissues, participation points, or behavior points all which could affect the grade (O'Connor, 2011; Reeves, 2016). None of those considerations were differentiated in this study.

The fifth limitation was that there was no attempt made to measure student growth. For example, a student who entered the 8th grade class on a third grade level could have made significant gains to a fifth or sixth grade level and still could have conceivably failed the 8th grade exam but received an A grade for their progress. The state of Ohio does use a value-added score to measure student growth, but this was not factored into this study.

The sixth limitation was there was no consideration given to test fatigue or measuring the motivation level of the individual students. The Science End-of-Course exam is given in the last testing window in Ohio. By the time the test is given, the students have been inundated with testing. In addition, the students in eighth grade had not ever been retained or made to fail a course due to their test results, so students who lack intrinsic motivation may have not put forth the preparation or effort that would reflect their ability. Also, this was the first time this test was given meaning the content and questions on the test were unfamiliar to the teachers preparing the students for the test. Teachers were limited in their ability with regards to "teaching to the test" as it was the first time this test had been given.

Lastly, attendance was not looked at in this study. Days of instruction lost or hours of school missed due to arriving tardy plays a large negative role in academic achievement. We could assume that students who missed a greater number of days would not fare as well in either grades or the test scores as they would have had they missed fewer days, but this was not tracked for this study.

Recommendations for Future Research

Additional research needs to be conducted to more fully understand the relationship between teacher-assigned grades and standardized test scores. A nearly endless list of possibilities are available for further study. Additional research related to this study include but is not limited to:

1. Replicating this study in a different geographic area or in other states where the tests differ from those in Ohio.
2. Disaggregating the data further to look at race, ethnicity, or ELL status to compare with the general population of students in the study to look for achievement gaps, cultural fairness, overrepresentation of minorities in the disadvantaged group, and equity in honors classes.
3. Replicating the study using grade percentages to produce richer data and more specific points of comparison.
4. Looking at the data with the added factor of comparing teachers at different stages in their career with the student achievement in both grades and standardized tests.
5. Comparing the data with single parent homes and two parent homes as a predictor variable.
6. Comparing schools that use standard based grades with schools where teachers assigned grades are left up to the conventions of the individual teachers.
7. Examining the grade inflation factor to replicate previous studies that found grade inflation prevalent in urban schools and schools with high poverty rates.
8. Investigating the relationship between teacher-assigned grades and standardized test scores for students receiving their education via an online school.

REFERENCES

Allen, J. (2005). Grades as a valid measure of academic achievement of classroom learning. *The Clearing House*, *78*(5), 218-223.

Bali, V. A., Anagnostopoulos, D., & Roberts, R. (2005). Toward a political explanation of grade retention. *Educational Evaluation and Policy Analysis*, *27*(2), 133-155.

Beck, J. S. & Santoyo, C. J., Using mastery learning to help all students achieve" (2017). *UNLV Best Teaching Practices Expo*. 27. https://digitalscholarship.unlv.edu/btp_expo/27

Borich, G. D., Peck, R. F., & Texas Univ., A. E. (n.d). Relationships between teachers' marks, achievement test scores and aptitude as a function of grade, ethnicity and sex.

Bostic, J., Vostal, B., & Ruffer, B. (2014). Promoting freedom through problem-based learning. *Ohio Social Studies Review*, *51*(1), 73-87.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: more than just academic knowledge. *Journal of Educational Administration*, *47*(5), 609-629. doi:10.1108/09578230910081080

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research & Evaluation*, *17*(3), 141-159. doi:10.1080/13803611.2011.597112

Breiner, J. (2015, March). Is high-stakes testing the answer?. *School Science & Mathematics*. pp. 103-104. doi:10.1111/ssm.12112.

Brimi, H. M. (2011). Reliability of grading high school work in English. *Practical Assessment, Research, and Evaluation*, 16(17), 1-12.

Brookhart, S. M. (2009). *Grading* (2nd ed.). Upper Saddle River, NJ: Pearson.

