THE PREDICTIVE VALIDITY OF THE GRADE TEN NEW STANDARDS REFERENCE EXAMINATION IN ENGLISH/LANGUAGE ARTS IN AN URBAN SCHOOL DISTRICT

BY

LORRAINE EBERHARDT

B. S., ELEMENTARY AND SPECIAL EDUCATION, CHENEY UNIVERSITY, 1976

M. S., LANGUAGE COMMUNICATIONS, UNIVERSITY OF PITTSBURGH, 1982

SUBMITTED TO THE GRADUATE FACULTY OF

ADMINISTRATIVE AND POLICY STUDIES IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF EDUCATION

UNIVERSITY OF PITTSBURGH

2005

University Of Pittsburgh

School Of Education

This Dissertation Was Presented

By

Lorraine Eberhardt

It Was Defended On

July 13, 2005

and approved by

Joseph S. Werlinich, Associate Professor

Isabel Beck, Ph.D., Professor

Otto Graf, Jr., Ph.D., Clinical Professor

Charlene Trovato, Ph.D., Clinical Assistant Professor

Dissertation Director: Joseph S. Werlinich, Associate Professor

THE PREDICTIVE VALIDITY OF THE GRADE TEN NEW STANDARDS REFERENCE EXAMINATION IN ENGLISH/LANGUAGE ARTS IN AN URBAN SCHOOL DISTRICT

Lorraine Eberhardt, Ed.D.

This study describes how the Grade Ten New Standards Reference Examination is used in an urban school district to predict student achievement on the Grade Eleven Pennsylvania State System of Assessment. It was hypothesized that proficient scores on the New Standards Reference Examination are closely associated with high scores on the Pennsylvania System of School Assessment. It was found that a high correlation of agreement exists between the two assessments on proficiency levels which support the hypothesis that the Grade Ten New Standards Reference Examination is a valid predictor of how well students will perform on the Grade Eleven Pennsylvania System of School Assessment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am grateful for a loving family and caring friends. I thank my wonderful husband, Vence for his support throughout this process. I could not have persevered to the end without his words of encouragement and positive attitude. To Tina Still and Dr. Gloria Walton, sincere gratitude is expressed for words of motivation during challenging times. I wish to thank the committee members and Drs. Robert Strauss and John Garrow who all contributed to the completion of this study. I am appreciative for the joy of serving children as my life's work. Above all, I give thanks to God who makes all things possible.

1.0 CHAPTER

1.1 Background

Although the push for school reform is not new, the implementation of standards and assessments to drive the promise of the educational system is a more recent contribution to academic achievement. In *Brown v. Board of Education* (1954), the U.S. Supreme Court ruled that state-mandated segregation in public schools violated the Fourteenth Amendment and was, therefore, unconstitutional. Central to the Court's ruling was the notion that segregated school systems adversely affected student achievement and overall quality of education.

Years after the Brown decision, the legal system remained an avenue for advocates of racial, economic, and qualitative excellence to influence the development of academic achievement. Yet, the courtroom has not been the sole avenue. The United States Congress has also initiated legislation to level population disparities in the nation's school systems.

In 1965, Congress enacted the Elementary and Secondary Education Act (the "ESEA") to provide supplementary federal funds for disadvantaged students (Reichbach, 2004). Eligible students were selected based on test scores (Reichbach, 2004). Although Congress aimed to bring economically disadvantaged children up to basic levels of achievement, many did not feel the Act accomplished its goal.

Nearly 30 years later, Congress, along with President Clinton, passed the Improving America's Schools Act of 1994 (IASA), a revision of the ESEA (20 U.S.C. § 6301). Nominally, it changed the ESEA's Chapter 1 funding program and renamed it Title I (Dougherty, 1998).

Focused on filling the gaps in instruction and support for educationally disadvantaged children, Title I called for clear statements defining the outcomes of student learning, or "standards," as well as assessments to measure student progress (Dougherty, 1998). Economically, Title I allowed schools receiving federal funds to budget resources to aid eligible students. As Dougherty explains, the goal was to transform the federal program "from a remedial track for low-achievers to an accelerated, high performance for low-income and minority students" (Dougherty, 1998). As with the original ESEA, critics believed the gaps in academic achievement between rich and poor, white and non-white, grew wider (Reichbach, 2004).

The Goals 2000: Educate America Act was also passed under the Clinton Administration (Dougherty, 1998). Goals 2000, like the IASA, is a comprehensive act that provides for the development of academic standards. Participants can use their funds to sponsor activities that "involve the writing or implementation of academic standards, to focus on teaching and learning, to take a comprehensive rather than piecemeal approach to reform, to use more flexibility in the use of funds and resources, to develop links with parents and the community, or to target resources where they are most needed" (Dougherty, 1998).

On January 8, 2002, President George W. Bush signed the No Child Left Behind Act (NCLB), a dramatic revision to the ESEA (Reichbach, 2004). Among its provisions, the NCLB called for stronger accountability for results and increased federal funding for reading programs (Kucerik, 2002). In order to comply with the Act, states must follow two steps. First, each state

must establish “challenging” curriculum content and performance-based standards in reading and mathematics (The No Child Left Behind Act, 2001). Next, each state must implement an annual testing system to determine whether these standards for academic achievement are being met (Kucerik, 2002). Further, the NCLB Act requires states to disclose and report data from the yearly tests in annual report cards on school performance and statewide progress. These reports provide information based on race, ethnicity, disability, and limited English proficiency (Kucerik, 2002). Finally, states must develop statewide proficiency and progress objectives and make quantifiable progress in bringing students up to these minimum levels within 12 years (Kucerik, 2002).

In addition to federal legislation, state-based initiatives have been developed to create, meet, and sustain scholastic achievement.

In this regard, The New Standards Project (NSP), founded in 1990, has been most comprehensive. By 1995, The New Standards Project had grown into a partnership of over twenty states and urban school districts (Spalding, 1995). In total, this amounted to nearly half of all United States school children (Spalding, 1995). Spearheaded by Lauren Resnick of the Learning Research and Development Center at the University of Pittsburgh and Marc Tucker of the National Center on Education and the Economy, the New Standards Project explored alternative ways to assess student learning: portfolios, performance tasks, and projects (Spalding, 1995). The project's goal was to develop a performance-based assessment system linked to a set of high national standards. Accordingly, the New Standards Project sought to enhance curriculum, instruction, and student learning as teachers and students developed a shared understanding of the standards, how they are embodied in student work, and how the quality of that work should be judged (Spalding, 1995).

In 1998, Pennsylvania joined the assessment movement when, in October, the Pennsylvania State board of Education adopted Chapter IV of the School Code. Under this Chapter, annual assessments of all public school students are mandatory (Pennsylvania Administrative Code, 22 Pa. Code §4.51, 2004). The assessments were to be based on state-determined standards of performance in academic subjects and skills (Brunner, 2003). The tests, known as the Pennsylvania System of School Assessment (PSSA), were intended as a means of providing students, parents, educators, and citizens with an understanding of student and school performance, especially in terms of student attainment of state-set standards (Pennsylvania Administrative Code, 22 Pa. Code §4.51, 2004). School districts can also use these test results to assess student proficiency and map new strategies for achievement (Brunner, 2003).

The PSSA tests student performance in reading, writing, and mathematics. All school districts participate in the reading and writing assessments each year (Brunner, 2003). Math and reading skills have been assessed at grades five, eight, and 11; writing has been formerly tested at grades six, nine, and 11 (Brunner, 2003). Currently the PSSA writing is field-tested at grade five and eight and an operational test administered at grade 11. Participation in the writing assessment occurs before a district's six-year planning cycle begins, after three years, and at the end of the planning cycle, although districts may participate off-cycle on a voluntary basis (Brunner, 2003).

1.1.1 Organization Of The Study

This study describes how selected interim assessments and the Pennsylvania state assessments are used and compared in an urban school district to predict student achievement and to inform instruction. The study contains the five major sections described below:

Chapter I provides an overview of the issues to be investigated, outlines the research that will be undertaken and identifies the research questions to which responses will be sought. This chapter also identifies the statement of the problem, problem significance, and the theoretical framework upon which the study will be based. Finally, Chapter I specifies the limitations of the study and defines the terms used.

Chapter II presents a review of relevant literature and research about the usefulness of standards-based assessments. The Chapter begins with a general overview of standards and federal and state-mandated policies regarding assessments; the focus then narrows to locally selected assessments at the school district level. It offers research responses from experts regarding the use of standards-based assessments to measure student achievement and improve classroom instruction.

Chapter III describes the research methodology. Explanations will detail the research methodology, the research population, and the method of data collection and analysis.

Chapter IV presents the data collected and explains the results in relation to the research questions and also notes unexpected results and concrete findings of the study.

Finally, Chapter V presents conclusions and implications of the study.

1.1.2 Purpose of the Study

The purpose of this study is to explore the ability of the NSRE English/Language Arts to predict PSSA Reading scores. Considerable data has been published on standards-based assessments and how those assessments are used to inform instruction with regards to improving student achievement. Many school districts have mandated curriculum standards and tests for their students. This study will attempt to compare two standards-based assessments in an effort

to help teachers understand and use those assessments to improve classroom instruction and raise student achievement.

