AN EXAMINATION OF THE RELATIONSHIP BETWEEN THE FREQUENCY OF STANDARDIZED TESTING AND ACADEMIC ACHIEVEMENT

by

ERIC W. BERGMANN

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

June 2014

DISSERTATION APPROVAL PAGE

Student: Eric W. Bergmann

Title: An Examination of the Relationship Between the Frequency of Standardized Testing and Academic Achievement

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Methodology, Policy, and Leadership by:

Yong Zhao	Chairperson
Keith Hollenbeck	Core Member
Charles Martinez	Core Member
Brigid Flannery	Institutional Representative

and

Kimberly Andrews Espy	Vice President for Research and Innovation; Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2014

DISSERTATION ABSTRACT

Eric W. Bergmann

Doctor of Education

Department of Educational Methodology, Policy, and Leadership

June 2014

Title: An Examination of the Relationship Between the Frequency of Standardized testing and Academic Achievement

Over the past twenty years, there has been significant research conducted on the effects of large-scale standardized tests on academic achievement. Policy makers around the world have developed policies and allocated substantial sums of money in order to increase the frequency of large-scale standardized tests, although existing research offers inconclusive findings as to whether the use of large-scale standardized tests leads to higher achievement. This study was intended to empirically examine the use of standardized testing and its relationship with student achievement. The study focused on two questions: first, why do some nations require their students to take large-scale standardized tests more frequently than others? And second, is there a correlation between the frequency of large-scale standardized tests frequency and academic achievement? This study examined data from the 2003 and 2009 administrations of the Program for International Student Assessment (PISA) in order to address these questions. Results from this study indicated the frequency of large-scale standardized tests is most likely to be associated with testing consequence or stake (e.g., data are made public, etc.). Additionally, results suggest that the frequency of large-scale standardized tests is not significantly related to academic achievement.

CURRICULUM VITAE

NAME OF AUTHOR: Eric W. Bergmann

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
University of Utah, Salt Lake City, UT

DEGREES AWARDED:

Doctor of Education, Education Leadership, 2014, University of Oregon
Master of Science, Education, 1996, University of Utah
Bachelor of Arts, History, 1991, University of Utah

AREAS OF SPECIAL INTEREST:

K-12 Leadership
Project-Based Learning
Teaching and Learning for the 21st Century

PROFESSIONAL EXPERIENCE:

Principal, Newberg Public Schools, Newberg, OR, 2011-2014

Principal, Portland Public Schools, Portland, OR, 2008-2011

Principal, Granite School District, Salt Lake City, UT, 2006-2008

Assistant Principal, Granite School District, Salt Lake City, UT 2000-2006

Teacher, Granite School District, Salt Lake City, UT, 1992-2000

PUBLICATIONS:

Bergmann, E. (2003). Taylorsville's students find taking risks empowering. *The Utah Special Educator, 23*(5), 6-7.

ACKNOWLEDGEMENTS

I thank Professor Zhao and my committee for their guidance, encouragement and patience throughout the writing of this manuscript. Special thanks are due to Professor Hollenbeck for his help in the final stages of my writing, as well as helping assemble my committee in the eleventh hour. Finally, I wish to thank my EMPL cohort for their constant friendship and support.

For my wife, for always believing in me, my children who I hope are inspired to follow in my footsteps, and my parents who have always supported my dreams.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The impact that the standards and accountability movement and the corresponding increase in the use of standardized testing have had on education has been massive. Sacks (1999), asserted that those responsible for the efforts to introduce more accountability into the school systems have "accomplished a near-complete makeover of American schools" (p. 68). He went on to state the standards and accountability movement has "fundamentally altered the very nature and meaning of education in America: what it means to teach, to learn, and to achieve" (ibid).

Sacks's statement was prophetic as the publishing of his book occurred three years before the passage of the 2002 re-authorization of the Elementary and Secondary Education Act, commonly referred to as No Child Left Behind (No Child Left Behind [NCLB], 2003). What many consider to be the culmination of the standards and accountability movement (Nichols, Glass & Berliner, 2006), NCLB increased the degree to which all stakeholders would be held publically accountable for making academic progress (Chubb, 2009). The legislation brought with it substantial controversy, not just about what some felt were the seemingly impossible expectations laid out for schools by the law, but also by the law's requirement to utilize standardized tests in order to determine whether students, schools, and, for that matter, entire education systems made progress toward the goals set forth by NCLB (Simpson, Lacava, & Sampson-Graner, 2004).

Research has shown that standardized testing has served multiple purposes. Sacks (1999) stated that the prevalence of standardized testing is a result of America's long-held fascination with measuring and comparing American minds. Wiliam suggested that the purpose of standardized testing is "deceptively simple" (p. 110, 2010), in that such testing is simply meant to measure the quality of schools. Linn (2001) asserted that standardized tests are often used as an intervention in and of themselves; that is, as a "tool" for improving the efficiency of schools (p. 29). Shepard (2008) brought these purposes together:

> "Predominantly, tests have been used to make decisions about individual students, especially to place students in special programs…Accountability testing—focused on judging the quality of schools—is a more recent phenomenon, but it has its roots in the technology of IQ testing and the ardent belief among Americans that tests can scientifically determine merit and worth" (p. 25)

Linn further suggested that the prevalence of standardized testing is due to a desire by contemporary policy-makers to have schools run more like businesses, and the use of testing would ensure more corporate-like efficiency, quality control and accountability. Regardless of the purpose of standardized testing, Shepard (ibid.) pointed out that since the 1970s, there has been a " huge burgeoning" in the amount of standardized testing conducted in the United States (p. 27).

The use of standardized testing has become more popular not just in the United States, but also in countries around the world (Kamens & McNeely, 2010). Since the inception of the standards and accountability movement in the early 1980s, the use of standardized testing among high-school aged students in countries around the world has increased substantially (Morris, 2011). However, that trend stands in stark contrast to the fact that existing research is inconclusive on whether standardized testing has any

positive effect on academic achievement (Mehrens 2002; Linn, 2001). According to the oft-cited research of Carnoy and Loeb (2002), there is little evidence of a relationship between accountability based on standardized testing and academic achievement.

Despite the lack of research, countries are investing huge sums of money in order to support standardized testing systems. For example, in 2012 in the United States alone, $1.7 billion was spent on educational testing (Chingos, 2012).

What triggers more frequent use of standardized testing is an unclear question, as no research currently exists that addresses that question. However, identifying reasons why countries test students more frequently seems to be an important preliminary question in the present research as it will help identify countries whose testing policies are worthy of the secondary analysis regarding testing frequency and academic achievement. An understanding of what factors lead to increases in standardized testing will help identify nations worthy of a closer of examination. Ultimately, answers to both questions would be of great importance to policy makers around the world.

In summary, this study contributes to the existing literature by examining two distinct but interrelated questions. First, why do some countries require their students to take more standardized tests than others? Second, are the changes in correlational strength between the frequency of standardized testing and academic achievement significant?

CHAPTER II

LITERATURE REVIEW

Existing literature offered a variety of definitions of standardized testing (Popham, 1999; Wang, Beckett & Brown, 2006; Morris, 2011), but most definitions tended to include several four central components, such who designs them, how they are scored, and purpose they serve. For the purposes of my research, I am defining large-scale standardized tests was defined using the following elements laid out by Morris (2011):

1. Standardized tests are designed and scored externally; that is, they are assessments developed by someone outside of the institution where the learning takes place;
2. They are typically given to large groups of students at once;
3. They follow a uniform procedure in administering, scoring and interpreting the test;
4. The results they generate are used for a variety of reasons, including assessment of learning as well as evaluation.

A Brief History of Large-Scale Standardized Tests in the United States

Standardized testing has been a regular part of the American educational landscape since the nineteenth century (Resnick, 1982). Originally, standardized tests were utilized as a method of sorting and classifying the massive influx of students that American schools were experiencing at the turn of the century (Linn, 2001). Linn pointed out that standardized testing slowly became a means to examine school efficiency

(ibid.). This second purpose was key as it demonstrated that testing in the United States has long been about holding schools accountable for their performance. As Haertel and Herman (2005) stated, "[since the turn of the century] policymakers have used tests in an attempt to discover which schools and districts are fulfilling their responsibilities and which are not" (2005, p. 28-29).

Throughout its long history, despite the stated purpose of educational testing, the effect and use of standardized testing seems to have been a source of polarizing, perpetual debate, winning its share of both supporters and critics; in fact, Cronbach (1975) demonstrated that the public debate sparked by educational testing has lasted at least 50 years. Linn (2001) suggested that the nature of the debate always seems to revolve around the notion that if schools were run more like businesses, their performance would go up and "testing is seen as a tool to prod educators into making the desired transformation" (p. 31).

Commenting on the polarizing effect of education testing, Madaus (1985) stated that standardized testing and assessment have been both the focus of controversy as well as the darling of policy makers. Finally, Linn (2001) echoed that sentiment 16 years later when he said, "Americans have had a love-hate relationship with educational testing" (p. 29). Despite the conflict, it appears that the place and use of testing in the United States seems permanent. According to Gunzenhauser (2003) standardized testing can now be considered a core aspect of the new default educational philosophy in this country.

Large-Scale Standardized Tests as an International Phenomenon

Driven by the desire to increase the achievement levels of their students, countries have invested significant levels of time and resources developing educational policies and procedures that include standardized testing. According to the United Nations Educational, Scientific and Cultural Organization (UNESCO) 2008 Global Monitoring Report for its Education for All initiative, between 2000 and 2006, worldwide use of standardized testing increased dramatically. Developed countries have seen a 23% increase, developing countries saw a 22% increase and countries in transition have seen an increase of 17%. It is worth noting that the 2008 UNESCO report focused solely on national assessments and as such it did not include any locally developed and administered standardized assessments. Further, the report made no attempt to calculate how frequently students were tested.

However, the report does lend support to the idea that the use of standardized testing is on the rise around the world. Kamens and McNeely (2010) offered three reasons that explain the surge in the use of standardized testing in the world. First, education has been re-prioritized, especially since the 1990s. Countries are not only viewing education as an engine to drive economic gain, but it has become an instrument "to promote the public good" (p. 10).

Second, as Kamens and McNeely (2010) point out, the increase in the use of testing is a reflection of the "hegemony of science" (p. 11) found around the world. That is to say, there is a now pervasive belief in many countries that no human endeavor is beyond the scope of rational analysis. Standardized testing became a means to

understand, measure, manage, and in the minds of some to improve educational achievement. Karpicke and Roediger (2006) asserted "testing is a powerful means of improving learning, not just measuring it" (p. 249).

In offering a third explanation for the increase of standardized testing, Kamens and McNeely pointed to the research of Meyer (2005) who suggested that countries are more and more resorting to philosophies of organizational, or business-like management when making decisions about the welfare of their of their citizens; meaning that there are "standard solutions to education problems" (Kamens and McNeely, 2010, p. 14). They asserted that standardized testing fits this notion remarkably well. Morris (2011) referred to this management philosophy as "New Public Management" (p. 7). Once again referring to an effort to run public systems with business-like efficiency, New Public Management focuses more on outputs such as test scores than inputs such as funding, and places a premium on efficiency, cost-effectiveness, quantifying results, and making decision makers accountable for their actions (Mons, 2009 as cited in Morris, 2011).

Morris (2011) offers four additional reasons for the rise in the use of standardized testing worldwide: (a) an increase in the use of standards-based assessment (see "Uses and Effects of Standardized Testing" for further discussion); (b) increased international competition, as measured and reported through the PISA and through the Trends in International Mathematics and Sciences Study (TIMSS); (c) a heightened focus on a particular subject matter, such as 21st century skills; and finally (d) pressure exerted on school systems by a growing and profitable testing industry.

Cost of Large-Scale Standardized Tests

Data that illustrates how much countries (other than the United States of America) spend specifically on educational assessment is hard to find. Some forms for such data exist for the United States, however. According to Dan Lips (2007), a researcher at the Heritage Foundation, a think tank in Washington D.C., in 2007 American state and local governments paid an estimated $141 million in administrative costs related to NCLB. More recently, the state of Texas signed a contract with Pearson to manage all standardized testing needs for the state between 2010 and 2015 at the cost of nearly $470 million (Cargile, 2012).