Brookhart, S. M. (2017). *How to use grading to improve learning*. Alexandria, VA: ASCD.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A century of grading research. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

Burrowes, C. P. (1993). *The Functionalist Tradition and Communication Theory*.

Carey, T., & Carifio, J. (2012). The minimum grading controversy: Results of a quantitative study of seven years of grading data from an urban high school. *Educational Researcher, 41*(6), 201-208. doi:10.3102/0013189X12453309

Clark, J. S. (2013). Educational inequality in Indiana: The impact of socioeconomic status and race on ISTEP+ exam performance. *Journal of the Indiana Academy of the Social Sciences, 16*(1), 106-117.

Cornue, J. (2018). *Changing the Grade: A Step-by-Step Guide to Grading for Student Growth*. ASCD.

Duckworth, A. L., & Seligman, M. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades and achievement test scores. *Journal of Educational Psychology, 98*(1), 198-208. doi:10.1037/0022-0663.98.1.198

Duckworth, A. L., & Seligman, M. E. (2017). The science and practice of self-control. *Perspectives on Psychological Science, 12*(5), 715-718. doi:10.1177/1745691617690880

Duckworth, A. L., Tsukayama, E., & Quinn, P. D. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology, 104*(2), 439-451. doi: 10.1037/a0026280

Dueck, M. (2014). *Grading smarter, not harder: Assessment strategies that motivate kids and help them learn*. Alexandria, VA: Association for Supervision & Curriculum Development.

Edwards, N. T. (2006, March). The historical and social foundations of standardized testing: In search of a balance between learning and evaluation. *JALT Testing & Evaluation SIG Newsletter, 10*(1), 8-16.

Ellison, B. J. (2008). *Does getting A's really matter? A conceptualization of grades as a measure of educational outcomes* (Doctoral dissertation, University of Notre Dame).

Erskine, J. L. (2014). It changes how teachers teach: How testing is corrupting our classrooms and student learning. *Multicultural Education*, *21*(2), 38-40.

Finefter-Rosenbluh, I., & Levinson, M. (2015). What is wrong with grade inflation (if anything)?. *Philosophical Inquiry in Education*, *23*(1), 3-21.

Froman, T., Brown, S., Tirado, A., & Miami-Dade County Public Schools, R. S. (2008). Classroom grades and their relationship to FCAT scores. Research brief. Volume 0705. *Research Services, Miami-Dade County Public Schools*,

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston: Pearson

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, 15 (1), 83-99

Garcia, D. R. (2011). The lingering influence of high-stakes testing polices on the postsecondary prospects of Latino students. *JEP: Journal of Education Policy*, 3-9.

Gencel, I. E., & Saracaloglu, A. S. (2018). The effect of layered curriculum on reflective thinking and on self-directed learning readiness of prospective teachers. *International Journal of Progressive Education, 14*(1), 8-20. doi:10.29329/ijpe.2018.129.2

Godfrey, K. E. (2011). *Investigating Grade Inflation and Non-Equivalence*. (Research Report 2011-2). College Board: New York, NY.

Groen, M. (2012). NCLB – the educational accountability paradigm in historical perspective. *American Educational History Journal, 39*(1/2), 1-14.

Guskey, T. R. (2001). Helping standards make the grade. *Educational Leadership, 59*(1), 20=27.

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

Guskey, T. R. (2010). Lessons of mastery learning. *Educational Leadership, 68*(2), 52.

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

Guskey, T. R. (2015). *On your mark: Challenging the conventions of grading and reporting*. Bloomington, IN: Solution Tree Press.

Guskey, T. R., & Jung, L. A. (2016). Grading: Why you should trust your judgment. *Educational Leadership, 73*(7), 50.