1.1.3 Statement of the Problem

This study is being conducted to describe how one standards-based assessment can be used as a valid predictor of how well students will perform on a similar standards-based assessment in the same content area. Specifically, the problem statement is

> What is the predictive validity of the Grade Ten New Standards Reference Examination (NSRE) in English/Language Arts, in relation to the Grade Eleven Pennsylvania System of School Assessment (PSSA), Reading Test, in the Pittsburgh Public Schools?

1.1.4 Research Questions

The central question to be answered by this investigation is the following:

> What is the predictive validity of the Grade Ten New Standards Reference Examination, English/Language Arts in relation to the Grade Eleven Pennsylvania System of School Assessment, Reading Test in the Pittsburgh Public Schools?

The following sub-questions will be used to fully explore the central question:

a. Is there a difference between male and female scores?
b. Is there a difference in scores between the students of various socioeconomic groups?
c. Is there a difference in the relationship between African-American and White student scores?

1.1.5 Hypotheses

The following hypotheses have been developed with regards to the previous questions presented:

1.1.5.1 Central Question Hypothesis

There is a strong relationship between student scores on the New Standards Reference Examination, English/Language Arts given in grade 10and the Pennsylvania System of School Assessment, Reading administered in grade 11. Proficient scores on the New Standards Reference Examination are closely associated with high scores on the Pennsylvania System of School Assessment.

1.1.5.2 Sub-question Hypotheses

Females are more likely to show higher agreement on proficiency levels than males.

Scores on the New Standards Reference Examination can be used to predict the results on the Pennsylvania System of School Assessment for low-income students.

African-American students will significantly agree on proficiency and non-proficiency levels with regards to the New Standards Reference Examination and the Pennsylvania System of School Assessment.

1.1.6 Significance of the Study

Pittsburgh Public Schools (PPS), as part of its strategic plan, adopted the policy of measuring student achievement for all grades (K-12). Since the PSSA is only given at grades 3, 5, 8 and 11, Pittsburgh Public Schools was left to determine which interim assessments would be used to provide a continuous flow of measuring student achievement in the non-PSSA assessed grades. In response to this need, the PPS selected multiple assessments. One of those assessments was the NSRE in English/Language Arts at grade 10. The district utilizes the NSRE in grade 10 as a predictor to determine student success on the 11^{th} grade PSSA. This assessment is also used to evaluate student instructional needs in an effort to determine student

strengths and weaknesses in the English Language Arts. To that end, a differentiated intervention plan is developed for individual student needs. This strategy is significant in that it provides a seamless overlay of assessment that ultimately moves students toward achieving proficiency on the PSSA. This state assessment is the measurement tool selected by the Pennsylvania Department of Education (PDE) to hold schools accountable for adequate yearly progress as mandated by No Child Left Behind (NCLB). The results of this study will give the PPS a vital understanding of the relationship between its various assessment tools and their impact on instruction. This body of knowledge may also benefit nationally in that school districts throughout the country may acknowledge and utilize this PPS strategy. Additionally, this study will be of incalculable significance in that it will describe through systematic scientific observations another dimension whereby standards-based assessments, when compared, are actually indicators of student performance against the standards. To that end, to the profession of teaching reading, it is expected that this study will yield data that will make a contribution as to whether different standard based assessments are valid and reliable predictors of student performance.

1.1.7 Theoretical Framework

Within the past two decades, a number of judicial and scholarly pronouncements have been made concerning the ability of performance-based assessments to bring about sweeping educational reform. Of all the standards and assessment projects of the 1980s and 1990s, the New Standards Project was unquestionably the most ambitious. A national coalition of approximately 17 states and seven urban school districts, co-directed by Lauren Resnick of the Learning, Research, and Development Center of the University of Pittsburgh and Marc Tucker of the National Center on Education and the Economy in Washington, D.C., the New Standards

Project explicitly aimed to create tests worth taking. To that end, the theoretical framework of this study has been established in concert with those concerns outlined by the authors of the New Standards Project. Both assessments compared in this study attempted to create a bank of performance tasks that would yield similar information about the reading skills of tenth and 11th graders when compared to the standard. Both the NSRE and the PSSA claim to report how well students are doing relative to a pre-determined performance level on a specified set of educational goals or outcomes included in state or federal standards. Therefore, the inspiration to acquire data on the specific assessments noted is sanctioned by the prominence of the New Standards Project (Spalding 1995).

1.1.8 Limitations of the Study

There are several limitations in this study. The first limitation relates to the sample of the study. The sample is single institutional data. Therefore conclusions cannot be generalized to the national sample. Second, the sample in this study overly represented the African-American students as the sample size for other groups was considerably small. A third limitation is the scope and sequence and pacing of the curriculum in that all concepts may not be taught uniformly and may not be taught based upon the mandated timelines prior to testing. To that end, in certain instances, teachers may teach objectives measured on the PSSA following administration of the test. Another limitation worthy of mention is that the scale scores of the PSSA are not comparable to those of the NSRE. However, since they are both horizontally scaled and psychometrically grounded, valid comparisons can be made. A final limitation of the study is that the NSRE measures national standards compared to the PSSA which measures Pennsylvania Academic Standards. Therefore, until studies are done that equate scores from the

PSSA with other states' exams as well as national exams, the exact correlations between NSRE and other state examinations will remain unknown.

1.1.9 Definition of Terms

Anchoring—An approach to equating where an anchor test is administered along with various forms of a test in order to provide a basis of comparison between the two forms (Linn, 1996).

Calibration—A form of linking in which two different types of tests may be compared although each test may assess performance at different levels. Typically, calibration involves the desire to compare scores from a short form of a test to those from a longer form (Linn, 1996).

Construct validity—A measure of how well a set of test results compare with those from other high-quality assessments of similar or dissimilar skills (Stecher, et. al., 1997).

Content validity—The ability of experts to review the content of an assessment and confirm that it is measuring the desired skills or behaviors under review (Stecher, et. al., 1997).

Criteria-referenced tests—Assessments that measure a student's performance against a rigid set of curricular standards.

Equating—A fundamental approach to linking in which scores from two tests can be used interchangeably such that any use or interpretation for one test will also work for the equated scores on the other test (Linn, 1996).

Fairness—An assessment is unfair, or biased, if students of relatively equal skill before differently on a particular question because of experience or knowledge not related to the underlying skill (Stecher, et. al., 1997).

Linear Regression—A form of regression analysis in which the function is a linear (straight-line) equation that expresses the best prediction of the dependent variable (*Y*), given the independent variables (*X*) (Draper, 1980).

Linking—A generic term that refers to making statistical comparisons between the results from one test or set of assessment tasks to those of another (Linn, 1996).

Norm-referenced tests—A category of assessment tests, typically of the multiple-choice variety, where a representative group of students ("the norm group") is given a test and, after the test is published, the scores of students who take the test are then compared to the "norm" (Bond, 1996; Kelly, 1998).

Prediction—A methodology for linking assessments that attempts to anticipate scores on a test based on performance on a previous assessment (Linn, 1996).

Rater reliability—An approach to reliability that asks whether the same scores would be assigned if a different group of experts were to read the student responses (Stecher, et. al., 1997).

Regression Analysis—Statistical analysis used to determine the values of parameters for a function that cause the function to best fit a set of data observations (Draper, 1980).

Reliability—The measure of the degree to which an individual measurement is free from error. Reliability measurements can occur in various forms: test-retest reliability, parallel-forms reliability, and rater reliability (Stecher, et. al., 1997).

Scaling—An approach to linking that becomes useful in situations where it cannot be assumed that the population taking one achievement test is equivalent to the group taking another.

Statistical moderation—An approach to linking assessments most often used to help make comparisons among students who have taken different combinations of achievement tests.

Test-retest reliability—A form of reliability that informs the researcher whether the same results would be produced if the assessment were given again (Stecher, et. al., 1997).

Validity—The inferences being drawn from the applicable test score. Techniques for establishing validity include the following approaches: content validity, concurrent or predictive validity, construct-validity, and consequential validity.

2.0 CHAPTER

2.1 Literature Review

2.1.1 Introduction

This chapter presents a review of the relevant literature and research about the usefulness of standards-based assessments. The Chapter begins with a general overview of standards and federal and state-mandated policies regarding assessments; the focus then narrows to locally selected assessments at the school district level. Additionally, in its discussion of standards and the methodology for linking assessments, it offers research responses from experts regarding the use of standards-based assessments to measure student achievement and improve classroom instruction.

2.1.2 Federalism and the Assessment Movement

2.1.2.1 Phase I: The Establishment of the Elementary and Secondary Education Act

In 1998, Arthur Coleman, Secretary of Civil Rights for the U.S. Department of Education, suggested that meeting the needs of students was the most basic obligation of educators. (Coleman, 1998). As a means of accomplishing this, teachers were told to take students as they find them,

> with their different backgrounds and abilities, and to inculcate values and teach skills to allow them to grow to maturity with meaningful expectations of a productive life in the workforce and elsewhere (Coleman, 1998).