Some estimates are much higher when the United States is looked at as whole. According to a recent Brookings Institute report (Chingos, 2012), the United States is now spending approximately $1.7 billion a year on testing. If the United States is spending significant sums of money in support of the use of standardized tests, it can be assumed that there are other countries whose education budgets have gone up due to the burden of testing costs.

Connecting Stakes to Large-Scale Standardized Tests.

One aspect of standardized testing that is frequently discussed and debated is the consequences of the test, that is, the *stakes* of the test. Madaus (1988) explained that a test becomes high-stakes when its results are used to make important decisions about students, teachers, administrators, schools, and/or systems. For example, Au (2007) stated that a high-stakes test might be tied to student graduation or in some cases teacher or principal salaries. McNeil (2000) added that another factor that makes a test high-

stakes is that the results are reported to the public. Conversely, the terms no- or low-stakes standardized tests are used as well. The results of such tests carry no consequences for students. However, Morris noted that in some cases, while a test may carry no- or low-stakes for students, it still might carry serious consequences for teachers, schools, or systems of schools (Morris, 2011).

Wiliam (2010) made a key observation in looking at how countries differ in the use of stakes in testing systems. He asserted that in the United States, standardized tests tend to carry high stakes for teachers, schools and school systems, but those same tests tend to carry no- or low-stakes for students. However, in Europe and Japan, the inverse is the case. In those cases, standardized tests tend to carry no- or low-stakes for teachers, schools or systems and carry high-stakes for students. Despite this difference, it is important to note that for the purposes of this study, no difference will be made between low- and high-stakes accountability testing. If there are any consequences attached to the results of a test, they will be considered tests with stakes.

The attachment of stakes to standardized tests has brought forth many arguments. There is a large body of evidence (Wang, et al., 2006) that suggests that attaching stakes to standardized tests brings forth a broad spectrum of unintended negative consequences such as lower motivation, lack of engagement, higher drop out rates, unethical test preparation, questionable ethics during test administration and so forth. Linn (2000) raised the often overlooked, but widely voiced question that pertains to validity – are standardized tests measuring what they purport to measure or are they in fact measuring something else, like the effects of test preparation?

The Purposes of Large-Scale Standardized Tests

Prior to the onset of the standards and accountability movement, assessment had a much simpler meaning than it does today (Morris, 2011). Black (1990) offered this purpose of educational assessment, "Assessment is at the heart of the process of promoting children's learning. It can provide a framework in which educational objectives may be set and pupil's progress charted and expressed" (p. 27). However, as Linn (2001) pointed out, over the last two decades, educational assessment, and specifically, standardized testing, has taken on a broader goal. Before introducing literature on the pros and cons of standardized testing, understanding the intended purposes of standardized testing offers a set of lenses through which each study may be examined. Those lenses, those intended purposes, offer a framework around which the literature may be studied. Wang, et al. (2006) offered four purposes, four "interlocking cornerstones" (p. 305) of standardized testing: (a) their use as an assessment-driven reform, that is, as an instrument to spur academic improvement; (b) their use in standards-based assessment; (c) their use in assessment-centered accountability; and (d) their use in determining high-stakes consequences. In the following two sections, existing literature on standardized testing has been divided between those who support and those who are critical of the use of standardized testing in schools, and then organized according to this framework.

Support for Large-Scale Standardized Tests

As mentioned previously, the use of standardized tests as a reform, that is, as an instrument to increase learning and not just measure learning is not uncommon. The

1980s saw the emergence of a new assessment philosophy referred to as Measurement-Driven Instruction (MDI). According to Popham (1987), if standardized tests were designed to measure important skills and carried enough weight with the students, then the tests could become an instrument to improve instruction. Popham, Cruse, Rankin, Sandifer and Williams (1985) and Popham (1987), pointed to an increase in tests scores on minimum competency tests to prove their point. Additionally, Karpicke and Roediger (2008) offered evidence that repeated testing, specifically, repeated testing that practiced rote memorization, enhanced long-term retention.

Many supported the use of standardized testing in pursuit of standards-based assessment, a hotly debated topic since the inception of NCLB in 2001 (Wang, et al., 2006). According to Wang, et al. (2006), standards-based assessment describes the process of assessing student performance relative to a set of standards rather than a normed group; it is an attempt to bring "all children to the same set of high standards" (p. 313).

Since the passage of NCLB in 2002, using standardized tests in order to hold teachers, schools, districts and states accountable has become a norm. In a strong call for using tests for this purpose, Phelps (2000) cited research that in countries where standardized tests were dropped, quality of curricula went down, students were less engaged, and evidence for student promotion to advanced programs became less evident. Additionally, when comparing the specific use of externally developed standardized tests compared to local created exams, it is clear that the "quality of teacher-made tests pales compared with more rigorously developed large-scale counterparts" (Cizek, 2001, p. 25).

Finally, supporters of standardized testing often associate their use with the now common practice of attaching stakes to the tests. According to Shanker (1995), "the United States has an education system in which very little counts" (p. 147). Cizek (2001) added to that notion that by attaching high-stakes consequences to standardized tests, teachers are now more "reflective, deliberate and critical" in their practice (p. 24).

Criticisms of Large-Scale Standardized Tests

The notion that standardized tests alone can effect change is an idea that has its critics. Roeber (1995) cut straight to the heart of the matter by asserting that tests alone do not create improvement. He stated, "[A] program of systemic change begins with the content standards" (p. 284). Amrein-Beardsley and Berliner (2003) echoed that sentiment by arguing that testing does not lead to an increase in learning.

Additionally, while the use of standardized testing has continued in schools, it is losing its influence among the general population as a viable educational reform. In 1984, Airasian highlighted a Gallup poll when he pointed out "tests are trusted and desired by a majority of Americans" (p. 394.) In 2007, 28% believed that standardized testing has helped academic achievement, and only 28% of Americans believed that standardized testing has hurt academic achievement (Bushaw & Lopez, 2013). According to Bushaw and Lopez (ibid.), in the 2013 Gallup poll, only 22% of American believe that testing has helped academic achievement, and 36% believe that standardized testing has hurt academic achievement. Based on these data, it can be surmised that the public support for standardized testing as a viable educational reform has been in a state of decline for nearly 30 years.

Regarding their use in standard-based assessment, there are two primary counter-arguments. The first is that the process robs local agencies of the ability to make autonomous decisions based on the individual needs of their constituents (McDonnell, 2008). The second is that standards-based assessment forces a uniform set of expectations on every student, regardless of their abilities. In so doing, Koretz (1995) argued that schools would have to either "[dumb] down instruction down to the lowest common denominator or condemn low-ability students to frequent failure" (p. 159).

Those opposed to the idea of using standardized tests for the purpose of supporting assessment-based accountability offered several reasons, but the literature is clear that it is not because opponents are opposed to accountability. Kohn (2000) offered, "endorsing the idea of accountability is quite different from holding students and teachers accountable specifically for raising test scores" (p. 46).

Perhaps the most common line of criticism regarding this purpose of standardized testing is that in the opinion of many, no test can serve the multiple purposes of being a summative assessment and at the same time offering diagnostic information on students (Madaus, 1995). Popham (1999) highlighted this point by suggesting that while educators do need to develop practices to generate feedback on their instruction, the use of standardized achievement tests to accomplish that task is fundamentally flawed. As Shepard (1989) suggested, the results from such tests are skewed due to the tendency to narrow the curriculum in order to focus on tested information.

Examining Frequency of Large-Scale Standardized Tests

The crux of the debate concerning the benefits and drawbacks of standardized testing hinged on the relationship between standardized testing and academic achievement, which is at best inconclusive (Linn, 2000). However, while many aspects of testing have been studied, there are still variables to be explored that may contribute to a better understanding of the role of standardized testing. For instance, little research existed on the effects of the frequency that standardized testing are used in schools around the world. We know that there has been a significant increase in testing over the last decade (Bushaw and Lopez, 2013), but how frequently are those tests given? Morris (2011) is one of a few sources that discussed the frequency of standardized testing, and even then the author simply stated that there is great variation in testing frequency around the world. Given that there are those who believe that tests increase academic achievement, it stands to reason that among pro-testing advocates there may reside a belief that the more frequently we test students, the higher student achievement will climb.

Greaney and Kellaghan (2008) mentioned that testing national standards typically occured annually, but may occur "more often where the system allows for repeats" (p. 15). These authors do caution their readers from over assessment as they claim it is unnecessary and costly. Black (2010) briefly mentioned the negative consequences of frequent testing if tests that are designed for the use of summative purposes are used diagnostically. Finally, there are multiple studies that report the benefits of frequent, in-class testing, using teacher-developed instruments, but none of these studies referred to

standardized assessments in the way that standardized assessments were defined for the present study (Bangert-Drowns, Kulik & Kulik, 1991; Basol and Johanson, 2009).

This study consists of two interrelated questions. The first question asks why some countries test students more frequently than others. Certain factors will be identified that lead to higher instances of standardized testing. The second question asks whether the change in correlational strength between the frequency of standardized testing and academic achievement from 2003 to 2009 is statistically significant from zero.

In addition to the policy implications and the exorbitant costs associated with standardized testing, the paucity of research that identify factors that lead to more frequent use of standardized tests as well as research that examine the relationship between the frequency of standardized testing and educational outcomes serve as the rationale for this study.

CHAPTER III

METHOD

This chapter explained the methods that were used in my study of the relationship between the frequency of standardized testing and academic achievement levels. The sections of this chapter provide a description the research questions and hypothesis, the research design, participants and sampling methods used, the independent and dependent attribute variables used in the analysis, and strategies used for missing data.

Research Questions and Hypotheses

The present study was designed to explore two distinct but interrelated questions. The first research question was why some countries test students more frequently than others. Additionally, a second question examined the change in correlational strength between the frequency of standardized testing and academic achievement in sample nations from 2003 to 2009 statistically significant from zero.

In regards to the first question, considering trends that are discussed in the existing literature, I hypothesized that the results will show a relationship between the use of some form of stake or consequence with a test and how frequently a country tests its students. As for the second question, given the mixed results that exist in current literature regarding the correlation between standardized testing and academic achievement, I hypothesized that the change in correlational strength between testing frequency and academic achievement will not prove statistically significant from zero.

Variables and Measures

For the purposes of this study, the independent attribute variable – frequency of standardized testing – was defined as the rate that students in a particular educational system (in this case, nations) were required to participate in standardized testing. As mentioned before, standardized testing were those tests that are designed and scored externally, are uniformly administered and scored, were typically given to large groups at once, and had results that are used for a variety of reasons.

The independent attribute variable was measured by data collected from the 2003 and 2009 PISA administrations. Established by the OECD, the PISA is an assessment designed to evaluate school systems around the world in three key areas: reading, math and science (Sjøberg, 2012). Developed in 1997 and first administered in 2000, the PISA has been used to assess reading, math and science levels among 15-year old students in nations around the world every three years since 2000 (OECD, 2003; OECD, 2009).

In addition to the academic assessment that the PISA administers to fifteen-year old students in each participating nation, students also completed questionnaires that gathered information about their study habits, opinions of school, and home life. Additionally, school leaders in participating schools were also asked to complete questionnaires that ask questions pertaining to the school's demographics, curriculum, and so forth (OECD, 2003; OECD, 2009). The present study is focused on the responses to several of the questions that were found in the 2003 as well as the 2009 PISA school questionnaires.

Data used to answer the study's preliminary question of what factors cause a

country to test students more frequently was drawn from responses to the 2003 and 2009 PISA school questionnaires. Data regarding testing frequency was pulled from question 12a of the 2003 school questionnaire (see appendix A). The same question appeared as question number 15a in the 2009 school questionnaire (see appendix B). Questions that offer data regarding the independent attribute variables used in the first research question were drawn from questions 16 and 22 from the 2009 PISA school questionnaire (see appendices C and D, respectively).

Respondents were then given the option of selecting one of the following five levels for each method: (a) Never, (b) 1-2 times a year, (c) 3-5 times a year, (d) Monthly, or (e) More than once a month (OECD, 2009; OECD, 2003). Responses to the PISA school questionnaire were delivered in the form of descriptive statistics, specifically with the percentage of the total number of respondents in each country who responded to each level.