Harvey, K. E., Suizzo, M., & Jackson, K. M. (2015). Predicting the grades of low-income–ethnic-minority students from teacher-student discrepancies in reported motivation. *The Journal of Experimental Education, 84*(3), 510-528. doi:10.1080/00220973.2015.1054332

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

Hazel, C. (2010). Interactions between bullying and high-stakes testing at the elementary school level. *Journal of School Violence*, *9*(4), 339-356. doi:10.1080/15388220.2010.507142

Heppt, B., Haag, N., Bohme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low- SES families. *Reading Research Quarterly*, *50*(1), 61-82. doi:10.1002/rrq.83

Higdem, J. L., Kostal, J. W., Kuncel, N. R., Sackett, P. R., Shen, W., Beatty, A. S., & Kiger, T. B. (2016). The role of socioeconomic status in SAT-freshman grade relationships across gender and racial subgroups. *Educational Measurement: Issues & Practice*, *35*(1), 21-28. doi:10.1111/emip.12103

Higgins, M. J., (2009). Standardized tests: Wristwatch or dipstick? *Research in Education*, 81(1), 1-11.

Hiss, W. C., & Franks, V. W. (2014). Defining promise: Optional standardized testing policies in American college and university admissions. *Report of the National Association for College Admission Counseling (NACAC). http://www.nacacnet.org/research/research-data/nacac-research/Documents/DefiningPromise. pdf.*

Houser, L. C., & An, S. (2014). Factors affecting minority students' college readiness in mathematics. *Urban Education, 50*(8), 938-960. doi:10.1177/004208591453699

Huddleston, A. P. (2014). Achievement at whose expense? A literature review of test-based grade retention policies in U.S. schools. *Education Policy Analysis Archives*, *22*(18), 1-31. doi:10.14507/epaa.v22n18.2014

Jakee, K., & Keller, E. (2017). The price of high-stakes educational testing: Estimating the aggregate costs of Florida's FCAT exam. *Journal of Education Finance*, *43*(2), 123-151.

Jamgochian, E. M., & Ketterlin-Geller, L. R. (2015). The 2% transition: Supporting access to state assessments for students with disabilities. *Teaching Exceptional Children*, *48*(1), 28-35.

Jensen, E. (2009). *Teaching with poverty in mind: What being poor does to kids' brains and what schools can do about it*. Alexandria, VA: ASCD.

Jimerson, S. R. (2001). A synthesis of grade retention research: Looking backward and moving forward. *California School Psychologist*, *6*47-59.

Jimerson, S. R., & Renshaw, T. L. (2012). Retention and social promotion. *Principal Leadership*, *13*(1), 12-16. Retrieved from http://ezproxy.liberty.edu/login?url=https://search-proquest-com.ezproxy.liberty.edu/docview/1035369546?accountid=12085

Jung, L., & Guskey, T. (2012). *Grading exceptional and struggling learners*. Thousand Oaks: Corwin Press.

Katsiyannis, A., Zhang, D., Ryan, J. B., & Jones, J. (2007). High-stakes testing and students with disabilities. *Journal of Disability Policy Studies*, *18*(3), 160-167.

Kettler, R. J., Elliott, S. N., Kurz, A., Zigmond, N., Lemons, C. J., Kloo, A., Shrago, J., Beddow, P. A., Williams, L., Bruen, C., Lupp, L., Farmer, J., & Mosiman, M. (2013). Predicting end-of-year achievement test performance: A comparison of assessment methods. *Assessment for Effective Intervention*, *39*(3), 156-169.

King, J. A., & Rohmer-Hirt, J. A. (2011). Internal evaluation in American public school districts: The importance of externally driven accountability mandates. *New Directions for Evaluation*, (132), 73-86. doi:10.1002/ev.397

Kling, K. C., Noftle, E. E., & Robins, R. W. (2012). Why do standardized tests underpredict women's academic performance? The role of conscientiousness. *Social Psychological and Personality Science*, 1948550612469038.

Knight, G. R. (2006). Philosophy & education: An introduction in Christian perspective. Berrien Springs, MI: Andrews University Press.