In other words, the educator's task is to assist students in achieving their full potential, during classroom instruction as well as during the administration of tests and evaluations (Coleman, 1998). While the classroom initiatives of teachers have generally been praised, much criticism has attended the increasing federalization of school reform and its push for assessment-centered reform through standardized testing.

Secretary Coleman's view echoes a longstanding value in society exalting the virtues of an attentive and robust educational system. Indeed, U.S. Supreme Court Chief Justice Earl Warren, in the landmark case of Brown v. Board of Education (1954), stated the matter thusly:

> Today, education is perhaps the most important function of state and local governments. Compulsory school attendance laws and the great expenditures for education both demonstrate our recognition of the importance of education to our democratic society. It is required in the performance of our most basic public responsibilities, even service in the armed forces. It is the very foundation of good citizenship. Today it is a principal instrument in awakening the child to cultural values, in preparing him for later professional training, and in helping him to adjust normally to his environment. In these days, it is doubtful that any child may reasonably be expected to succeed in life if he is denied the opportunity of an education.

Despite the principles enunciated by the Court, the nation has struggled with, first, how to ensure the availability of a quality education to all children and, second, how to measure the concept of "quality" and achievement. As a result, the Brown ruling sparked national dialogue about the quality of education afforded to African-American children, as well as a broader discussion about the needs of all other children raised in poor families or with other disadvantages (Zamora, 2003).

In the early 1960s, Congress blocked most of the proposals for federal education initiatives generated by President John F. Kennedy, including general aid for school construction and teachers' salaries. As President, Lyndon Johnson, hailing from an impoverished

background, sought to earmark federal aid for the "compensatory" education of the disadvantaged student as part of his "anti-poverty" campaign (Caro, 1982).

In 1965, perhaps as a response to the inability of the states to comply with integration mandate of the Brown's decision, the federal government entered the educational fray by passing the 1964 Civil Rights Act (Chadsey, 2002). The 1964 Civil Rights Act gave the Department of Health Education and Welfare (HEW) the authority to set regulations which would determine local school districts' eligibility for receiving federal educational funds (Chadsey, 2002). As a means of enforcing this authority, it gave HEW authority to terminate funds in the event of a school district's failure to comply (Chadsey, 2002).

The importance of this authority became apparent when Congress passed the Elementary and Secondary Education Act ("ESEA"), which made large sums of federal dollars available to local school districts for the first time (Chadsey, 2002). According to Gerald Rosenberg (2001), the massive influx of federal educational dollars through the ESEA, and HEW's consequent ability to withhold those funds, as provided in the 1964 Civil Rights Act, eventually helped to erode the widespread and defiant practice of openly segregated schools. Court decisions, in the spirit of Brown v. Board of Education, also contributed. Perhaps an unintentional consequence of the 1964 Civil Rights Act, the ESEA, and the judicial activism of the Warren Court was the ushering in of a new era of federalized school reform marked by a tremendous focus on accountability and standardized testing. From a Constitutional standpoint, this is significant because it launched a debate between two historically opposed factions: on the one hand, a faction that contemplates federal activism to promote the "general welfare" of the country and, on the other, a staunchly anti-federal faction, pointing to the 10th Amendment of the Constitution, advocating limited government regarding areas clearly within the police powers of the states

(Zamora, 2003). President Johnson, whose efforts to galvanize both factions in pursuit of his educational goals helped facilitate the swift passage of the ESEA, declared, "I will never do anything in my entire life, now or in the future, that excites me more, or benefits the Nation I serve more…than what we have done with this education bill" (Jennings, 2001).

Kenneth Wong (2002) has discerned three distinct policy phases regarding the implementation of the ESEA and Title I. The first is a decidedly anti-poverty policy period, occurring between 1965 and continuing into the 1980s (Wong, 2002). During this time period, local decision-making was often challenged by federal anti-poverty goals. In fact, when President Johnson proposed the ESEA, he issued the following pronouncement: "Poverty has many roots, but the taproot is ignorance. . . . Just as ignorance breeds poverty, poverty too often breeds ignorance in the next generation" (Jennings, 2001). In the second phase, occurring from the mid-1980s to the mid-1990s, educators and policymakers began to place greater emphasis on quality of instruction and curriculum. Third phase, beginning in the mid-1990s, saw reformers making greater attempts to "restructure Title I in different directions, namely, whole school reform, district-based support, annual testing, and consumer-based or voucher programs" (Wong, 2002). Overall, and as a trend, Title I has reduced its focus on regulatory compliance while increasing its emphasis on outcome-based accountability (Wong, 2002).

Phase I began with the passage of the ESEA. According to its legislative history, the Act was designed to

> …provide financial assistance ... to local educational agencies serving areas with concentrations of children from low-income families to expand and improve their educational programs ... which contribute particularly to meeting the special educational needs of educationally deprived children (Elementary and Secondary Education Act, 1965).

The philosophy behind the ESEA was that poverty imparts a deleterious effect upon educational progress for disadvantaged children (Zamora, 2003). This lack of educational achievement resulted in a continuing "cycle" of poverty. The Senate Education Committee's report on the ESEA stated that

> the conditions of poverty or economic deprivation produce an environment which in too many cases precludes children from taking advantage of the educational facilities provided . . . . Under [Chapter 1] of this legislation the schools will become a vital factor in breaking the poverty cycle by providing full educational opportunity (Jennings, 2001).

The aim then is to compensate for the effects of group poverty on an individual student rather than the eradication of poverty in general or individual poverty (Zamora, 2003).

Thus, Title I, as it was conceived in 1965, did not target to low-income students directly, but spread its benefits to all students, regardless of income, who suffered from the conditions of poverty that negatively impacted their schooling (Zamora, 2003). This philosophy was reflected in Congress' recognition that "the correlation between poverty and educational access did not apply in every case: a student can be educationally disadvantaged without being economically disadvantaged, and vice versa" (Jennings, 2001). By addressing poverty through the proxy of educational opportunity, the ESEA

> distributed the federal funds to school districts and schools on the basis of poverty data for children, but then made the services available to all students at these schools on the basis of educational need (Jennings, 2001).

In this way, Title I was more of a funding stream than a comprehensive educational program (Zamora, 2003). The ESEA bundled a federal aid program for "educationally deprived children" with an array of smaller programs of federal aid specifically earmarked for the purchase of library books, the creation of supplemental education centers, and the development of state departments of education (Zamora, 2003). Early on, Title I was solely concerned with

providing a measure of equity in levels of funding, not with the types of programs implemented with Title I funds. It equated educational opportunity with rough equality of inputs (funding). Thus, Title I took the form of federal grants to state education agencies and local school districts without any prescription for their use other than that they benefit the "educationally-deprived child" whose educational deprivation stems from contact with poverty. These funds were meant to remedy funding disparities, not to alter the substance of curriculum or methods of instruction (Zamora, 2003). Quality of instruction and curriculum was not a major focus until the 1980s or, as Wong dubs it, Phase II.

Title I funds were to be awarded to state departments of education, which were to distribute the funds to local school districts upon the district's submission of a program application describing local needs and plans for program funds (Zamora, 2003). The amount of money that a local district received was to be calculated using a formula that considered the number of children from low-income families residing within the district's boundaries (Zamora, 2003). Within the district, funds were to be allocated to schools on the basis of the poverty rate of the area surrounding the school: an elementary or secondary school was to receive Title I funds if the estimated percentage of children from low-income families was as high as the percentage of such children residing in the whole of a school district (Zamora, 2003). Federal regulations further required that funds within districts be concentrated upon the highest poverty areas within the district (Zamora, 2003). One of the most important targeting provisions to Title I was the requirement that funds "supplement, not supplant" local funds spent on disadvantaged students. This was meant to ensure that Title I did not just free up state and local monies that were spent on disadvantaged children so that the state could spend more money on non-Title I students (Zamora, 2003).

Unfortunately, the ESEA did not adequately meet its original goals of economic parity. According to Peter Zamora (2003), the reason for its inadequacy rested with fallacies of philosophy and wavering resolve with respect to implementation. With regard to philosophy, the ESEA's strategy rested on several unsupported assumptions, notably:

1) Poverty was an inherent educational disadvantage;

2) Educational underachievement had negative economic consequences;

3) This cycle could be broken through the targeted use of limited federal funds to partially counterbalance operational funding disparities between schools and income disparities in the home;

4) The curricula, pedagogy, and local expertise employed in local schools at the time were all potentially effective in improving achievement among the "educationally deprived;"

5) The needs of local school districts are coextensive with the needs of low-achieving students; and

6) Equalizing educational opportunity would create increased economic opportunity" (Zamora, 2003).

On the subject of wavering resolve, Zamora cites insufficient Congressional funding of the Title I program, misappropriated and/or wasted resources by local authorities, poor pedagogy at the local level, lack of parental involvement, lack of federal enforcement, and the lack of meaningful evaluation measures (Zamora, 2003).