The independent attribute variable was measured using a Testing Frequency Index (TFI) that had been created for the purposes of this study. Each nation's TFI, essentially a weighted mean, was intended to be a comparable value of how frequently its 15 year-old students must take standardized tests. The TFI was calculated using the following method: the percentage of responses to each level was multiplied by the value of the level (1 through 5). Those five values were added together and then divided by 100 in order to obtain the TFI, which was a number between 1-5. The TFI were compared to the values of the original levels in order to gain a sense of how frequently a nation's 15-year olds must take standardized tests. Table 1 demonstrates the process used to determine the 2003 TFI for the United States.

Table 1

2003 Testing Frequency Index Calculation for the United States

Response Level (A)	Percent Responding (B)	Weighted Values (AxB)
1	1.5	1.5
2	75.88	151.76
3	18.95	56.85
4	1.87	7.48
5	.18	.9
Sum of Weighted Values		218.49 (C)
TFI (C/100)		2.1849

The 2003 TFI for the United States is 2.1849, which means that on average, in 2003 15 year-old students in the United States were required to sit for standardized tests somewhere between 1-2 and 3-5 times a year.

The dependent attribute variable of the present research will be the academic achievement level in math, reading and science for each of the 40 sample nations. The variable will be measured using the 2003 and 2009 PISA aggregated scores of each participating nation's sample of 15 year-old students.

Design

A descriptive design was used in this study (Anastas, 1999). The current study analyzed extant data from a convenience sample of nations that participated in the 2003 and 2009 PISA administrations.

My study involved asking two distinct, yet interrelated questions: first, why do some countries require their students to take more standardized tests than other countries? Second, is the change in correlational strength between testing frequency and academic achievement between 2003 and 2009 among sample nations statistically significant from zero? Data for each variable was collected from the OECD database located on the organization's website.

Prior to any computation, the underlying assumptions for a correlational analysis were verified. They were: (a) the studied variables must be interval or ration nature; (b) the variables must be approximately normal in distribution; (c) there is a linear relationship between the two variables; (d) outliers are kept to a minimum or are removed entirely; and finally, (e) there is homoscedasticity of the data (Morgan, Leech, Gloeckner & Barrett, 2011).

The first question identified reasons why countries utilize standardized testing frequently. No theories currently exist that explain why some countries test students more frequently than others. Given the paucity of a substantiated theory, trends identified in existing research were selected as the independent attribute variables in the present study. One group of variables that several studies (Amrein & Berliner, 2003; Au, 2007; Madaus, 1988; McNeil, 2000; Morris, 2011; Wang, et al., 2006) have suggested may have unintended or negative effects on academic achievement are those that involve the use stakes, or consequences, in testing situations. The present research selected variables that were used the 2009 PISA School Questionnaires that represented what could be considered a consequence or some form of stake. Those eight variables included:

1. Using assessment results to make decisions about a students' retention or promotion (question 16b).
2. Using assessment results to compare the school to national performance (question 16d).
3. Using assessment results to make judgments about teacher effectiveness (question 16f).
4. Using assessment results to compare the school with other schools (question 16h).
5. Achievement data are posted publicly (e.g., in the media) (question 22a).
6. Achievement data are used in evaluation of the principal's performance (question 22b)
7. Achievement data are used in evaluation of teachers' performance (question 22c).
8. Achievement data are used in decisions about instructional resource allocation to the school (question 22d)

Using Statistical Package for Social Sciences software (SPSS), version 21.0, a scatterplot was created and a correlation coefficient was computed to assess the relationship and the corresponding significance between each of the independent attribute variables and the 2009 TFI.

For the second question (whether the change in correlational strength between testing frequency and academic achievement between 2003 and 2009 is statistically significant from zero), a correlational analysis was conducted between the 2003 TFI and each of the 2003 PISA testing scores (math, reading, and science) for each participating country. Using Statistical Package for Social Sciences software (SPSS), version 21.0, a scatterplot was created and a correlation coefficient was computed to assess the

relationship and the corresponding significance. The same process will take place between the 2009 TFI and the corresponding 2009 PISA scores for each participating country.

Subsequently, the significance of the change between the three 2003 correlation coefficients (between 2003 TFI and 2003 math, reading, and science scores from PISA) and the three 2009 correlation coefficients (between 2009 TFI and 2009 math, reading, and science scores from PISA) were analyzed using a Fisher's r to z transformation. Information was plugged into the applet found at www.vassarstats.net/rdiff.html. The calculated z scores revealed whether the difference in correlation between the 2003 and 2009 TFI and corresponding PISA math, reading and science scores were statistically significant from zero.

Validity and reliability issues. In his seminal work, Cronbach (1957) made a distinction between experimental and correlational research in that they used different samples, measures, analyses and inferences. Cook and Campbell (1976) built on Cronbach's work by offering a set of concepts that were helpful for evaluating the validity of correlational research. Among those concepts, and most pertinent to the first stage of the present study were the concepts of internal validity and construct validity.

In experimental designs, internal validity has to do with some spurious event confounding the relationship between the treatment and the dependent variable (Mitchell, 1985). The majority of confounds offered by Campbell, Stanley and Gage (1963) do not tend to fit into the conceptual framework of a descriptive design using correlation. The most common threat is the " 'third variable' that may be correlated with X or Y or both

but is not a conceptual replacement for X or Y" (Mitchell, 1985, p. 196). Mitchell (1985) suggested that in order to strengthen internal validity, "systematic thinking and measuring should be done to check for alternative explanations" (ibid.).

As for construct validity, the meanings of the constructs were brought to question as well as whether those involved in the study of the group from which they are sampled. Debate exists surrounding the construct validity of the PISA test. Sjøberg (2012) summarized the main thrust of that debate. He stated, "A fundamental premise for the PISA project is that it is indeed possible to measure the quality of a country's education by indicators that are common, i.e. universal, independent of school systems, social structure, traditions, culture, natural conditions, ways of living, modes of production etc." (p. 7). In other words, PISA claimed to be able to compare the quality of education across national lines, paying little regard to the cultural, social and political differences between countries (Sjøberg, 2012). Such claims naturally brought questions of validity to light, but Sjøberg does add that abundant evidence exists to support PISA's validity as well as to undermine it. He suggested that the problem with PISA is that the results were often selectively used for political expediency. He stated, "The reference to PISA to justify and legitimize educational reforms is widespread. This influence ought to be better researched and scrutinized" (p. 17). The present study moved forward with the caution surrounding the construct validity of PISA as a potential limitation.

Participants and Sampling Method

According to the OECD (2003), the organization that developed and administers the PISA, in 2003, 41 countries participated in the PISA. In 2009, (OECD, 2009) the number of participating countries jumped to 74 (65 countries took the assessment in 2009,

and another nine countries took the same assessment in 2010). All 41 of the countries that took participated in the 2003 PISA also participated in the 2009 PISA. Of the 41 countries that participated in both administrations, I used 40 as the subjects of study in this study's quantitative analysis of the research question. The only country that participated in the 2003 PISA that was excluded from this study was France and that was because the country did not publish any survey data regarding testing frequency. Table 2 listed the sample countries along with pertinent information.

Missing Data

It should be noted that both the 2003 and the 2009 PISA data sets that were used in the present study have missing data. For instance, during the 2003 PISA administration, achievement test data for the United Kingdom were not included in the final results because it was determined that sampling standards had not been met (OECD, 2003). Additionally, the Chinese provinces of Hong Kong and Macao did not provide testing frequency data in 2003 (ibid.); consequently, a 2003 TFI could not be calculated. Finally, Thailand and Lichtenstein did not provide data on response level 4 or 5 in the 2003 and 2009 testing frequency question. In all cases, missing data was handled by utilizing a listwise deletion strategy when appropriate.

Table 2

Sample Countries

Country	2013 Population[a]	UN Region[b]
Australia	23,344,735	Australia and New Zealand
Austria	8,485,272	Western Europe
Belgium	11.098,609	Western Europe
Brazil	200,674,130	South America
Canada	35,163,430	North America
China - Hong Kong	7,187,476	Eastern Asia
China - Macao	591,900	Eastern Asia
Czech Republic	10,676,248	Eastern Europe
Denmark	5,617,144	Northern Europe
Finland	5,425,553	Northern Europe
Germany	82,656,067	Western Europe
Greece	11,115,778	Southern Europe
Hungary	9,939,402	Eastern Europe
Iceland	329,807	Northern Europe
Indonesia	250,585,668	South-Eastern Asia
Ireland	4,627,491	Western Europe
Italy	60,891,838	Southern Europe
Japan	126,981,371	Eastern Asia
The Republic of Korea	49,158,901	Eastern Asia
Latvia	2,046,784	Northern Europe
Lichtenstein	36,713	Western Europe
Luxembourg	529,914	Western Europe

Table 2 continued

Country	2013 Population[a]	UN Region[b]
Mexico	122,730,392	Central America
Netherlands	16,752,511	Western Europe
New Zealand	4,512,546	Australia and New Zealand
Norway	5,042,200	Northern Europe
Poland	38,161,569	Eastern Europe
Portugal	10,614,640	Southern Europe
Russian Federation	142,572,794	Eastern Europe
Serbia	9,518,138	Southern Europe
Slovak Republic	5,442,195	Eastern Europe
Spain	46,853,796	Southern Europe
Sweden	9,567,347	Northern Europe
Switzerland	8,069,376	Western Europe
Thailand	67,108,507	South-Eastern Asia
Tunisia	11,002,329	Northern Africa
Turkey	75,087,121	Western Asia
United Kingdom	63,134,171	Western Europe
United States of America	320,526,920	North America
Uruguay	3,410,763	South America

Note. [a] Data collected from www.worldpopulationreview.com. [b] Information gathered from http://unstats.un.org/unsd/methods/m49/m49regin.htm.

CHAPTER IV

RESULTS

This chapter presents the results of the quantitative analyses and is divided into two sections: Question One and Question Two results. Within each section, research questions guide the presentation of the results. The chapter concludes with a summary of the results from both sections.

Analysis

My study's results consisted of data aimed at answering two research questions: first, why do some countries test more frequently than others? Second, is the difference in correlation between the 2003 and 2009 testing frequency and PISA tests among sample nations statistically significant? The second question examined whether changes in the correlation between the frequency of standardized testing and academic achievement were significant.

Research question 1: Why do some countries test more frequently than others? In order to determine why some countries require their students to test their students more frequently than others, the TFI for each of the sample nations was examined. The 2003 and 2009 TFI for each sample nation is listed in Table 3.

A country's TFI will range from 1 to 5, with 1 indicating that students are *never* required to take standardized tests, and 5 indicating that students are required to take standardized tests more than once a month. As Table 3 indicates, there was a mean of 4.59% increase in the frequency of standardized testing among sample nations from 2003 to 2009. The largest increase occurred in Indonesia where there was a 47.88% increase

Table 3

2003 and 2009 Testing Frequency Indices

Nation	2003 TFI	2009 TFI	% Change
Australia	1.7360	1.8709	7.77
Austria	1.3509	1.3923	3.06
Belgium	1.4025	1.3854	-1.22
Brazil	2.2178	2.1258	-4.15
Canada	1.9212	2.0138	4.82
Czech Republic	1.8885	1.9921	5.49
Denmark	2.0108	2.4098	19.84
Finland	2.1332	2.1765	2.03
Germany	1.4826	1.3143	-11.35
Greece	2.1244	2.1914	3.15
Hungary	1.5072	1.8090	20.02
Iceland	2.0725	1.8729	-9.63
Indonesia	1.8710	2.7668	47.88
Ireland	1.6352	1.8091	10.63
Italy	2.3384	2.1498	-8.07
Japan	1.8160	2.0391	12.29
The Rep. of Korea	2.5722	2.1015	-18.30
Latvia	2.8492	2.9425	3.27
Lichtenstein	1.9290	1.7853	-7.45
Luxembourg	2.0795	2.0531	-1.27
Mexico	2.1889	2.2903	4.63
Netherlands	2.2252	2.8368	27.49

Table 3 continued

Country	2003 TFI	2009 TFI	% Change
New Zealand	2.6625	2.3979	-9.94
Norway	2.2406	2.2422	0.07
Poland	2.1894	2.6326	20.24
Portugal	1.8259	2.2103	21.05
Russian Federation	2.1646	2.4674	13.99
Serbia	1.5014	1.5543	3.52
Slovak Republic	1.9596	2.3052	17.64
Spain	2.1488	1.3452	-37.40
Sweden	2.3561	2.4773	5.14
Switzerland	1.6239	1.8739	15.40
Thailand	1.8669	1.8073	-3.19
Tunisia	2.4737	2.7025	9.25
Turkey	2.4253	2.0684	-14.72
United Kingdom	1.7532	1.7462	-0.40
United States	2.1849	2.4189	10.71
Uruguay	1.377	1.4323	4.02
Mean	2.002	2.095	4.59

Note. TFI = Testing Frequency Index; calculations based on data from the 2009 PISA School Questionnaire. * = China did not provide any frequency data for the 2003 PISA, consequently, China – Hong Kong and China – Macao were eliminated from the list.

in the frequency of standardized testing from 2003 to 2009. The largest decrease took place in Spain where there was a 37.4% decrease in the frequency of standardized testing.