Koutsoftas, A. D., Gray, S., Nippold, M., & Schneider, P. (2012). Comparison of narrative and expository writing in students with and without language-learning disabilities. *Language, Speech & Hearing Services in Schools*, *43*(4), 395-409. doi:10.1044/0161-1461(2012/11-0018

LaSalle, T. P., Roach, A. T., & McGrath, D. (2013). The relationship of IEP quality to curricular access and academic achievement for students with disabilities. *International Journal of Special Education*, *28*(1), 135-144.

Lai, S. A., & Berkeley, S. (2012). High-stakes test accommodations: Research and practice. *Learning Disability Quarterly*, *35*(3), 158-169.

Lane, S., & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education, 39*(1), 165-214. doi:10.3102/0091732x14556073

Laurie, R. (2009). Raising the bar: A data-driven discussion on grade inflation. *Education Canada*, *49*(4), 32-34.

Layton, L. (2014, September 11). Common Core 2.0: Common Core by another name. *The Washington Post*. Retrieved 2016, from http://www.highbeam.com/doc/1P2-37170097.html?refid=easy_hf

Learning in Ohio. (n.d.). Retrieved April 8, 2018, from http://education.ohio.gov/Topics/Learning-in-Ohio

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

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

Minarechova, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy*, *3*(1), 82-100.

Moffatt, G. K. (1993). The validity of the SAT as a predictor of grade point average for nontraditional college students.

Mora, R. (2011). "School is so boring": High-stakes testing and boredom at an urban middle school. *Perspectives on Urban Education*, *9*(1), 1-9.

Muir, M. (2005). Standardized tests and grades. Research brief. *Education Partnerships, Inc.*

Munro, J. (2008). *Roundtable Viewpoints: Educational Leadership* (1st ed.). New York: McGraw-Hill.

Nagel, D. (2015). *Effective grading practices for secondary teachers: Practical strategies to prevent failure, recover credits, and increase standards-based/referenced grading*. Thousand Oaks: Corwin Press.

Nichols, S. L., & Berliner, D. C. (2008). Testing the joy out of learning. *Educational Leadership*, *65*(6), 14

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, *20*(20), 1-30

Nichols, S. L., & Valenzuela, A. (2013). Education policy and youth: Effects of policy on practice. *Theory into Practice*, *52*(3), 152-159. doi:10.1080/00405841.2013.804306.

O'Connor, K. (2011). *A repair kit for grading: 15 fixes for broken grades* (2nd ed.). Boston: Pearson.

O'Connor, K. (2013). *The school leader's guide to grading*. Bloomington, Ind.: Solution Tree Press.

O'Connor, K. (2017). A case for standards-based grading and reporting. *School Administrator, 74*(1), 24-28. Retrieved from http://ezproxy.liberty.edu/login?url=https://search-proquest-com.ezproxy.liberty.edu/docview/1862643582?accountid=12085

Ohio Quick Facts from the US Census Bureau. (2015). Retrieved December 6, 2015, from http://quickfacts.census.gov/qfd/maps/ohio_map.html

Ohio's State Tests. (n.d.). Retrieved July 12, 2016, from http://oh.portal.airast.org/ocba/

Pan, Y. (2014). Learner washback variability in standardized exit tests. *Tesl-Ej*, *18*(2),

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher, 42*(5), 259-265.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the big five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, *36*(2), 239-249.

Philipp, J. G. "End of course grades and end of course tests in the virtual environment: A study of correlation." Liberty University, 2014.

Powell, S. R. (2012). High-stakes testing for students with mathematics difficulty: Response format effects in mathematics problem solving. *Learning Disability Quarterly*, *35*(1), 3-9. doi:10.1177/0731948711428773

Pratt, A. M. (2013). *The relationship between student performance on the Missouri Assessment Program and students' final numerical grades*. Union University.