2.1.2.2 Phase II: the Push for Quality in Instruction and Curriculum

Between the 1960s and 1980s, the period Wong refers to as Phase I, dissatisfaction with the lack of success of the ESEA in meeting the needs of impoverished and minority students paved the way for greater emphasis on quality of instruction and curriculum. Khattri and Sweet (1996) attribute this newfound emphasis to "three related phenomena, all gaining momentum during the late 1980s:

- the reaction on the part of educators against pressures for accountability based on norm-referenced testing;
- the development in the cognitive sciences of a constructivist model of learning; and
- the concern on the part of the business community that students entering the workforce were not competent enough to compete in an increasingly global economy.

The impetus for reform may have begun earlier than the late 1980s. In its 1983 report entitled *A Nation at Risk*, the National Commission on Excellence in Education warned that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people" (National Commission on Excellence in Education, 1983). According to the report, U.S. students performed poorly compared to students in other countries. The report also offered the following statistics: 13% of all 17-year-olds were functionally illiterate, with rates among minority groups running as high as 40%; large percentages of 17-year-olds were also said to be without sufficient intellectual skills to draw inferences from written material, solve complex mathematical problems, or write persuasive essays; scores on tests measuring achievement of high school and college students were said to be in consistent decline; and business and military leaders complained that they were obliged to spend millions on costly remedial courses in basic skills such as reading, writing, spelling, and mathematics (National Commission on Excellence in Education,1983). As a result of the national concern surrounding *A Nation at Risk*, parents, schools, and policy makers focused on education reform (Elul, 1999).

At the same time, Americans became more likely to accept departures from localized educational authority (Heise, 1994). Alarmed that “our students were leaving our public schools ill-prepared for the world as citizens and as workers,” states such as California and Kentucky

increased state control of curriculum content and promoted statewide standards (Heise, 1994). Similar changes were being contemplated in North Carolina, Illinois, Minnesota, Wisconsin, and Vermont (Heise, 1994). According to a 1987 Gallup Poll, 84% of the American public approved of the idea of federal regulations requiring state and local educational authorities to meet minimum federal standards (Heise, 1994). Following this wellspring of support, President George H. W. Bush invited the governors of the fifty states to Charlottesville, Virginia for an Education Summit in September 1989. At the Summit, the President and the governors agreed to produce a set of national education goals, the first of its kind in the country's history (Heise, 1994). Notably, Governor and eventual President, William Jefferson Clinton of Arkansas was among those present (Gergen, 1990).

Two years later, Congress passed the Education Council Act of 1991 (Education Council Act, 1991). Title IV of that Act established the National Council on Education Standards and Testing, which was authorize to advise Congress, the Secretary of Education, and the National Education Goals Panel on issues relating to the desirability and feasibility of establishing national educational standards and a uniform system of student examinations.

In 1994, Congress revised the ESEA and renamed it the Improving America's Schools Act (IASA) (Dougherty, 1998). Not only did the revision's new structure call for definitive statements about what students should learn, it also mandated assessments on state created standards (Dougherty, 1998). Flexibility accompanied receipt of Title I funds such that schools could allocate budgeting resources so that Title I students could obtain much needed support. Eleanor Dougherty (1998) described the effort as an attempt to "change the federal program from a remedial track for low-achievers to an accelerated, high performance educational experience for low-income and minority students" (Dougherty, 1998).

Under the Clinton Administration, Congress also passed the Goals 2000: Educate America Act (1994). Unlike Title I, Goals 2000 was shorter and less complex and did not focus as heavily on providing monies to develop specific educational programs. Under Goals 2000, participants could use their funds to sponsor a wide variety activities, including: the writing or implementing academic standards, focusing on teaching and learning, taking a comprehensive rather than incremental approach to reform, or targeting resources where they were needed most (Dougherty, 1998).

More specifically, Goals 2000 set up national education goals. Among these goals are: By the year 2000, all children in America will start school ready to learn.

By the year 2000, all students will leave grades 4, 8 and 12 having demonstrated competency over challenging subject matter including English, mathematics, science, foreign languages, civics and government, economics, arts, history and geography, and every school in America will ensure that all students learn to use their minds well so they may be prepared for responsible citizenship, further learning, and productive employment in our Nation's modern economy.

By the year 2000, United States students will be the first in the world in mathematics and science achievement (Educate America Act, 1994).

Goals 2000 left the achievement of these goals, including the methods for doing so, to local educators. States applying for federal funds under Goals 2000 had to establish content standards and state improvement plans for meeting those standards. The improvement plans were also required to describe their plans for student assessments (Educate America Act, 1994). In this way, the 1990s saw the pendulum of academic reform swing in favor of assessments and accountability under a sweeping federalized model.

Goals 2000 described criteria-referenced tests that must be aligned to curricula, as well as performance-based measure. However, the statute did not prescribe the purposes of these tests. The language of the statute is broad enough to include individual student assessment or regional comparisons (Kelly, 1998). Goals 2000 did provide details on the form of these assessments. Among other things, the assessments had to: be aligned with the State's content standards; use multiple measures of student performance; be accessible to students with diverse learning needs; allow for accommodations and adaptations for students with those diverse learning needs; be consistent with nationally recognized professional and technical standards for such assessments; provide the state with coherent information about student attainment of the standards; and support effective curriculum and instruction (Educate America Act, 1994). Additionally, Goals 2000 created the Goals Panel to work with and assist the states with technical support, largely with regard to early childhood assessments used to gauge school readiness (Educate America Act, 1994). Between 1994 and 1998, $1,270,270,000 had been allocated to the states from the federal government to support state-submitted improvement plans (Kelly, 1998).

In his 1999 State of the Union Address, President Clinton called for an end to social promotion in primary and secondary schools. Social promotion occurs when students are promoted with their peer without regard to whether they have attained the skills to succeed in the next grade level. School districts have been sued in several states under the legal theory that social promotion deprives students of an adequate education. In his Address, President Clinton stated that

> . . . no child should graduate from high school with a diploma he or she can't read. We do our children no favors when we allow them to pass from grade to grade without mastering the material (Clinton, 1999).

Further, the President promoted his Education Accountability Act, requiring schools receiving federal funds to end social promotion, adopt higher education standards, hold school districts and teachers responsible for poor student achievement, and inform parents of school quality.

In passing Goals 2000 and in reauthorizing Title I in 1994 through the passage of the Improving America's Schools Act (IASA), Congress and President Clinton incorporated the core ideas of standards-based reform. In doing so, they fundamentally changed the nature of Title I. Instead of providing funds to support remedial instruction for disadvantaged students, Title I funds had to be used to create standards for all students.

2.1.2.3 Phase III: No Child Left Behind

The No Child Left Behind Act (NCLBA, 2001) follows the same basic approach as the IASA, but it establishes more ambitious goals and places greater constraints on the state education system. States must still develop "challenging" content and performance standards, now not only in reading and math, but also in science (NCLBA, 2001). States must still use assessments that are aligned with those standards, and must hold schools and school districts accountable for failing to meet ambitious achievement goals (NCLBA, 2001).

The most significant changes have to do with teachers, testing, and accountability. As for teachers, the NCLBA (2001) requires that Title I schools hire only "highly qualified" teachers for all subjects and that veteran teachers in such schools demonstrate that they are "highly qualified" (NCLBA, 2001). The Act also reaches beyond Title I schools and requires that all teachers of "core academic subjects" in non-Title I schools must be or become "highly qualified" (NCLBA, 2001). Pursuant to the Act and accompanying regulations, teachers are considered "highly qualified" if they are fully certified and have demonstrated competency in the subjects they teach.

As for testing and accountability, whereas the IASA required testing in math and reading at three points in a student's school career, the NCLBA requires annual testing in reading and math in grades three through eight, beginning in the 2005-6 school year. At least one additional test in reading and math must be given in grades 10 through 12. Beginning in 2007-8, students must also be tested in science at least three times between grades three through 12 (NCLBA, 2001).

Test scores are central to the NCLBA. Scores are tabulated for schools in a bundle and must be dissembled into subgroups, including migrant students, disabled students, English-language learners, and students from all major racial, ethnic, and income groups. All of these scores are then used to determine whether schools are making "adequate yearly progress." Adequate yearly progress (AYP), in turn, is the linchpin of the NCLBA (2001).

Adequate yearly progress is tied to whether a sufficient percentage of students are performing proficiently on state tests. The NCLBA requires states to bring all students to the proficient level within 12 years of the Act's passage (i.e., by 2014), and states must ensure that their definitions of adequate yearly progress will enable the ultimate 12-year goal to be met. To accomplish this, states must set a proficiency goal each year, and that percentage must rise periodically so that by 2014, it hits 100%. For a school to make adequate yearly progress, the student population as a whole, as well as each identified subgroup of students, must meet the same proficiency goal. For example, if in the year 2004-5, the state determines that 65% of students must be "proficient" on the tests, 65% of all the students within a school and 65% of the students within each subgroup (e.g., disabled students, poor students, minority students) must be performing proficiently for a school to be making adequate yearly progress (NCLBA, 2001).