To answer the first research question, why some countries require more

standardized testing of their students than others, correlations were conducted between the 2009 TFI and a variety of independent attribute variables drawn from the 2009 PISA school questionnaire. The independent attribute variables selected were those that most closely resembled what existing literature consider a consequence or some kind of "stake" that is attached to the outcome of the assessment. Those eight variables included:

1. Using assessment results to make decisions about a students' retention or promotion (question 16b).
2. Using assessment results to compare the school to national performance (question 16d).
3. Using assessment results to make judgments about teacher effectiveness (question 16f).
4. Using assessment results to compare the school with other schools (question 16h).
5. Achievement data are posted publicly (e.g., in the media) (question 22a).
6. Achievement data are used in evaluation of the principal's performance (question 22b)
7. Achievement data are used in evaluation of teachers' performance (question 22c).
8. Achievement data are used in decisions about instructional resource allocation to the school (question 22d)

Variables that led to a high 2009 TFI. Figures 1 through 8 are scatterplots that show the association between each attribute variable and the 2009 TFI. The scatterplots are followed by Table 4, which describes the correlational values associated between each of the independent attribute variables and the 2009 TFI of the sample nations.

Assessment data are used to make decisions about student retention or promotion and 2009 TFI. Figure 1 details the 2009 TFI values and the percentage of schools in sample nations who reported that assessment data was used to make decisions about students' retention or promotion. The scatterplot demonstrates a nonlinear relationship, no evidence of either a positive or negative association, and that a weak correlation exists between the two variables.

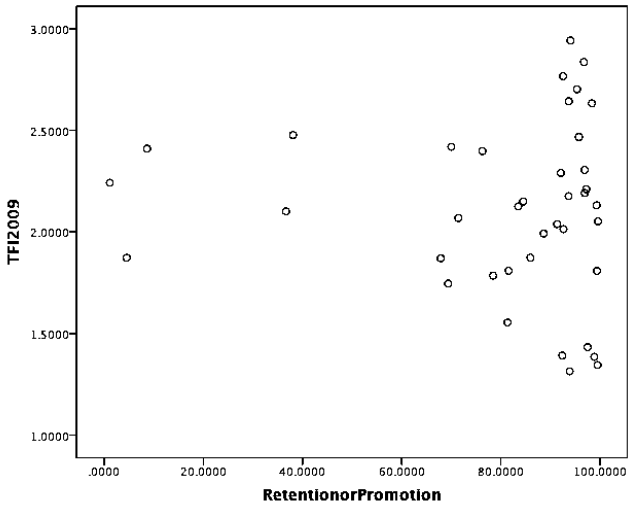

Figure 1. Scatterplot demonstrating the relationship between 2009 TFI values and the percentage of schools in sample nations who report that assessment data are used to make decisions about students' retention or promotion.

Assessment data are compared to national performance levels and 2009 TFI. Figure 2 shows the 2009 TFI values and the percentage of schools that reported that assessment data was compared to national performance levels. The scatterplot indicates that the relationship is linear, that a positive association between the two variables, and a moderate correlation exists between the two variables.

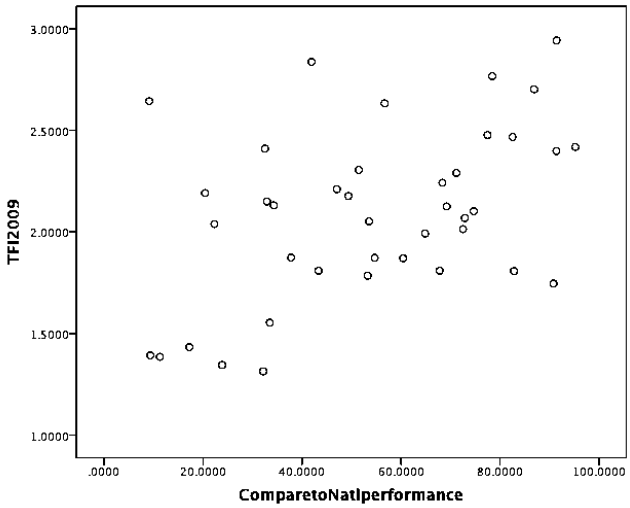

Figure 2. Scatterplot demonstrating the relationship between 2009 TFI values and the percentage of schools in sample nations who report that assessment data are compared to national performance levels.

Assessment data are used to assess teacher effectiveness and 2009 TFI. Figure 3 indicates the 2009 TFI values and the percentage of schools that reported that assessment data was used to judge teacher effectiveness. The scatterplot shows that the relationship is linear, that there is a positive association between the two variables, and that a weak correlation exists between the two variables.

Assessment data are compared to other schools and 2009 TFI. Figure 4 shows the 2009 TFI values and the percentage of schools that reported that assessment data was compared to other schools. The scatterplot indicates that the relationship is linear, that there is a positive association between the two variables, and that a moderate correlation exists between the two variables.

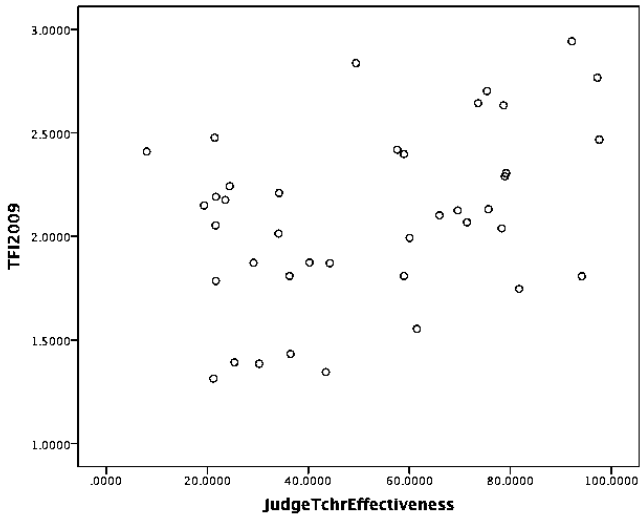

Figure 3. Scatterplot demonstrating the relationship between 2009 TFI values and the percentage of schools in sample nations who report that assessment data are used to make judgments about teachers' effectiveness.

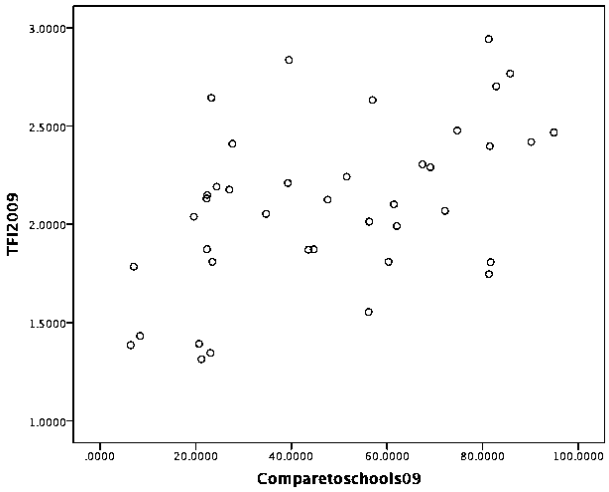

Figure 4. Scatterplot demonstrating the relationship between 2009 TFI values and the percentage of schools in sample nations who report that assessment data are compared to other schools.

Assessment data are made public and 2009 TFI. Figure 5 details the 2009 TFI values and the percentage of schools that reported that assessment data was made public (e.g., in the mass media). The scatterplot indicates that the relationship is linear, that

there is a positive association between the two variables, and that a moderate to weak correlation exists between the two variables.

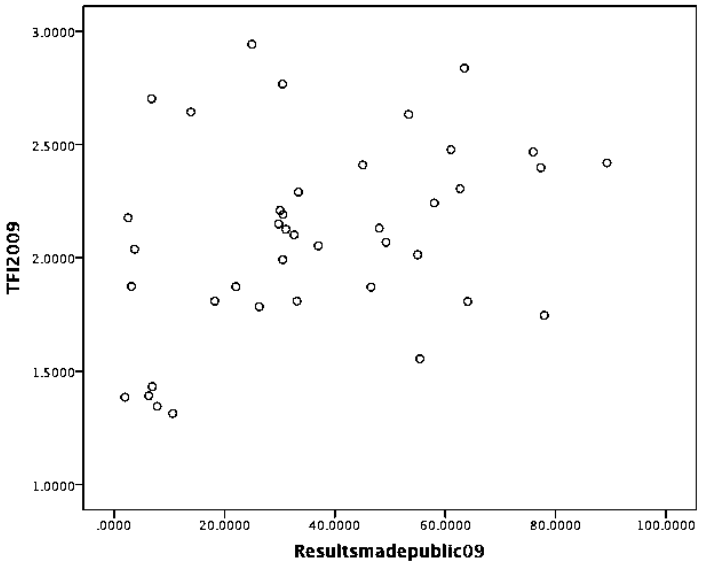

Figure 5. Scatterplot demonstrating the relationship between 2009 TFI and the percentage of schools in sample nations who report that achievement data are posted publicly (e.g., in the media).

Assessment data are used to evaluate principal performance and 2009 TFI.

Figure 6 shows the 2009 TFI values and the percentage of schools that reported that assessment data was used to evaluate principal performance. The scatterplot indicates that the relationship is linear, that there is a positive association between the two variables, and that a weak to moderate correlation exists between the two variables.

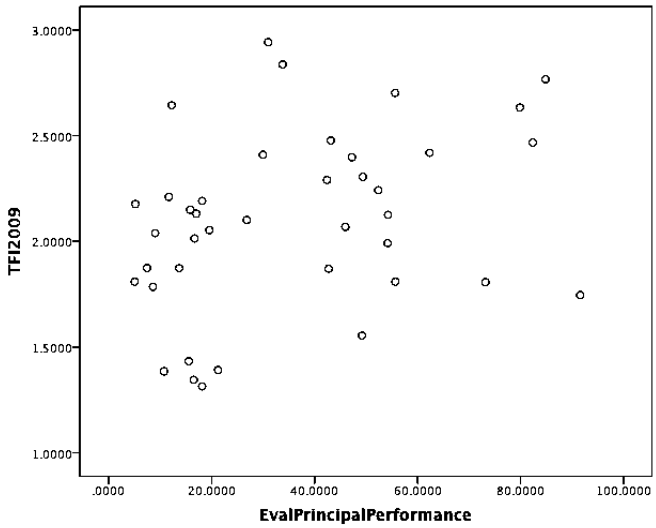

Figure 6. Scatterplot demonstrating the relationship between 2009 TFI and the percentage of schools in sample nations who report that achievement data are used to evaluate principals' performance.

Assessment data are used to evaluate teacher performance and 2009 TFI. Figure 7 shows the 2009 TFI values and the percentage of schools that reported that assessment data was used to evaluate teacher performance. The scatterplot indicates that the relationship is linear, that there is a positive association between the two variables, and that a weak correlation exists between the two variables.