Reardon, S.F., Arshan, N., Atteberry, A., & Kurlaender, M. (2008). *High stakes, no effects: Effects of failing the California high school exit exam.* (Working Paper 2008-10). Stanford, CA: Stanford University, Institute for Research on Education Policy & Practice.

Reeves, D. (2016). *Elements of grading: A guide to effective practice*. (2nd ed.) Bloomington, IN: Solution Tree Press.

Reeves, D., Jung, L. A., & O' Connor, K. (2017, May). What's worth fighting against in grading? *Educational Leadership, 74*(8), 42-45.

Ricketts, C. R. (2010). *End of course grades and standardized test scores: Are grades predictive of student achievement?* ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106.

Rodriguez, J. M., & Arellano, L. (2016). The impact of high-stakes testing on Latina/o students' college aspirations. *Journal of Hispanic Higher Education, 15*(2), 113-135. doi:10.1177/1538192715627192

Sackett, P. R., Kuncel, N. R., Beatty, A.S., Rigdon, J.L, Shen, W., & Kiger, T. B. (2012). The role of socioeconomic status in SAT-grade relationships and in college admissions decisions. Psychological Science, 23(9), 1000-1007. doi: 10.1177/0956797612438732.

Scaled Score Ranges for Ohio State tests. (n.d.). Retrieved September 5, 2016, from http://education.ohio.gov/getattachment/Topics/Testing/Testing-Results/Results-for-Ohios-State-Tests/Scaled-Score-Ranges-for-Ohio-State-Tests.pdf.aspx

Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A-F marking scheme. *Journal of Curriculum Studies, 46*(2), 201-224.

Science State Test Specifications and Resources. (2015). Retrieved September 05, 2016, from http://education.ohio.gov/Topics/Testing/Ohios-State-Test-in-ELA-Math-Science-SocialStudies/State-Tests-Content-Preparation-for-Teachers/Science-Test-Specifications-and-Resources.

Stark, D., & Elliott, E. C. (1912). Reliability of the grading of high-school work in English. *The School Review*, *20*(7), 442–457. Retrieved from http://www.jstor.org/stable/1076706

Steffan, A. L. (2010). *An examination of the correlation between ninth grade mathematics grades and student performance on the mathematics portion of the California high school exit examination*. Pepperdine University.

Stiggins, R. J. (1999). Barriers to effective student assessments. *Education Digest, 64*(6), 25.

Stiggins, R. (2007). Assessment through the student's eyes. *Educational Leadership, 64*(8), 22-26.

Stiggins, R. (2014). Improve assessment literacy outside of schools too. *Phi Delta Kappan*, *96*(2), 67. doi: 10.1177/0031721714553413

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *Journal of Educational Change*, *10*(2/3), 211-227. doi:10.1007/s10833-009-9105-2

Sutinen, A. (2008). Constructivism and education: Education as an interpretative transformational process. *Studies in Philosophy & Education*, *27*(1), 1-14. doi:10.1007/s11217-007-9043-5

Swan, G. M., Guskey, T. R., & Jung, L. A. (2014). Parents and teachers perceptions of standards-based and traditional report cards. *Educational Assessment, Evaluation and Accountability*, *26*(3), 289-299.

Tanis, B. (2014). Pushing back against high stakes for students with disabilities. *American Educator*, Winter 2014, p. 19-23.

Thompson, G. L., & Allen, T. G. (2012). Four effects of the high-stakes testing movement on African American K-12 students. *The Journal of Negro Education*, *81*(3), 218-227.

Toldson, I. A., & McGee, T. (2014). What the ACT and SAT mean for black students' higher education prospects. *Journal of Negro Education*, *83*(1), 1-3

Turnbull, M. (2016). Adverse impact of high-stakes testing. *Education, 97*(7), 6-7. Retrieved from http://ezproxy.liberty.edu/login?url=https://search-proquest-com.ezproxy.liberty.edu/docview/1845980089?accountid=12085

Ullman, E. (2005). Study: High-stakes tests have no effect on achievement. *District Administration*, *41*(11), 18.