Adequate yearly progress is thus less about yearly achievement gains than it is about hitting uniform benchmarks. All states must set a uniform bar for achievement for all schools and all subgroups of students within a school. The first benchmarks were based on test scores from 2001-2. Using these test scores, states had to establish a starting point for AYP that was the higher of the following two values: (1) the percentage of students in the lowest-achieving subgroup, statewide, who were performing proficiently; or (2) the threshold percentage of students performing proficiently in the lowest-performing quintile of schools statewide (NCLBA, 2001).

If 30% of a state's poor students, for example, scored at the proficient level in 2001-2002, while 40% of all students in the school at the 20th percentile of achievement scored at the proficient level, the initial AYP bar must be at least 40% for all schools and all subgroups of students (Ryan, 2004). According to the language of the Act, the percentage of students performing proficiently must rise every two or three years, like stair steps, until the 2013-2014 school year, when all students must be scoring at the proficient level (Ryan, 2004).

Although the Act is quite strict in defining AYP, it is remarkably loose with regard to state standards and tests, resulting in state-level freedom in determining their own standards, to create their own tests, and to determine for themselves the scores that individual students must receive in order to be deemed "proficient." The harder the tests or the higher the scores needed, the harder it will be for schools to meet the NCLBA's definition of adequate yearly progress (Ryan, 2004). For the same reasons, some states have much farther to travel than others in order to meet the goal of 100% proficiency. The starting percentages in Massachusetts, for example, were roughly 40% proficiency in reading and 20% proficiency in math. In Colorado, the starting

percentages ranged, depending on the grade level, from roughly 75%-90% in reading and 50%-80% in math (Ryan, 2004).

The Act requires all schools within a state, regardless of whether they receive Title I funding, to make adequate yearly progress. It also requires states and districts to disseminate information about each school's AYP status. The stricter accountability mechanisms, however, are reserved for schools receiving Title I funding (NCLBA, 2001).

Those schools that receive federal funding and fail to make adequate yearly progress are identified as in need of improvement. They are also subject to a range of progressively more serious actions. After two consecutive years of failure, schools must develop a plan for improvement and are supposed to receive "technical" assistance. Students in those schools are also allowed to choose another public school, including a charter school, within the same district. After three years, students who have not already departed for greener pastures must be provided with tutoring services from an outside provider, public or private. Those schools that fail to make AYP for four consecutive years must take one of several measures, including replacing school staff or instituting a new curriculum, and those that fail for five years in a row must essentially surrender control to the state government, which can reopen the school as a charter school, turn over management to a private company, or take over the school (NCLBA, 2001).

Additionally, the NCLBA requires that the National Assessment of Educational Progress (NAEP) reading and math tests be administered every two years to fourth and eighth graders (Ryan, 2004). The NAEP is an extensive testing program that has been used for over 30 years to collect data about student achievement (Ryan, 2004). The test is essentially national, in that it is not aligned with any state standards. Instead, the NAEP attempts to measure content and skills thought common to all state educational systems. Prior to the NCLBA, participation in the NAEP

was voluntary, but now all states must participate. Nonetheless, only a random sample of students within each state must take the test, and scores are not reported for individual students or individual schools (Reckase, 2002). The NCLBA does not indicate what is supposed to be done with the results of the NAEP, but supporters of the Act suggest that results on the NAEP will ensure the rigor of standards and tests used in each state (Ryan, 2004). Whether this use of the NAEP will be successful in keeping state standards and tests rigorous is subject to serious question from scholars, parents, and educators alike.

2.1.3 Pennsylvania Standards & Assessments

In 1998, the Pennsylvania State board of Education adopted Chapter IV of the School Code, which mandated annual assessments, or tests, of all public school students are mandatory (Pennsylvania Administrative Code, 2004). The tests, known as the Pennsylvania System of School Assessment, are based on state-determined standards of performance in academic subjects and abilities. All school districts participate in the reading and writing assessments each year (*Writing Assessment Handbook: Overview*, 2004). Math and reading skills have been assessed at grades five, eight, and 11; writing has been formerly tested at grades six, nine, and 11. Currently, the PSSA writing is field-tested at grade five and eight and an operational test administered at grade 11. Participation in the writing assessment occurs before a district's six-year planning cycle begins, after three years, and at the end of the planning cycle, although districts may participate off-cycle on a voluntary basis (*Writing Assessment Handbook: Overview*, 2004).

The purpose of the PSSA include, but are not limited to, providing students, parents, educators, and citizens with an understanding of student and school performance, especially in terms of student attainment of state-set standards (Pennsylvania Administrative Code, 2004).

Accordingly, each school district can use the test results to assess student proficiency and map new strategies for achievement (Brunner, 2003).

Specifically, Chapter IV states that information regarding student performance must include student names to ensure that such information "is available to parents and teachers" (Pennsylvania Administrative Code, 2004). The regulations clarify that individual [PSSA] results shall be used in planning instruction only by parents, teachers, administrators and guidance counselors with a need to know based upon local board policy on testing and in reporting academic progress (Pennsylvania Administrative Code, 2004). The regulations also state that the Department of Education is prohibited from collecting individual student test scores, and may collect school and district scores only in the aggregate. Students not achieving the proficient level in the grade eleven assessments may be given another chance to demonstrate proficiency in grade 12 (Pennsylvania Administrative Code, 2004).

Pennsylvania educators, as part of Advisory Committees, choose the concepts that form the basis for the assessments (*Writing Assessment Handbook: Overview*, 2004). The advisory committees are themselves composed of teachers, supervisors, curriculum directors, and college specialists. Often, these educators either write the test questions, tasks, and writing prompts themselves or they replicate them from outside examples specifically designed and created for Pennsylvania students (*Writing Assessment Handbook: Overview*, 2004). The Pennsylvania Department of Education contracts with Data Recognition Corporation (DRC), a private Minnesota-based enterprise founded in 1978, for scoring of the PSSA (Data Recognition Corporation 2004, hereinafter "DRC"). Other states under contract with DRC are Alaska, Alabama, Arkansas, Georgia, Louisiana, Minnesota, North Carolina, Ohio, and South Carolina (DRC, 2004). In addition, the Pennsylvania Department of Education's "partnership" with the

company has grown in 2004 to include a new battery of Reading and Mathematics tests for grades four, six, and seven, encompassing approximately 450,000 students (DRC, 2004)

In accordance with Pennsylvania's writing assessment regulations, student performance on PSSA writing tests must be demonstrated by "the quality of students' written compositions on a variety of topics and modes of writing" ((Pennsylvania Administrative Code, 2004). In 1989, as part of a continuing review of the conceptual bases for statewide testing, the Writing Assessment Advisory Committee (WAAC), a composite of over sixty education professionals from across the state, was formed to design a writing test that would measure students' ability to write for different purposes. To do this, they examined writing research, investigated various types of writing assessments being used by several other states and the National Assessment of Educational Progress (NAEP), and studied the implications of the Pennsylvania Department of Education's framework for integrating communication skills across the curriculum (*Writing Assessment Handbook: Overview*, 2004).

WAAC identified three types of writing - informational, narrative, and persuasive - that it deemed most appropriate for state assessment because they are "among the most important in school and life" (*Writing Assessment Handbook: Overview*, 2004). WAAC also has developed guidelines for the scoring of the essays and provides assistance in aligning the items within the assessment to meet state writing standards. Among its other responsibilities, WAAC also is required to attend scoring sessions with the independent entities with which it contracts to score the PSSA writing assessment (*Writing Assessment Handbook: Overview*, 2004).

The Division of Evaluation and Reports of the Pennsylvania Department of Education publishes annually the Writing Assessment Handbook, which provides an overview of the test and its administration and gives detailed examples of how each standard is measured (by

providing sample evaluations of actual student essays from past years' assessments). The Handbook describes WAAC's "Scoring Guide," which has two components: the six-point scoring rubric and the five characteristics of effective writing. The five characteristics of effective writing are focus, content, organization, style, and conventions (including grammar, mechanics, spelling, usage, and sentence formation). Displayed graphically by the Pennsylvania Assessment Holistic Scoring Guide are the five characteristics located in Appendix A of the Handbook (*Writing Assessment Handbook: Overview*, 2004).

The first items students encounter on the writing assessment relate to specific strategies used in the writing process. They are asked to answer "Yes" or "No" to whether they have been taught how to brainstorm, about different types of writing, how to revise and edit, how to conference with teachers and peers, and how to use a computer or word processor. Next, students are presented with several statements related to opportunities to practice specific writing strategies. For example,

- I plan, brainstorm, list, or read to gather ideas before I write.
- I write (stories, papers, etc.) in school.
- I have the opportunity to share my writing with my classmates (for example, peer conferencing, reading aloud, etc.).
- I have the opportunity to discuss my writing with my teacher.
- I revise and edit my writing.

The response choices to these statements are:

- Every day
- Every week
- Every month
- Every grading period
- Never (Writing Assessment Handbook: Overview, 2004).