Assessment data are used to make decisions about instructional resource allocation and 2009 TFI. Figure 8 shows the 2009 TFI values and the percentage of schools that reported that assessment data was used to make decisions about instructional resource allocation. The scatterplot indicates that the relationship is linear, that there is a positive association between the two variables, and that a weak correlation exists between the two variables.

Correlation values between attribute variables and 2009 TFI. To investigate the strength of the association between the listed independent attribute variables and the

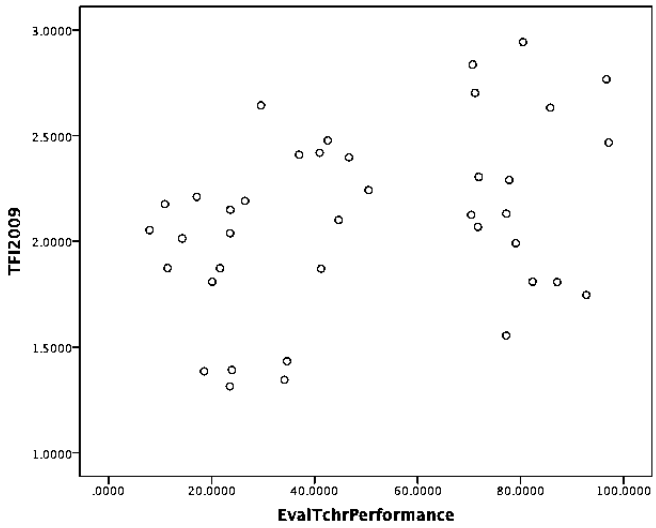

Figure 7. Scatterplot demonstrating the relationship between 2009 TFI and the percentage of schools in sample nations who report that achievement data are used to evaluate teachers' performance.

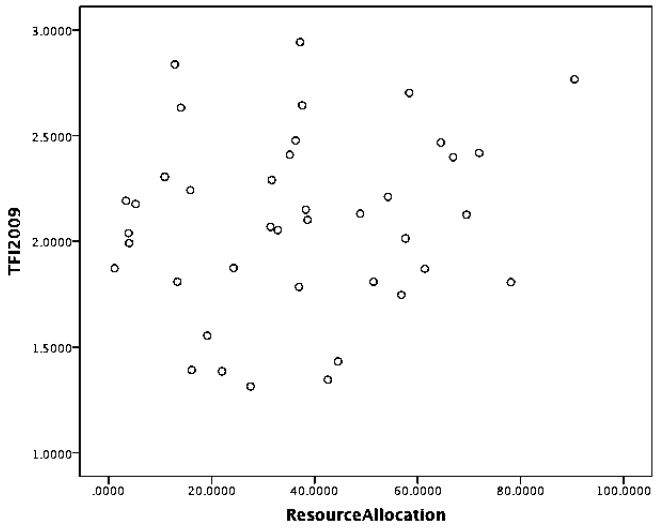

Figure 8. Scatterplot demonstrating the relationship between 2009 TFI and the percentage of schools in sample nations who report that achievement data are used to make decisions about instructional resource allocation to the school.

2009 TFI, each of the independent attribute variables was skewed negatively with the exception of "assessment data is used to judge teacher effectiveness," and " achievement data are used in evaluation of the principal's performance" which were both positively

skewed. As such, the assumption of normality was violated for each computation. Thus, the Spearman rho statistic was used with each variable.

There was a range of variances present in the data. Four of the correlations were not statistically significant from zero. In those instances, there is a high probability that the TFI was affected as a result of chance rather than the attribute variable. Those correlations include "assessment data is used make decisions about students' retention or promotion" (p =.951); "assessment data is used to make judgments about teachers' effectiveness" (p = .070); "achievement data are used in evaluation of teachers' performance" (p = .057); and "achievement data are used in decisions about instructional resource allocation to the school" (p = .473). Four of the correlations were statistically significant from zero, meaning that there is a chance that changes to the 2009 TFI came as a result of a relationship with the attribute variable. Those correlations include: "assessment data is compared to national performance levels" (p = .012); "assessment data is compared to other schools" (p = .001); "assessment data is posted public (e.g., in the media)" (p = .048); and "achievement data are used in evaluation of the principal's performance" (p = .046).

In each of the four statistically significant correlations, the direction of the correlation was positive, which means that in the sample nations where those variables exist, students tend to be tested more frequently. Using Cohen's (1988) guidelines, the effect sizes are typical for "assessment data is compared to national performance levels", "assessment data is posted public (e.g., in the media)" and "achievement data are used in evaluation of the principal's performance." The effect size was larger than typical for

Table 4

Correlation Values between independent attribute variables and 2009 TFI[a]

Variable	Correlation Coefficient (Spearman's rho)	Sig. (2-tailed)
Assessment data is used make decisions about students' retention or promotion.	-.010	.951
Assessment data is used to compare the school to national performance	.392	.012*
Assessment data is used to make judgments about teachers' effectiveness	.289	.070
Assessment data is compared to other schools.	.508	.001**
Achievement Data are posted publicly (e.g., media).	.315	.048*
Achievement data are used in evaluation of the principal's performance.	.318	.046*
Achievement data are used in evaluation of teachers' performance	.307	.057
Achievement data are used in decisions about instructional resource allocation to the school.	.117	.473

Note. *p < .05. **p < .001

a. Listwise N=40

"assessment data is compared to other schools". Consequently, the null hypothesis for research question 1 is rejected.

Research question 2: Is the difference in correlational strength between standardized testing frequency and academic achievement between 2003 and 2009 among sample nations statistically significant from zero? The second research question was designed to determine if the change in correlational strength (if one exists) between testing frequency and academic achievement is significant. In order to answer the second research question, correlations were conducted that examined the association between the 2003 PISA math, reading and science scores in sample nations and the nations' 2003 TFI. Table 5 displays the 2003 correlation values. The same process was done for 2009 PISA math, reading and science scores and the nations' 2009 TFI. The 2009 correlation values can be seen in Table 6.

Correlation values between 2003 PISA math, reading and science means and 2003 TFI. To investigate if the association was statistically significant from zero between the 2003 TFI among sample nations and the corresponding means for the 2003 PISA math, reading and science scores, a correlation was computed. Every variable was skewed, which violated the assumption of normality. Thus, the Spearman rho statistic was calculated in all three computations. In all cases, the resulting correlation between the 2003 TFI and 2003 means for PISA math, reading and science assessments were smaller than typical (r = -.113, .033 and -.101) and none were statistically significant from zero (p = .507, .846, .554).

Table 5

Correlation values between 2003 PISA math, reading and science means and 2003 TFI[a]

PISA Assessment	Correlation Coefficient (Spearman's rho)	Sig. (2-Tailed)
Math	-.113	.507
Reading	.033	.846
Science	-.101	.554

Note. a. Listwise N=37

Table 6

Correlation values between 2009 PISA math, reading and science means and 2009 TFI[a]

PISA Assessment	Correlation Coefficient (Spearman's rho)	Sig. (2-Tailed)
Math	-.100	.540
Reading	-.022	.891
Science	-.103	.526

Note. a. Listwise N = 40

Research question #2 summary. Because each of the variables was skewed to some degree, the assumption of normality was violated. Accordingly, the Spearman rho statistic was calculated in all three correlations. As was the case in 2003, the correlations between the three 2009 PISA assessment mean scores and the 2009 TFI were not significant (p = 540, .891, .526). Additionally, all correlations were best described as weak (r = -.100, -.022 and -.103).

Fisher's *r* to *z* transformation. The final step in answering the second research question is to determine if the difference between the correlation values in 2003 and 2009 is statistically significant from zero or if the change occurred more likely as a result of

chance. To answer that question, the correlation values between the 2003 PISA math test and 2003 TFI as well as the 2009 PISA math test and the 2009 TFI a Fisher's *r* to *z* transformation was conducted. The calculation has one assumption, that the two means come from independent groups, and that assumption was not violated in this case.

Z-Scores assessed the significance of the difference between the 2003 and 2009 PISA math test. The same process was conducted for the 2003 and 2009 reading and science tests as well. The null hypothesis of the test is that there is no significant difference between the 2003 and the 2009 correlations. Figures 9 through 11 are images of the online tool with data and results shown.

Fisher's *r* to *z* transformation for 2003 and 2009 PISA math scores and TFIs. A *z* score was computed for the difference between the correlation strength of PISA math scores and TFI in 2003 and in 2009. Figure 9 shows the result of that calculation, which was $z = -0.06$. This *z* score tells us that the difference was not statistically different from zero at the .05 level of confidence. The non-significant statistical analysis ($p = .48$) comparing the two *z*-scores showed that the null hypothesis for PISA math data and the TFIs was retained. Hence sampling error is a plausible explanation for the difference between the two correlations. Figure 9 shows that the PISA math and the TFIs were not significant correlated.

Fisher's *r* to *z* transformation for 2003 and 2009 PISA reading scores and TFIs. A *z*-score was computed for the difference between the correlation strength of

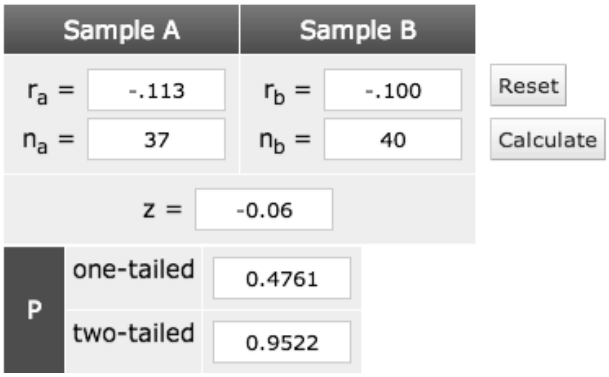

Figure 9. Image of online tool found at http://vassarstats.net/rdiff.html with results from Fisher's *r* to *z* transformation for 2003 and 2009 PISA math scores and TFIs.

PISA reading scores and TFI in 2003 and in 2009. Figure 10 shows the result of that calculation which was $z = 0.23$. The statistical analysis using the two *z*-scores tells us that the difference was non-significant ($p = .41$) and the null hypothesis between PISA reading data and TFIs was not rejected. Thus, sampling error or pure chance are plausible explanations for the difference between the two correlations. Figure 10 confirms that the PISA reading data and the TFIs were not statistically correlated.

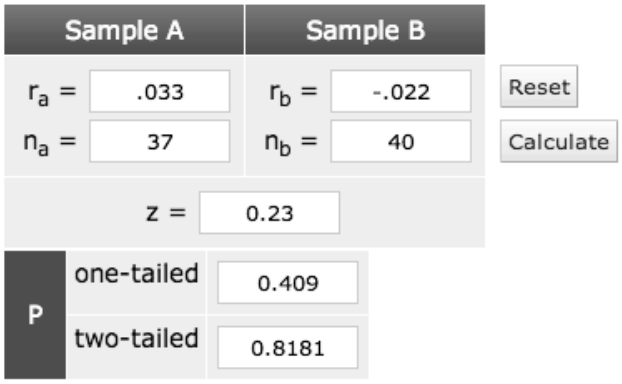

Figure 10. Image of online tool found at http://vassarstats.net/rdiff.html with results from Fisher's *r* to *z* transformation for 2003 and 2009 PISA reading scores and TFIs.

Fisher's *r* to *z* transformation for 2003 and 2009 PISA science scores and TFIs. A z-score was computed for the difference between the correlation strength of PISA science scores and TFI in 2003 and in 2009. Figure 11 shows the result of that calculation, which was $z = 0.01$. The statistical analysis using the two z score showed us that the difference was non-significant ($p = .50$) and the null hypothesis between the PISA Science and TFIs was not rejected. Thus, sampling error is a plausible explanation for the difference between the two correlations. Figure 11 illustrates that the PISA science data and the TFIs were not significantly correlated.