Vatterott, C. (2015). *Rethinking grading: Meaningful assessment for standards-based learning*. Alexandria, Va.: Association for Supervision and Curriculum Development.

Vernaza, N. A. (2012). Teachers' perceptions of high-stakes accountability in Florida's Title I elementary schools. *Current Issues in Education*, *15*(1), 1-11.

Watson, C. E., Johanson, M., & Dankiw, J. (2014). Effects of high-stakes testing on third through fifth grade students: Student voices and concerns for educational leaders. *E Journal of Organizational Learning & Leadership*, *12*(1), 1-11.

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. (2013). Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues & Practice*, *32*(2), 26-36. doi:10.1111/emip.12009

Wongsurawat, W. (2009). Does grade inflation affect the credibility of grades? Evidence from US law school admissions. *Education Economics*, *17*(4), 523-534. doi:10.1080/09645290802470061

Woodruff, D. J., Ziomek, R. L. (2004). *High School Grade Inflation from 1991 to 2003*. (Research Report Series 2004-04). Iowa City, IA: ACT Inc.

Wormeli, R. (2006). *Fair isn't always equal: Assessing and grading in the differentiated classroom* (1st ed.). Portland, ME: Stenhouse Publishers.

Yell, M. L., Katsiyannis, A., Collins, J. C., & Losinski, M. (2012). Exit exams, high-stakes testing, and students with disabilities: A persistent challenge. *Intervention in School and Clinic*, *48*(1), 60-64.

Yilmaz, K. (2008). Constructivism: Its theoretical underpinnings, variations, and implications for classroom instruction. *Educational Horizons*, *86*(3), 161-172.

Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D. R., Dennison, A., Triezenberg, H., Sharpe, M., & Hawes, M. (2004). What we know and need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children*, *71*(1), 75-95.

Zellmer, M. B., Frontier, A., & Pheifer, D. (2006). What are NCLB's instructional costs?. *Educational Leadership*, *64*(3), 43-46.

Zhang, Q., Sanchez, E. I., & ACT, I. (2013). *High school grade inflation from 2004 to 2011*. (ACT Research Report Series, 2013 (3)). Iowa City, IA: ACT Inc.

Zwick, R. (2013). Disentangling the role of high school grades, SAT scores, and SES in predicting college achievement. *ETS Research Report Series*, *2013*(1), 1-20. doi:10.1002/j.2333-8504.2013.tb02316.x

Zwick, R., & Greif Green, J. (2007). New perspectives on the correlation of SAT scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement, 44*(1), 23-45. doi:10.1111/j.1745-3984.2007.00025.x

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement, 48*(2), 101-121. doi:10.1111/j.1745-3984.2011.00136.x

APPENDIX

Appendix A: IRB Approval

LIBERTY UNIVERSITY.
INSTITUTIONAL REVIEW BOARD

July 12, 2017

Mark Robert Potts
IRB Application 2856: The Relationship between Teacher-Assigned Course Grades and the Ohio Air End of Course 8th Grade Science Test

Dear Mark Robert Potts,

The Liberty University Institutional Review Board has reviewed your application in accordance with the Office for Human Research Protections (OHRP) and Food and Drug Administration (FDA) regulations and finds your study does not classify as human subjects research. This means you may begin your research with the data safeguarding methods mentioned in your IRB application.

Your study does not classify as human subjects research because it will not involve the collection of identifiable, private information.

Please note that this decision only applies to your current research application, and any changes to your protocol must be reported to the Liberty IRB for verification of continued non-human subjects research status. You may report these changes by submitting a new application to the IRB and referencing the above IRB Application number.

If you have any questions about this determination or need assistance in identifying whether possible changes to your protocol would change your application's status, please email us at irb@liberty.edu.

Sincerely,

G. Michele Baker, MA, CIP
Administrative Chair of Institutional Research
The Graduate School

LIBERTY
UNIVERSITY.
Liberty University | Training Champions for Christ since 1971