Students randomly receive one of the nine prompts on which to write. Two 40-minute sessions over two consecutive days are set aside for completing the assignment. During that time, students are encouraged to use learned processes in order to develop and produce their "final

copy" (second draft), which is transcribed onto the two pages of the assessment folder by the end of the second 40-minute session. Students are permitted to use a dictionary, thesaurus and electronic spell checker. Although elements of the writing process are incorporated into this assessment, it is a large-scale, on demand performance assessment, and they are not permitted to get assistance from teachers or peers (*Writing Assessment Handbook: Overview*, 2004).

2.1.4 Linking Assessments

Since the purpose of this study is to explore the ability of the NSRE English/Language Arts to predict PSSA Reading scores, it is necessary to understand how educators judge the quality of assessments as well as the methodology for linking them. Three questions govern the quality of an assessment:

- How accurate is the information?
- How confident can one be in the conclusions drawn about students or programs?
- Is the assessment fair to all students who take it? (Stecher, et. al., 1997)

These questions correspond to the domains of reliability, validity, and fairness.

2.1.4.1 Reliability

First, there is reliability, defined as the measure of the degree to which an individual measurement is free from error. The importance of reliability is reflected in the simple fact that every measurement tool is imprecise and, in order to ascertain a realistic view of one's results, researchers need to know how much error there is likely to be (Stecher, et. al., 1997).

Reliability measurements can occur in various forms: test-retest reliability, parallel-forms reliability, and rater reliability. In essence, techniques for measuring reliability rely on repetition as a means of validating test results.

Test-retest reliability informs the researcher whether the same results would be produced if the assessment were given again.

Parallel-forms reliability refers to whether the same results would be produced if students responded to a highly similar assessment.

Rater reliability asks whether the same scores would be assigned if a different group of experts were to read the student responses (Stecher, et. al., 1997). This form of reliability improves as educators become more familiar with the assessment and develop more accurate scoring procedures.

Not only do these three factors contribute error to a student's score, they may also have a cumulative effect.

Reliability can be represented as a numerical index, estimated mathematically on a scale from zero to one, with one representing the highest possible reliability. Commercially distributed multiple-choice tests usually produce a reliability score of .80 and above using test-retest or parallel-forms methodologies (Stecher, et. al., 1997).

On the other hand, performance tasks tend to be less reliable.

First, performance assessments—such as essays, projects, and portfolios—are scored by human raters rather than the machine-scoring technique that usually attends multiple-choice or selected-response assessments (Stecher, et. al., 1997).

Another reason for lack of performance assessment reliability relates to student performance. In general, students do not perform consistently on performance tasks and their responses are more varied. As the complexity and scope of the task increases, the consistency of student performance declines. To alleviate the error of inconsistency, more tasks are needed to

produce a reliable score, which translates into more development time, more classroom time, and greater cost.

Third, performance tasks take more time to complete than traditional examinations and, consequently, less information can be gathered in a given amount of time (Stecher, et. al., 1997).

2.1.4.2 Validity

Validity has been considered the most important consideration in the evaluation of any assessment (Linn, 1996). Validity refers to the inferences being drawn from the applicable test score. An inference from a score is considered valid if it is justified. While reliability illustrates the accuracy of measurement, validity illustrates the way scores are interpreted by users (Stecher, et. al., 1997). The significance of this is that assessments that are valid for one purpose may not be valid for another. For instance, a student may consistently make mistakes on written word problems. The resulting score would be reliable as to the consistency of the errors, but the score would not necessarily be valid as an indication that the student does not know how to solve word problems in general. The score may have been a function of reading difficulties instead of technical mathematical know-how (Stecher, et. al., 1997).

Yet again, the choice between multiple-choice examinations and performance assessments highlights validity issues. Multiple-choice and other selected-response tasks limit the assessment to a rigid format, which can in turn limit the types of skills that are measured. Performance tests, or constructed-response assessments, present challenges that more closely match the activities performed in practice (Stecher, et. al., 1997). Proponents of performance assessments refer to this feature when they say that a test offers "authentic" tasks.

Techniques for establishing validity include the following approaches: content validity, concurrent or predictive validity, construct validity, and consequential validity (Stecher, et. al, 1997).

Content validity refers to the ability of experts to review the content of the assessment and confirm that it is measuring the desired skills or behaviors.

Concurrent, or predictive, validity means that results can be compared with performance in a work setting. Concurrent validity is a simultaneous comparison while predictive validity is a future comparison. In both cases, the idea is that a meaningful score will positively relate to real performance (Stecher, et. Al., 1997).

Construct validity measures how well the test results compare with those from other high-quality assessments of similar or dissimilar skills. When assessments of similar skills produce similar results while assessments of dissimilar skills do not, greater confidence can be placed in the new assessment's ability to accurately measure skills. This approach may be appropriate when the concepts under observation are complex and difficult to define or when successful performance is a matter of interpretation and judgment (Stecher, et. Al., 1997).

With consequential validity, consequences from using the assessment can be examined to shed light on the meaning of the results (Stecher, et. Al., 1997).

2.1.4.3 Fairness

An assessment is unfair, or biased, if students of relatively equal skill before differently on a particular question because of experience or knowledge not related to the underlying skill. Under this definition, it is important to understand how factors like family background and experience might affect the scores of test takers (Stecher, et. al., 1997).

Detecting bias is not easy. The most common technique involves expert reviews, wherein committees trained to spot troublesome factors spend time reviewing proposed assessments.

A common issue of fairness has been expressed in terms of access to instructional opportunities. The issue may also be expressed as the degree to which students are provided with the needed instructional supports to prepare them for the assessment.

For example, In *Debra P. v. Turlington* (1981), the plaintiffs brought a class action lawsuit against the State of Florida, challenging the state's right to impose the passing of the examination as a condition precedent to the receipt of a high school diploma. More specifically, the plaintiffs asserted that the state could not constitutionally deprive public school students of their high school diplomas on the basis of an examination that covered material not taught in the curriculum. Since the group failing the examination included a disparate number of African-American students, the plaintiffs alleged violations of the due process and equal protection clauses of the 14th Amendment, as well as violations of Title VI of the Civil Rights Act and the Equal Educational Opportunities Act. The Fifth Circuit Court of Appeals held in favor of the plaintiffs, that the state could not constitutionally deprive its students unless it had submitted proof of the curricular validity of the test.

2.1.4.4 Linking

Linking is a generic term that refers to making statistical comparisons between the results from one test or set of assessment tasks to those of another (Linn, 1996). The comparisons can take a variety forms and serve different purposes.

There are a variety of approaches for linking, including but not limited to: anchoring, benchmarking, calibration, equating, prediction, projection, scaling, statistical and social moderation, and verification. Several of these approaches require a basic explanation.

An understanding of these concepts is best understood using the equating approach as a starting point. If two assessments satisfy the assumptions of equating, then the results can be used for nearly any comparative scenario (Linn, 1996). Equated scores can be used interchangeably such that any use or interpretation for scores on Test A will also work for the equated scores on Text B (Linn, 1996). Generally, when an assessment consists of a small number of tasks, it is more difficult to approach the goal of equating. An example of equating is where new versions of a state test are introduced each year and research is undertaken to make sure the requirements in one year are equivalent to those of previous years (Linn, 1996).

Anchoring operates as a form of equating. In anchoring, an anchor test, A, is administered along with Form B and Form C. Forms B and C are versions of the same test given to separate groups of students. The anchor test would also be given to each group and its purpose is to increase the precision of the equating and to adjust for differences in the results of students taking Forms B and C. The closer its relationship to the two forms being equated, the better the anchor test works. Where the anchor test has a stronger equivalence to one test over the other, the two forms cannot be strictly equated (Linn, 1996).

Calibration provides a method of comparing scores on tests that satisfy relatively less stringent requirements than those for equated tests. The classic calibration situation involves the desire to compare scores from a short form of a test to those from a longer form. Although the two forms of the test presumably measure the same skills, calibration assumes that they may be designed to assess performance at different levels or with different degrees of reliability (Linn, 1996).

Statistical moderation describes two different situations. In one situation, it refers to the use of an external examination to adjust teacher-assigned grades. Outside of the United States,

this approach is used "to adjust scores on examinations in different subject areas or to compute a total score for students taking examinations in different subjects" (Linn, 1996). The second situation involves making comparisons among students who take different combinations of achievement tests. In this scenario, the statistics are used to adjust scores for variations in means and standard deviations. This results in comparable scores even though "Student A might have taken a examinations in mathematics, physics, and English, while Student B took examinations in history, political science, and English" (Linn, 1996). Unlike pure equating, this case does not render the scores equivalent. Thus, Student B's preparation for history is not inferentially linked to how well he or she can prepare for a mathematics test (Linn, 1996).

Scaling becomes useful in situations where it cannot be assumed that the population taking one achievement test is equivalent to the group taking another. Averages across the two groups will likely vary. The variation could be a result of the relative difficulty of the tests, different levels of preparation for the test, or some combination of the two (Linn, 1996). Like statistical moderation, scaling the scores will not lead them to be equivalent.