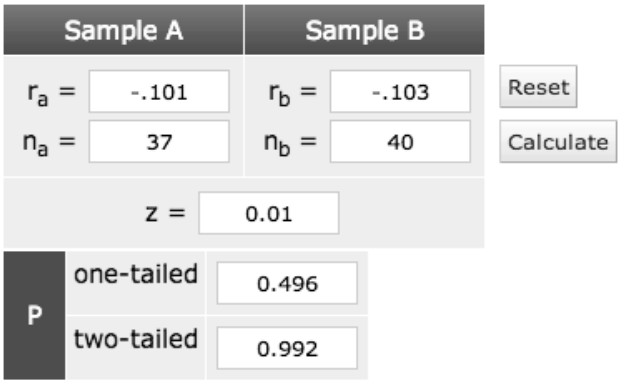

Figure 11. Image of online tool found at http://vassarstats.net/rdiff.html with results from Fisher's *r* to *z* transformation for 2003 and 2009 PISA science scores and TFIs.

different from zero at the .05 level of confidence, hence sampling error is a plausible explanation for the difference between the two correlations.

According to Kenny (1979), if z scores are above or equal to 1.96 or below or equal -1.96, the correlations are significantly different from zero at the .05 level of significance. None of the z scores in Table 7 meet those criteria and as a result are not considered significant at the .05 level. Therefore, the null hypothesis that the difference between the 2003 and the 2009 is not significant is not rejected.

CHAPTER V

DISCUSSION

The purpose of this study is to examine the association between the frequency of standardized testing and academic achievement. Using a sequential explanatory theory, my study explored two research questions: (a) why some countries require their students to take more standardized tests than other countries; and (b) whether the change in correlational strength between testing frequency and academic achievement between 2003 and 2009 among sample nations was statistically significant.

In regard to the first research question, it is hypothesized that countries that attach consequences, or stakes, to the results of standardized tests will engage in standardized more frequently. In regard to the second research question, whether the change in correlational strength between testing frequency and academic achievement between 2003 and 2009 is significant, it is hypothesized that the results will indicate no statistical significance.

Summary of Findings

Research question #1. All variables from the 2009 PISA school questionnaire that could reasonably be considered some form of consequence that could be attached to achievement results were selected as potential factors or conditions that lead to higher levels of testing frequency. A correlation analysis using the data from 40 sample nations and the findings of that analysis show that a statistically significant association exists between four variables and how frequently nations required 15-year old students to engage in standardized testing in 2009. As such, the null hypothesis was rejected. The

variables include: (a) data is used to compare the school to national performance; (b) data is used to compare the school to other schools; (c) achievement data is posted publicly (e.g. in the media); and (d) achievement data is used in the evaluation of the principal. In other words, in the sample nations where these conditions exist at high levels, there are also high levels of testing frequency for fifteen-year old students.

Research question #2: The second research question explores the association between testing frequency and academic achievement. To do so, a testing frequency index (TFI) has been generated for each sample nation for 2003 and 2009. The TFI is based on 2003 and 2009 PISA school questionnaire data that asks schools to indicate how frequently 15 year-old students are required to standardized tests. A correlation coefficient has been calculated to measure the strength of association between each nation's 2003 TFI and its 2003 PISA reading, math and science means. The same process has been used using 2009 data.

Subsequently, using an online tool found at vassarstats.net, the 2003 and 2009 *r* coefficients were transformed into *z* scores in order to determine if the change in correlation strength from 2003 to 2009 was statistically significant from zero. Results from that calculation indicate that the change in correlational strength is not significant. Therefore, the null hypothesis is not rejected.

Limitations

This study acknowledges limitations that decrease the generalizability of the results. This section describes three general limitations: two limitations affecting sample size and another limitation related to the test data gathered.

The first two limitations are related to the sample used in the study. The number of nations selected as participants in the study was limited to only those nations who participated in both the 2003 and the 2009 PISA. As mentioned previously, several nations did not have a complete set of data. During the 2003 PISA administration, achievement test data for the United Kingdom were not included in the final results because it was determined that sampling standards had not been met (OECD, 2003). Additionally, the Chinese provinces of Hong Kong and Macao did not provide testing frequency data in 2003 (ibid.); consequently, a 2003 TFI could not be calculated. Finally, Thailand and Lichtenstein did not provide data on response level 4 or 5 in the 2003 and 2009 testing frequency question. In all cases, a listwise deletion strategy was used when appropriate. Given the small number of countries who lack a complete set of data, I do not believe the final conclusions of the present data have been compromised. However, if future studies occur with similar research questions, authors may want to keep this limitation in mind for limitation purposes or to design research in such a way as to include the aforementioned countries.

Additionally, this research is limited to the study of a certain age group of students. PISA is an assessment of fifteen-year old students. Surveys that are completed by school leaders are done so with fifteen-year old students in mind. Had survey questions been asked about younger or older students, it is possible responses could have been different, thus yielding differing conclusions. Further, PISA achievement tests are designed with what is deemed to be an appropriate level of content for a fifteen year-old student. Combined, these reasons make it difficult to generalize any results beyond fifteen year-old students.

Finally, the present study's definition of academic achievement is limited to the assessments that are offered in each PISA administration. PISA administrations include reading, math and science assessments that consist of multiple-choice questions. There are large numbers of ways to measure academic achievement that vary both in content and in method, and no assumption is made that results of one set of achievement tests equate to the results of another.

Explanation of Findings Utilizing the United States as an Example

To contextualize the results of the first research question, I explore the findings in relation to the United States. The frequency of standardized testing in the United States (as reported on the 2003 and 2009 PISA school questionnaire by participating schools) went up by 10.71% (see Table 3). One possible explanation for that increase is the fact that the United States exhibits high levels of certain conditions that are statistically associated high testing levels. Those conditions are "data is used to compare the schools to national performance;" "data is used to compare to the schools to other schools;" "student achievement data is posted publicly;" and "student achievement data is used in the principal's evaluation." Findings suggest three reasons why the United States ranks high in these categories when compared to the other sample nations in this study: the effects of No Child Left Behind, the volume of tests that are frequently administered to freshmen in American high schools, and the final reason are similar policies that exist in multiple states that require student achievement data be used in the evaluation of principals.

An analysis of policies, laws and other pertinent documents pertaining to American education reveal that there are three answers to these questions: (a) because

the No Child Behind Law required them to do so; (b) because of an increase in the number of tests 15-year old student are required to take; and, specifically to the fourth question, (c) because many states require that achievement data be used in the evaluation of principals.

Before proceeding, it should be noted that in the United States, fifteen-year olds are typically in the ninth grade, or freshman year of high school. In most cases, the ninth grade is the first year of high school, but there are still some locations that include the ninth grade year as the final year of junior high school (NCES, 2012). Additionally, it is important to point out that PISA developers do not specify what kinds of achievement data or assessments they refer to in their school questionnaires (see appendices A through D). As such, respondents may answer in the affirmative for a wide variety of reasons.

According to the 2009 PISA school questionnaire, 95.22% of participating schools in the United States said yes when asked if their school's performance is compared to a national performance (OECD, 2010). That ranked first among the 40 sample nations. In the same survey, 90.17% of participating schools in those same schools said yes when asked if their school's achievement data is used to compare the school to other schools (ibid.). That data ranked second among sample nations. Finally, 89.32% of those same schools report that their school's achievement data is made public through the media. That ranks first among all sample nations.

The United States ranks high compared to other sample nations in the use of data to compare a school to national performance as well as to other schools, and also ranks high in reporting data publicly largely because in 2009 schools in the United States were

required by federal law to do so. NCLB required all school systems to post disaggregated achievement data publicly each year. Because data for all school systems, and more often than not for each school, was posted publicly on an annual basis, schools could easily be compared to one another. Additionally, in most cases, school or district achievement data was posted in such a way as to easily compare it to state and national data, thus making it easy to compare school data to national means.

According to the 2001 law, all states and Local Education Authorities (LEAs; e.g., districts) had to publicly post student achievement data (NCLB, 2003). Across the 50 states, report cards varied in style and composition, but the law required that certain information appear on every card (see Appendix E for an example). Data had to be presented in disaggregate form, showing data for student sub-populations (e.g., racial subgroups, special education students, English language learners, etc.). Additionally, each report card was required to contain a section that indicated whether the school had made adequately yearly progress (AYP), a certain level of growth that was meant to push the school to the law's ultimate goal of 100% mastery for all students in reading and math by 2014 (NCLB, 2003).

In many states, comparisons of schools are available on the report card. For example, the report cards that were used in Oregon in 2009 listed every school in the district on the front page with their NCLB classification (see Appendix E). The specific requirements of a law that effects every state in the country offers one clear explanation as to why these three categories ("data is used to compare the school to national performance;" "data is used to compare the school to other schools;" and "data is posted publicly") rank as high as they do in the United States.

Another factor in the answer to these questions is the sheer volume of tests that are administered to American fifteen year-olds, whose results are either publicly posted or are used for comparative purposes are required to take. While the mandated accountability tests that are required in all fifty states that generate the data posted on required report cards are not the only tests American ninth graders are required to take. End-of-course tests, Advanced Placement Tests, tests to assess college readiness, and others are regularly given to ninth graders and all create data that is either posted publicly or can be compared to other schools or to a national averages.

According to the National Center for Fair and Open Testing (2009), at the time the 2009 PISA school questionnaire was filled out by schools, there were already a handful of states that were requiring ninth grade students to take an end-of-course exam in one or more of their academic core classes (English, math, science social studies, etc.). Depending on the state, results from those tests at the least affected the grade in the class, but in some of the states, determined whether a student would graduate or not (National Center for Fair and Open Testing, 2009). In some states, results from those tests are only shared in aggregate form at the state level, in others the end-of-course test results are not only posted publicly as is the case in Washington but are presented in such a way as to allow for easy school to school comparison (State of Washington Office of Superintendent of Public Instruction, 2014

Another test that students may opt to take that may lead to a greater likelihood that schools would believe their ninth grade assessment data is used for comparative purposes or is posted publicly is the Advanced Placement Test. While the majority of American students who participate in Advanced Placement (A.P.) classes and testing are

seniors (1,422,635 students in 2009), there are still a high number of freshmen who also take the exams. According to the College Board, in 2009 43,454 ninth grade students took A.P. exams (College Board, 2009). While A.P. test results are not reported publicly in every district or in every state, they do offer local and national comparative information on the score reports that are available to school personnel (College Board, 2012).

The final research question focuses on why The United States ranks high in the use of achievement data in principal evaluations. Of the schools in the United States who responded to the 2009 PISA school questionnaire, 62.31% responded affirmatively that achievement data is used in the evaluation of the principal's performance. Among the 40 sample nations, the United States ranks sixth in this category.

Requiring the use of student achievement data as a part of a principal's evaluation is a practice that has become common since the onset of the standards and accountability movement in the United States. According to Linn (2000), in the late 1980s and early 1990s, school systems began to use test scores for accountability purposes. Linn states "Accountability programs took a variety of forms, but shared the common characteristic that they increased real or perceived stakes of results for teachers and educational administrators" (p. 7). Attaching student achievement scores to an evaluation is an example of that phenomenon.

As is the case with many matters pertaining to education, individual states have the authority to make decisions on how school administrators are to be evaluated. Consequently, according to Goldring, et al., (2009), there is no single method of principal

evaluation, contributing to wide variability across the country in how principals are evaluated. In that same study, the authors conducted an analysis of principal evaluation instruments in 43 states and the District of Columbia. In their analysis, they point out problems in the use of outcome-based performance (using achievement data to assess principal performance). Suggesting a reason why more states and districts don't base principal evaluation on school outcomes, they state, "Although this approach seems to be better aligned with performance accountability…it faces methodological hurdles, especially in assuming direct causal relations between what the principal does and school outcomes" (p. 22).

Despite this, many districts at the time the 2009 PISA school questionnaire was completed required the use of student performance data in principal evaluation even though their state department of education had not required it. For example, in a study by the Institute of Education Sciences it is reported that in 2010, 68% of the 1,013 (693) participating school districts in California report that achievement data is used in the evaluation of principals (White, Makkonen, Vince & Bailey, 2010).

In some states, there is a requirement to include student performance in principal evaluation. For instance, in 2008 Ohio adopted the Ohio Principal Evaluation System (OPES). According to the Ohio Department of Education website, OPES requires principal evaluations to consist of two components: principal performance on standards and student growth measures (Principal Evaluations, 2014). While Ohio districts have some freedom to choose what assessment data to use, the must include data in the evaluation.