Prediction methodology attempts to anticipate scores on a test based on performance on a previous assessment. Predictions lead to sound inferences as long as there is a strong relationship between the performance on one assessment and the performance of another. Interestingly, the predictions are context and group dependent (Linn, 1996). In this situation, group-dependence means that particular traits of the test group, such as gender, can influence the predicted score. Context and group-dependence introduce additional variables to the prediction calculus, as describing group characteristics is not the same thing as comparing the performance of individual students. The calculations can become more complex as they incorporate issues such as identifying specific demographic and educational variables (Linn, 1996).

2.1.5 The High-Stakes Testing Debate

The idea of implementing national standards for what students learn in public schools is the heart of the education debate and one of the most controversial issues in education reform. Not only has it sparked federal legislation and fierce public policy debates, it has also spawned extreme views from both ends of the spectrum. Proponents say standards will be the saving grace of the public educational system, while opponents fear that the implementation of standards will lead to a nationalized curriculum. All the while, researchers and educators in the field of instruction have pointed out that there is a "sameness to what is taught in public schools" (Lewis, 1995). Arguably, standards in education already existed although they were not necessarily codified and ritualized.

The codified focus on standards heralded a significant shift in policy toward education and training. Not only did standards policies attempt to spotlight the virtues of accountability, they institutionalized the idea that programs should be judged by their *outputs* rather than their inputs or the resources they used (Elmore, 1996). At the same time, evaluation and accountability external agents of the sponsoring government usually conducted the various evaluation and accountability processes. Therefore, despite the institutionalization of the idea of output- and outcome-determinativeness, evaluations almost never become a routine at the day-to-day operating level (Elmore, 1996). In fact, the idea of evaluation often requires evaluators to maintain an "objective distance" from the programs under evaluation (Elmore, 1996).

Proponents of standards initiatives rely on at least five of the following conditions to sustain their position: (1) schools will teach the skills and qualities that the standards as required by the standards; (2) the skills and qualities will be assessed in a valid way; (3) schools are focused on improving student performance as their most important goal; (4) knowledge and skill exists among educators to produce student performance; and (5) an assessment system is in place

to determine whether individuals meet the standards and the system commands the authority to influence curriculum, teaching practices, and organization in schools (Elmore, 1996).

These conditions are instrumental to the shift in ideology concerning the purpose of public schools. What these conditions imply is that the purpose of the public school is to impart a common body of knowledge, skill, and personal qualities, rather than simply providing access. The idea then is to provide a strong common set of academic experiences instead of accommodating diverse interests and aptitudes (Elmore, 1996). From this view, it also follows that various assessments may be used to measure this common set of academic experience since, after all, they are merely measuring a distinct and quantifiable corpus of information.

At the other end of the spectrum, standards are arguably "invisible" to colleges and employers, leading the skeptics of standards initiatives to question whether standards provide adequate incentives to students, teachers, parents, and administrators (Elmore, 1996). For college-bound students, upgraded standards may have a negligible impact on college admission prospects since most admission decisions are based on class rank, GPA, and aptitude tests. Although graduates exposed to a high standards education environment may be more successful in college, this benefit is postponed, at best, and mostly speculative (Elmore, 1996). Similar arguments face work-bound students, as employers are not always influenced by high school reputations or student achievement when making hiring decisions. Furthermore, the ones who do consider academic achievement are more likely to use indicators such as GPA and class rank, rather than the nebulous concept of which school possessed higher standards than another (Elmore, 1996).

The debate over national standards naturally leads into a discussion of exactly how teachers are to assist students in reaching and surpassing new education goals. Testing has been

the most common answer to the question of how to measure student achievement. As such, Anne Lewis (1995) describes several categories of standards, including content standards, performance standards, opportunity-to-learn standards, and "world class" standards (referred to as "world" standards in the Goals 2000 legislation).

Content standards establish what should be learned in various subject areas and, in general, tend to focus on measuring the degree to which students learn content through critical-thinking and problem-solving strategies rather than through rote memorization and recitation of particular facts (Lewis, 1995).

Performance standards define which levels of learning will be considered satisfactory. The governing board of the National Assessment of Educational Progress (NAEP) attempted to establish proficiency standards on it assessments by assigning numerical scales to levels of student performance (Lewis, 1995). Although controversial, some states have adopted the method for their tests. In particular, the Writing Assessment Handbook's Holistic Scoring Guide for the PSSA utilizes this idea of numerical scales and prescribing a score between the highest and the lowest possible scores as a "minimum" for student proficiency. Performance standards ask students to apply what they know in situations that mimic real life, through demonstrations, and through the compilation of portfolios (Lewis, 1995).

Opportunity-to-learn standards are concerned with the conditions and resources necessary to give students the adequate and equal opportunity to meet the performance standards (Lewis, 1995). The original Elementary and Secondary Education Act (ESEA) of 1965 also made this a priority. Opportunity-to-learn standards are critical in a "high-stakes" testing environment where test results affect promotion between grades, graduation, and job opportunities.

Finally, "world class" standards are based on the content presented to and the expectations held for students in other countries (Lewis, 1995). New Standards, formerly known as the New Standards Project, a joint effort of the Learning Research and Development Center (LRDC) at the University of Pittsburgh and the national Center on Education and the Economy, has studied actual curricula in other countries as well as student performance on international assessments. For example, New Standards catalogued its findings in 1995, focusing primarily on the education structure and subject-area content standards of school systems in France and the Netherlands (Resnick, et. al., 1995).

Standardized tests are used for a number of purposes, among which measuring student achievement is most often cited. The U.S. Congress, Office of Technology Assessment has defined a standardized test as one that "uses uniform procedures for administration and scoring in order to assure that the results from different people are comparable" (Bond, 1996). Thus, if uniform scoring and administration are used, any kind of test can be considered "standardize," whether the test involves multiple choice questions, essays, or oral examinations.

For countless students, teachers, and administrators, standardized tests wield incredible power. They can determine whether a student will be promoted to the next grade or retained; they can affect curriculum and instruction; and, under the current federal funding regime, standardized tests can influence the fiscal resources a school receives. Given the impact of standardized testing on these educational spheres, a concern of accuracy and validity from education reformers is not surprising. Regardless of the form of the test or its scoring idiosyncrasies, the questions surrounding the use of standardized tests often focus on accuracy of measurement and fair assessment of student abilities.

As a general proposition, state-level assessment regimes include one or more of the following components: criterion and/or norm-referenced multiple-choice items; several open-ended short-answer questions; and a writing sample, with the short-answer questions and the writing samples being the most recent additions. Since the 1920s, the basic forms of standardized tests have been norm-referenced and criterion-referenced multiple-choice examinations (Neill, 2000). Norm-referenced tests compare scores with a national sample, whereas criterion-referenced tests compare students against a predetermined standard of achievement.

Norm-referenced tests are often used to classify students by highlighting achievement differences between and among students "to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers" (Bond, 1996). A school system might employ this test to identify and place students in remedial or gifted programs. Teachers might also use these tests to help identify different ability levels among students for placement in various reading or mathematics instructional groups (Bond, 1996).

Norm-referenced exams are typically multiple-choice tests sold by commercial suppliers and are designed to produce curved results, with half of the students scoring above average and half below (Kelly, 1998). A representative group of students is given the test before it is made available to the public. This is the "norm group." After the test is published, the scores of students who take the test are then compared to the "norm" (Bond, 1996). Examples of norm-referenced tests are the California Achievement Test (McGraw-Hill), the Iowa Test of Basic Skills (Riverside), and the Metropolitan Achievement Test (Psychological Corporation). Each of these are normed using a national sample of students. Since norming a test is an expensive and

time-intensive undertaking, it is not unusual for test publishers to use norms for seven years before instituting a new one (Bond, 1996).

Since norm-reference exams work with national samples, these tests often have inherent problems, particularly concerning the fact that the "norm," created by using the national samples, is established before the test has actually been administered. As Lisa Kelly (1998), a Professor of Law at West Virginia College, explains:

> [P]erhaps [at this early stage] test results truly do reflect a student's position in the national continuum. However, over time, as the instrument becomes more familiar and teachers begin to teach to prepare their students to take the test, the original norm is no longer reflective of the current range of student performance. Unless re-normed frequently, test results become inflated.

This inflationary effect has been dubbed the "Lake Woebegone effect," named after Garrison Keillor's fictional town where "all the women are strong, all the men are good-looking, and all the children are above average" (Beck, 1991). John Cannell, a West Virginia physician, incidentally charted the phenomenon while treating children who suffered from depression and stress-related illnesses (Kelly, 1998). In 1987, after the standardized test scores of West Virginia students were announced as being far above the national average, Cannell was skeptical of the results since the state was known to have one of the highest illiteracy rates in the country. Further investigation revealed that most states were touting their students as way above the norm, prompting Cannell to remark, "Every state using [these exams] is testing above the national average." However, what appeared to be a statistical impossibility – that every child could be "above average" – became a reality because test scores tended to be higher when current students, whose teachers had become familiar with the exams, were compared to norming groups from the past (Beck, 1991). Ironically, this "Lake Woebegone effect" could be viewed as an inevitable consequence of conscientious teaching, as the best way to prepare

students for standardized tests is to devote class time to teaching the information found on the exams (Bond, 1996).