Similar to Ohio, North Carolina's principal evaluation system, which has been in place since 2008, also requires student achievement data to be used as a part of the process. However, the North Carolina policy states that district superintendents, in cooperation with principal to be evaluated, collectively select the data to be used in the evaluation (North Carolina School Executive, 2008)

My findings support the idea that the reason why schools in the United States require its students to test so frequently is because its educational "terrain" (Blake, 2012, p. 5) is well suited for the existence of certain conditions that have a significant correlation with testing frequency. Again, the conditions that were identified in the first research question include: a) data is used to compare the school to national performance; (b) data is used to compare the school to other schools; (c) achievement data is posted publicly (e.g. in the media); and (d) achievement data is used in the evaluation of the principal.

Implications

Among sample nations, the mean change in testing frequency from 2003 to 2009 went up 4.59%. In some nations, the increase was more drastic. Indonesia, for instance, saw an increase in testing frequency of 47.88%, the Netherlands saw an increase of 27.49%, and Portugal saw an increase of 21.05%. The increase in standardized testing represents not only a substantial investment in the financial resources required to administer more tests, but also an investment in commitment (political and pedagogical) to the notion that more frequent testing will lead to increased academic progress. The

latter of these two concepts is what this study has focused on, and offers some evidence that suggests that the logic is perhaps faulty.

In regard to the first research question, as to why some nations test more frequently than others, existing literature suggests that the presence of stakes or consequences lead to unintended consequences (Amrein & Berliner, 2003: Au, 2007; McNeil, 2000; Nichols, et al. 2006; Popham, 1999; Wang, et al. 2006). It is hypothesized that one unintended consequence of the attachment of stakes to testing results might be increased testing frequency. Establishing an association between the presence of stakes and elevated testing frequency would allow for a system to identify nations where testing frequency might be higher than normal. Knowing that certain practices lead to an increase in testing frequency is information that nations might be able to use to assist in the development of educational policies.

The second research question examines the association between testing frequency and academic achievement, as measured by PISA math, reading and science test scores. While existing literature on the subject of standardized testing offers ample research on the topic of standardized testing, there currently exists no study that examines the effects of the frequency of standardized testing.

Given the high cost of standardized testing (Chingos, 2012) and the fact that the frequency of standardized testing increased from 2003 to 2009 (see Table 3), from the standpoint of developing educational policy, determining whether there is a positive association between the frequency of testing and academic achievement would seem to be an important piece of information. Whether standardized testing has an appropriate

role in education at all is a question that is beyond the scope of this study. However, it does question the effect that increased frequency of standardized testing has on academic achievement. Evidence has been offered that suggests the change in correlational strength between the frequency of standardized testing, as reported by schools in sample nations, and academic achievement, as measured by the results of the PISA math, reading and science tests from 2003 to 2009 is not statistically significant from zero. Thus, there is a high probability that any variation in the data occurred as a result of chance.

Based upon the previous section where I utilized US as an example for my quantitative findings, I surmised that a primary reason why the US ranks so high (as compared to the other sample nations) in the listed conditions has to do with a pervasive educational philosophy that is based on educational accountability. Elmore (2002) defines educational accountability as:

> Broad-based and politically persistent over time. It involves state legislators, governors, advocacy groups and professional organizations. It stems from the belief that schools, like other public and private organizations in society, should be able to demonstrate what they contribute to the learning of students and that they should engage in steady improvement of practice and performance over time (p. 3).

Since the inception of the educational accountability movement, and certainly during its high water mark with the passage of NCLB in 2002, American schools have been asked to publicly validate their effectiveness on an annual basis. McDonnell (2008) points out that the primary method for that validation is through the use of standardized tests. She

offers four reasons for the trend: (a) in modern America, testing is a trusted method of measuring many aspects of our society; (b) testing has also been in use in education for many decades, and especially over the past thirty years; (c) for higher levels of government, testing carries with it a relatively low-cost; and (d) testing is seen as a method of measurement that would allow policy makers to avoid "excessive regulation and micromanagement" (p. 50).

Current literature consistently states that current national and local educational policies in the United States have reflected and will continue to reflect an emphasis on accountability (Linn, 2000; McDonnell, 2008; Shepherd, 2008; Wang, et al., 2006). As long as the conditions discussed in the first research question exist, that is, as long as stakes are attached to educational tests, I contend there will be a trend among schools to test more frequently. There is much riding on the outcomes of school testing data (Gunzenhauser, 2003). And to keep data as high as possible, school leaders will continue to seek methods – presumably within the confines of the policies that guide their decisions - to increase test scores. One such method is to increase the frequency of testing. Whether it's believed to be a strategy to make students more comfortable with the tests that are being used, to take advantage of policies that allow for multiple testing attempts to raise low scores, or for the misguided belief that somehow more frequent testing in and of itself leads to a deeper understanding of academic content, there's no denying the fact that levels of testing frequency continue to go up in the United States and abroad.

Of the six categories that were used, the United States ranked first or second in three of them. However, it ranked 9th and 11th in the 2009 TFI and the percentage

increase from the 2003 to the 2009 TFI, respectively. Clearly, increased testing frequency is not an issue that is limited to the borders of the United States. It is my speculation that should further analysis like this one be conduced in order to examine other countries, they would reveal conditions similar the ones found in the United States. That is, the presence of policies that require the public posting of data, data used in comparative ways, and data used to evaluate school leaders, in addition to a variety of tests that are regularly administered.

The second question goes right to the heart of the effect that testing frequency has on academic achievement. Findings from this research demonstrate that the change in correlational strength between testing frequency and academic achievement from 2003 to 2009 are not statistically significant from zero. In other words, as testing frequency changes, it cannot be dismissed that any similar change in academic achievement has occurred as a result of pure chance. Further, it has been established that testing frequency is often a by-product of educational policy. Therefore, it would seem logical that policy-makers in countries around the world would want to develop education policies that lead to practices that produce positive educational results.

Conclusion

Large-scale standardized tests can be effective tools to determine if students are learning. However, the results of this research show that increasing the frequency of testing may not - directly - result in higher academic achievement. The educational "terrain" (Blake, 2012) of certain countries (especially the U.S.) is well suited for conditions that are associated with elevated levels of testing frequency, such as the

presence of certain forms of stakes that are attached to test results. In light of this research, those who write educational policies that involve large-scale standardized tests should do two things: (a) consider the possible effects that attaching stakes to such tests have; and (b) they should examine whether the ends, that is, questionable increases in academic achievement, justify the means, the increased costs of more frequent use large-scale standardized testing.

APPENDIX A

QUESTION 12 2003 PISA

Q12 Generally, in your school, how often are <15-year-old> students assessed using:

(Please <tick> only one box in each row.)

	Never	*1 – 2 times a year*	*3 – 5 times a year*	*Monthly*	*More than once a month*
a) Standardised tests?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b) Teacher-developed tests?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c) Teachers' judgmental ratings?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d) Student <portfolios>?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e) Student assignments/ projects/homework?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

APPENDIX B

QUESTION 15 2009 PISA

Q15 Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using the following methods?

(Please tick only one box in each row)

	Never	*1 – 2 times a year*	*3 – 5 times a year*	*Monthly*	*More than once a month*
a) Standardised tests	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b) Teacher-developed tests	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c) Teachers' judgmental ratings	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d) Student <portfolios>	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e) Student assignments/ projects/homework	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

APPENDIX C

QUESTION 16 2009 PISA

Q16 In your school, are assessments of students in <national modal grade for 15-year-olds> used for any of the following purposes?

(Please tick only one box in each row)

	Yes	*No*
a) To inform parents about their child's progress	☐ 1	☐ 2
b) To make decisions about students' retention or promotion	☐ 1	☐ 2
c) To group students for instructional purposes	☐ 1	☐ 2
d) To compare the school to <district or national> performance	☐ 1	☐ 2
e) To monitor the school's progress from year to year	☐ 1	☐ 2
f) To make judgements about teachers' effectiveness	☐ 1	☐ 2
g) To identify aspects of instruction or the curriculum that could be improved	☐ 1	☐ 2
h) To compare the school with other schools	☐ 1	☐ 2

APPENDIX D

QUESTION 22 2009 PISA

Q22 In your school, are achievement data used in any of the following <accountability procedures>?

Achievement data include aggregated school or grade-level test scores or grades, or graduation rates.

(Please tick one box in each row)

	Yes	*No*
a) Achievement data are posted publicly (e.g. in the media)	☐ 1	☐ 2
b) Achievement data are used in evaluation of the principal's performance	☐ 1	☐ 2
c) Achievement data are used in evaluation of teachers' performance	☐ 1	☐ 2
d) Achievement data are used in decisions about instructional resource allocation to the school	☐ 1	☐ 2
e) Achievement data are tracked over time by an administrative authority	☐ 1	☐ 2

APPENDIX E

2008-2009 BEAVERTON SCHOOL DISTRICT REPORT CARD

ACCOUNTABILITY INFORMATION REQUIRED BY THE FEDERAL NO CHILD LEFT BEHIND ACT

The information below is used to determine the Adequate Yearly Progress designation for your district. A district is designated as *Not Meeting AYP* if any indicator is determined to be *Not Met*. The Student Achievement and Student Participation ratings are based on 2007-2008 and 2008-2009 Oregon Statewide Assessments for the students in your district identified as enrolled for a full academic year. The statewide goal for the minimum percentage of students expected to meet or exceed standards is 60% in English/Language Arts and 59% in Mathematics. Student Participation is expected to be 95% or greater. The statewide goal for the minimum graduation rate is 68.1%. The statewide goal for the minimum attendance rate is 92.0%. For more information, please view documents at **www.ode.state.or.us/data/reportcard/reports.aspx**

STUDENT GROUP	STUDENT ACHIEVEMENT		STUDENT PARTICIPATION		
Race/Ethnicity	ENGLISH/LANGUAGE ARTS	MATHEMATICS	ENGLISH/LANGUAGE ARTS	MATHEMATICS	GRADUATION
Am. Indian/Alaskan Native	MET	MET	MET	MET	NOT MET
Asian/Pacific Islander	MET	MET	MET	MET	MET
Black (not of Hispanic origin)	MET	MET	MET	MET	MET
Hispanic	NOT MET	NOT MET	MET	MET	NOT MET
White (not of Hispanic origin)	MET	MET	MET	MET	MET
Multi-Racial/Multi-Ethnic	MET	MET	MET	MET	MET
Students with Disabilities	NOT MET	NOT MET	MET	MET	MET
Limited English Proficient	NOT MET	NOT MET	MET	MET	MET
Economically Disadvantaged	NOT MET	NOT MET	MET	MET	MET
All Students	MET	MET	MET	MET	MET

NA Too few test scores or students to determine a rating.

The National Assessment of Education Progress (NAEP) is only in grades 4, 8, and 12. NAEP results are based on representative samples of students and NAEP does not provide results for individual students, schools, or districts in Oregon. The table below lists the most recently available NAEP data for reading and mathematics. Small differences between results for Oregon and the U.S. may not be statistically significant. For more information, see http://nces.ed.gov/nationsreportcard.

2007 NAEP RESULTS

			Advanced %	Proficient %	Basic %	Below Basic %	Participation Rates: Students with Disabilities	Participation Rates: English Language Learners
Reading	Grade 4	Oregon	6	22	34	38	72	85
		United States	7	24	34	34	66	80
	Grade 8	Oregon	3	31	43	23	77	86
		United States	2	27	43	27	66	77
Math	Grade 4	Oregon	4	31	44	21	85	93
		United States	5	33	43	19	80	92
	Grade 8	Oregon	9	26	38	27	76	90
		United States	7	24	39	30	70	89

DISTRICT INFORMATION

Financial Data

General Fund Expenditures

The table below shows dollars spent per student by your district for the 2007-2008 school year. For more information, visit the Database Initiative Project website: www.ode.state.or.us/data/reports/toc.aspx

Education Service District	District	State
ESD Support Per Student	$323	$392

General Fund	District	State
Direct Classroom	$4,520	$4,540
Classroom Support	$1,568	$1,389
Building Support	$1,210	$1,411
Central Support	$170	$383

Bond Levy / Local Option	Number of Elections	Election Result: Yes	Election Result: No
Election Year: 2008	0	0	0
Election Year: 2007	0	0	0
Election Year: 2006	2	1	1

2008-2009 District Report Card

BEAVERTON
SCHOOL DISTRICT

Dear Parents and Community Members, November 10, 2009

The Oregon Department of Education is proud to issue the 11th annual school Report Card. This year's report cards include two significant changes: a simplified rating system for schools and a new description of how much students have learned from year to year called a "Growth Model." The new rating system uses three categories: Outstanding, Satisfactory and In Need of Improvement. These ratings cannot tell you everything about your school but are a good starting point for talking about our successes and opportunities for improvement.

Federal Adequate Yearly Progress Rating: NOT MET

☐ MET ☒ DID NOT MEET

See rating details on back page

☐ Identified for District Improvement

Susan Castillo, State Superintendent of Public Instruction

SCHOOL RATINGS

SCHOOLS	Overall	School Improvement Status
Aloha High School	In Need of Improvement	
Aloha-Huber Park School	Satisfactory	
Arts & Communication High School	Outstanding	
Arts & Communication Middle	Outstanding	
Barnes Elementary School	Satisfactory	
Beaver Acres Elementary School	Outstanding	
Beaverton High School	In Need of Improvement	
Bethany Elementary School	Outstanding	
Bonny Slope Elementary School	Not Rated	
Cedar Mill Elementary School	Outstanding	
Cedar Park Middle School	Satisfactory	
Chehalem Elementary School	Satisfactory	
Community School	In Need of Improvement	
Conestoga Middle School	Satisfactory	
Cooper Mountain Elementary School	Outstanding	
Elmonica Elementary School	Satisfactory	
Errol Hassell Elementary School	Outstanding	
Findley Elementary	Outstanding	
Fir Grove Elementary School	Satisfactory	
Five Oaks Middle School	Satisfactory	
Greenway Elementary School	Satisfactory	
Hazeldale Elementary School	Satisfactory	
Health & Science School	Not Rated	
Highland Park Middle School	Satisfactory	
Hiteon Elementary School	Outstanding	

Continued on last page.

FEDERAL DESIGNATION FOR TITLE 1 SCHOOLS	DISTRICT		STATE	
REQUIRED BY THE FEDERAL NO CHILD LEFT BEHIND ACT	SCHOOLS	PERCENTAGE	SCHOOLS	PERCENTAGE
Identified for School Improvement (SI1 or SI2)	0	0.0%	56	4.3%
Identified for Corrective Action (CA) or Restructuring (R1)	0	0.0%	17	1.3%

For more information, contact your local school or district.

REFERENCES CITED

Airasian, P.W. (1984). Classroom assessment and educational improvement. *Classroom Assessment: A Key to Educational Excellence.* Session presented at Northwest Regional Educational Laboratory, Portland, Oregon.

Amrein, A.L., & Berliner, D.C. (2003). The testing divide: New research on the intended and unintended impact of high-stakes testing. *Peer Review*, *5*(2), 31, 32.

Anastas, J. W. (1999). Chapter 5 – Flexible Methods: Descriptive Research. In J.W. Anastas (Ed.) *Research Design for Social Work and the Human Services* (2nd ed.) New York: Columbia University Press.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36* (5), 258-267.

Bangert-Drowns, R.L., Kulik, C.L. C., Kulik, J.A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61* (2), 213-238.

Başol, G., & Johanson, G. (2009). Effectiveness of frequent testing over achievement: A meta analysis study. *International Journal of Human Sciences*, *6* (2), 99-121.

Black, P. (1990). APU science—the past and the future. *School Science Review*, *72* (258), 13-28.

Blake, J.E. (2012). High-stakes testing: A (mis)construed, normalizing gaze. *International Journal of Educational Policies*, *6*(1), 5-23.

Bushaw, W.J., & Lopez, S.J. (2010). A time for change: The 42nd annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan*, *92* (1), 9-26.

Campbell, D.T., Stanley, J.C., & Gage, N.L. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Cargile, E. (2012, May 3). *Tests' price tag $90 million this year* [news broadcast]. Retrieved from http://www.kxan.com/news

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305-331.

Chingos, M (2012). Strength in numbers: State spending on K-12 assessment systems. Retrieved from Brookings Institution website: http://www.brookings.edu/research/reports/2012/11/29-cost-of-ed-assessment-chingos

Chubb, J. & Hoover Institution on War, Revolution, and Peace. (2009). *Learning from no child left behind: How and why the nation's most important but controversial education law should be renewed*. Stanford, Calif: Hoover Institution Press, Stanford University.

Cizek, G.J. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988) *Statistical power and analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

College Board (2009). *AP Grade distributions for specific grade level groups.* Retrieved on February 27, 2014 from http://media.collegeboard.com/digitalServices/pdf/research/grade-dist-by-grade-level-09.pdf.

College Board (2012). *Online scores for schools and districts*. Retrieved on February 27, 2014 from http://professionals.collegeboard.com/testing/ap/scores/online-score-reporting.

Cook, T.D., & Campbell, D. (1976) The design and conduct of quasi-experiments and true experiments in field settings." *Handbook of Industrial and Organizational Psychology*.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671

Denzin, N.K. (1978). *The research act: A theoretical introduction to research methods.* New York, NY: McGraw-Hill

Elmore, R.F. (2002). Bridging the gap between standards and achievement. *Albert Shanker Institute.* Washington, DC, 2002

Gall, J.P., Gall, M.D., & Borg, W.R. (2003). *Educational research: A practical guide*. Boston, MA: Allen and Bacon.

Goldring, E., Cravens, X.C., Murphy, J., Porter, A.C., Elliott, S.N., & Carson, B. (2009). The evaluation of principals: What and how do states and urban districts assess leadership? *The Elementary School Journal*, *110*(1), 19-39.

Greaney, V., & Kellaghan, T. (Eds.). (2008). *Assessing national achievement levels in education* (Vol. 1). Retrieved from World Bank website: http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2007/12/10/000020953_20071210152818/Rendered/PDF/417890Ed0achie101OFFICIAL0USE0ONLY1.pdf

Gunzenhauser, M.G. (2003). High-stakes testing and the default philosophy of education. *Theory Into Practice*, *42*(1), 51-58.

Herman, J.L., Haertel, E., & National Society for the Study of Education. (2005). *Uses and misuses of data for educational accountability and improvement*. Chicago: NSSE.

Kamens, D.H., & McNeely, C.L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25.

Karpicke, J.D., & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966-968.

Kenny, D.A. (1979). *Correlation and causality.* New York: Wiley

Keppel, F. (1966). *The necessary revolution in American education*. New York: Harper & Row.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koretz, D. (1995). Sometimes a cigar is only a cigar, and often a test is only a test. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessments*? (154-166). Washington D.C.: The Brookings Institution (Brookings Dialogs on Public Policy Series).

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4-16.

Linn, R.L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, *7*(1), 29-38.

Lips, D., & Feinberg, E. (2007). The Administrative Burden of No Child Left Behind. Retrieved from the Heritage Foundation website: http://www.heritage.org/research/commentary/2007/04/the-administrative-burden-of-no-child-left-behind

Madaus, G. (1985), Public policy and the testing profession—you've never had it so good. *Educational Measurement: Issues and Practice,* 4: 5–11. doi: 10.1111/j.1745-3992.1985.tb00294.x

Madaus, G. (1988). The distortion of teaching and testing: High stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29-46.

McDonnell, L. (2008). The politics of educational accountability: Can he clock be turned back? In K. Ryan & L. Shepard (Eds.), *The future of test based educational accountability* (pp. 47-68). New York, NY: Routledge.

McNeil, L. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York, NY: Routledge.

Mehrens, W.A. (2002). Consequences of assessment: What is the evidence? In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for all students: validity, technical adequacy, and implementation*, 149-177. Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Meyer, J.W. (2005). Management models as popular discourse. *Scandinavian Journal of Management*, *21*(2), 133-136.

Mitchell, T.R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review*, *10*(2), 192-205.

Morgan, G.A., Leech, N.L., Gloeckner, G.W., & Barrett, K.C. (2011). *IBM SPSS for introductory statistics: Use and interpretation*. New York, NY: Taylor & Francis.

Morris, A. (2011). *Student standardised testing*. Paris: OECD.

National Center for Fair and Open Testing. (2009). *States continue move to end-of-course exams*. Retrieved February 26, 2014 from http://www.fairtest.org/states-continue-move-to-end-of-course-exams

Nichols, S.L., Glass, G.V., & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, *14*(1)

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 et seq. (West 2003)

North Carolina Principals and Assistant Principals Association (2008). *School executive: principal evaluation process*. Retrieved from http://www.ncpapa.org/forms/Evaluation%20Instrument.pdf on March 1, 2014.

OECD Programme for International Student Assessment. (2004). *PISA Learning for Tomorrow's World: First Results from PISA 2003* (Vol. 659). Paris, France: OECD Publishing.

OECD Programme for International Student Assessment. (2011). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science*. Paris, France: OECD Publishing.

Phelps, R.P. (2000). Trends in Large Scale Testing Outside the United States. *Educational Measurement: Issues and Practice*, *19*(1), 11-21.

Popham, W.J. (1987). The merits of measurement-driven instruction. *The Phi Delta Kappan*, *68*(9), 679-682.

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8-16.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, P.L. (1985). Measurement-driven instruction: It's on the road. *The Phi Delta Kappan*,*66*(9), 628-634.

Principal Evaluations, Ohio Department of Education (2014). Retrieved from http://education.ohio.gov/Topics/Teaching/Educator-Evaluation-System/Ohio-Principal-Evaluation-System-OPES on March 1, 2014.

Resnick, D.P. (1982). *History of educational testing* (pp. 173-194). Ability Testing: Uses, Consequences, and Controversies. Part II. Washington DC: National Academy Press.

Roeber, E. (1995). *Emerging student assessment systems for school reform*. ERIC Clearinghouse on Counseling and Student Services. ERIC Digest.

Ryan, K. & Shepard, L.A. (Eds.) (2008). *The future of test-based educational accountability*. New York, NY: Taylor & Francis. Kindle Edition

Sacks, P. (2000). *Standardized minds: The high price of America's testing culture and what we can do to change it*. New York, NY: Harper Collins.

Shanker, A. (1995). The case for high stakes and real consequences. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessments*, 145-153. Washington, D.C.: Brookings Institution Press.

Shepard, L.A. (1989). Why We Need Better Assessments. *Educational Leadership*, *46*(7), 4-9.

Simpson, R.L., Lacava, P.G., & Graner, P.S. (2004). The No Child Left Behind Act Challenges and Implications for Educators. *Intervention in School and Clinic*, *40*(2), 67-75.

Sjøberg, S. (2012). PISA: Politics, fundamental problems and intriguing results [English trans.]. *La Revue, Recherches en Education*, *14*, 1-21.

Stake, R.E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications.

State of Washington Office of Superintendent of Public Instruction (2014). *Washington State Report Card.* Retrieved on February 27, 2014 from http://reportcard.ospi.k12.wa.us/SideBySide.aspx?).

Thomas, J.Y., & Brady, K.P. (2005). The elementary and secondary education act at 40: Equity, accountability, and the evolving federal role in public education. *Review of Research in Education*, 51-67.

UNESCO (2007). *Global Monitoring Report 2008: Education for All by 2015. Will We Make It?* Paris, France: UNESCO.

U.S. Department of Education, National Center for Educational Statistics (2012). *Digest of Education Statistics, 2011* (NCES 2012-001). Retrieved from http://nces.ed.gov/fastfacts/display.asp?id=84

Vinovskis, M.A. (2001). *Overseeing the nation's report card: The creation and evolution of the national assessment governing board (NAGB)*. Washington D.C.: NAGB

Wang, L., Beckett, G.H., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education*, *19*(4), 305-328.

White, M.E., Makkonen, R., Vince, S., & Bailey, J. (2012) *How California's local education agencies evaluate teachers and principals* (REL Technical Brief 2012-023). Washington, D.C.:U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, *45*(2), 107-122.