Test publishers regularly tout their commercial, norm-referenced achievement tests solutions to the assessment dilemmas faced by the states. Since their tests are professionally developed, they offer their tests as having sound psychometric properties compared to state-specific efforts as well as being aligned with the state content standards (Plake, et. al., 2000).

Thus, there are three major criticisms of norm-referenced tests. On the one hand, the tests, if accurately normed, are "set up to stigmatize half of the children as being below average" (Kelly 1998). On the other hand, the norms can be so old that test results at the statewide level provide a false feel-good impression that nearly all of the children are above average (Kelly, 1998). A third criticism asserts that teachers, whether by necessity to achieve "better" student results or by "teaching to the tests," eventually place too much emphasis on low level skills (Bond, 1996).

Criteria-referenced tests, on the other hand, measure a student's performance against an unwavering set of curricular standards. These assessments compare a student's knowledge to the information he or she is expected to possess rather than information that other students may or may not know (Kelly, 1998). Most Department of Education policymakers favor criteria-referenced tests over norm-based tests. They urge that content standards should be set high, assessments should be aligned with these challenging standards, and educators and students should be held accountable for the results (Kelly, 1998). Criterion-referenced exams are likely to provide a better match between the test and the state's curriculum standards, as well as providing "better validity evidence for resultant scores" (Plake, et. al., 2000). The possible downside is the

heftier financial and resource burden on the state compared to the cost of purchasing a commercially available test.

A third form of assessment is by portfolio review. Under the portfolio review system, samples of the student's work throughout the year are collected and reviewed by scorers outside of the school to determine the student's evolving knowledge and abilities (Kelly 1998; Wiley & Haertel, 1996). In designing portfolios, some specifications must be made about the kinds of materials to be included, how many of each kind, and what criteria a set of materials must meet in order to be included (Wiley & Haertel, 1996). Since the entire portfolio must be evaluated, scoring presents a unique complexity. Schools have overcome the scoring hurdle by evaluating each component of the portfolio separately and combining these evaluations. Schools have also undertaken holistic assessments of the entire collection (Wiley & Haertel, 1996).

This form of assessment arguably provides the most accurate picture of the school's curriculum content as well as the student's demonstrated abilities. Among its disadvantages, however, is the subjectivity of the review, the cost of review, and, for those advocates of standardization, an inability to state the results in quantifiable, comparative terms (Kelly, 1998).

For these reasons, portfolio review is rarely used. For example, in the 1995-1996 academic year, only California, Michigan, New Mexico, and Vermont included portfolio assessments as a part of their testing apparatus. Among these few, California, New Mexico, and Vermont did not require statewide use of portfolios, choosing instead to leave the decision of employing them to the local school districts (Kelly, 1998).

Assessment questions can take two basic forms, regardless of whether the tests themselves are norm-referenced or criteria-referenced. These two forms are performance-based or multiple choice.

Under performance-based testing, the student is required to demonstrate directly what he or she can do (Kelly, 1998). For example, in a writing assessment, a multiple choice test may be able to test for knowledge of the rules of standard grammar or they may present a student with choices for the best missing sentence in a particular paragraph. However, they cannot measure the student's ability to compose, organize, and write his or her own paragraph on a given or self-selected subject. Performance-based assessments place the child in the position of demonstrating his or her abilities in these previously untested skills. Testing situations do not have to be all or nothing; some states and school districts choose test instruments that include both multiple choice and performance-based questions (Kelly, 1998). One problem with performance-based testing may be that it is more costly than the traditional multiple-choice tests.

An interesting example of a performance task is a mathematics and literacy exercise for fourth graders, known as the *Aquarium Task*. This assignment was developed in 1992 by the New Standards Project, a consortium of over 20 states and school districts. The task operates in the following manner: [A letter from the principal announces] (Resnick & Resnick, 1996).

> . . . that the fourth-grade classroom will be getting a 30-gallon aquarium. The students in that classroom have the responsibility of buying fish for the tank. The class will receive $25.00 to spend on fish and a *Choosing Fish for Your Aquarium* brochure. The brochure provides the necessary information about the size of each type of fish, how much each costs, and the special needs of each fish. The students are instructed to choose as many different kinds of fish as possible and then to write a letter explaining which fish were chosen

Interestingly, the students who worked on the problem wondered how anything that was so much fun could be a test. Meanwhile, their teachers were often surprised that their students could do anything so complex. For educators, such exercises are useful because they provide a direct means of investigating the ways in which students solve problems (Resnick & Resnick, 1996).

Unlike performance tasks, most multiple-choice questions reward students for memorizing factual information and learning to distinguish between stimuli with similar characteristics. Preparation for these questions tend to be segmented rather than integrated, meaning that the focus of the preparation is likely to be on single and possibly unrelated facts rather than on a "synthetic construction of inter-concept relatedness" (Crehan, 1991). Multiple-choice tests have a "dubious effect" on student motivation since the actual testing situation is "somewhat dissimilar to the normal instruction setting" (Crehan, 1991).

Although multiple-choice tests are time consuming to create, due to the necessity of having a large number of items, they are by nature very easy to score objectively. The downside is that students typically do not have access to their test papers to review following test administration and scoring. While this helps to maintain security, it makes it difficult for students to relate to feedback, further removing the assessment from the instructional process (Crehan, 1991).

The content validity of a multiple-choice test depends on well developed the content specifications are and how well the test items are written or selected to follow those specifications. Content standards should inform the test and not be limited by it. When content specifications are limited to the measurability of the multiple-choice items, then the instructional validity of the specifications is inherently suspect (Crehan, 1991).

Educators generally disfavor multiple-choice examinations, largely because the questions tend to measure "predominately lower level learning outcomes at a micro-level" (Crehan, 1991; Khattri & Sweet, 1996).

Theoretically at least, multiple-choice exams serve useful functions. Correctly constructed, they can be used to measure the outcome of higher level thinking, especially if "the

interpretive exercise or testlet format is used" (Crehan, 1991). An interpretive exercise consists of a small number of multiple choice items designed to measure "interpretation of a novel stimulus presentation," such as short reading passages, maps, steps in a laboratory experiment, or cartoons (Crehan,1991).

Also, multiple-choice tests are free "from the confounding of expressive skills and lack of knowledge" (Bock, 1996). With essay questions, it is not always clear when a poor response is the result of the student's failure to convey his or her thoughts in writing or an absence or confusion of ideas about the topic. Although multiple-choice problems may be confounded by reading comprehension issues, they are usually free from the ills that routinely plague open-ended written exams (Bock, 1996).

Another benefit of the multiple-choice exam is that it generally does not require a high level of motivation. Often this occurs because the items are short, require relatively little reading, and offer immediate positive or negative reinforcement when a correct (positive) or incorrect (negative) answer is chosen from the list of alternatives (Bock, 1996).

2.1.6 The New Standards Project

The New Standards Project (NSP) was founded in 1990. By 1995, The New Standards Project had grown into a partnership of over twenty states and urban school districts (Spalding, 1995). In total, this amounted to nearly half of all United States school children (Spalding, 1995). Spearheaded by Lauren Resnick of the Learning Research and Development Center at the University of Pittsburgh and Marc Tucker of the National Center on Education and the Economy, the New Standards Project explored alternative ways to assess student learning: portfolios, performance tasks, and projects (Spalding, 1995). The project's goal was to develop a performance-based assessment system linked to a set of high national standards. Accordingly,

the New Standards Project sought to enhance curriculum, instruction, and student learning as teachers and students developed a shared understanding of the standards, how they are embodied in student work, and how the quality of that work should be judged (Spalding, 1995).

New Standards has aspired to develop "tests worth teaching to" (Spalding, 2000). To reach this goal, the organization sponsored a multitude of meetings in the early 1990s, inviting thousands of educators to assist with building comprehensive assessments (Spalding, 2000). Its research and development units were housed at various sites across the country, among them the University of Pittsburgh, the University of California at Berkeley, and the National Council of Teachers of English in Urbana, Illinois, which housed the Literacy Unit (Spalding, 2000). Thus, the New Standards Project enabled educators, teachers, and reforms from Maine to California to "meet, pool resources, and share expertise" (Spalding, 2000).

At the outset, the Project initiated a series of national meetings, each focusing on a particular segment of the development process. These meetings included teachers, administrators, assessment directors, educators and measurement and subject-area specialists (Spalding, 2000). The teachers brought loads of student work, which were then passed around and read by all the other participants. The participants, in turn, discussed and argued about what constitutes high-quality work, how the traits of high-quality work might be incorporated into a rubric, and how such a rubric might eventually be scored (Spalding, 2000). Typically, emotions ran high as teachers and educators were challenged, perhaps for the first time in their careers, to explain and defend their instructional and assessment modalities to their peers (Spalding, 2000).

Students take the New Standards Reference Examination in grades 4, 8, and 10. New Standards sought to create an assessment system comprised of the "three P's" of performance examinations, portfolios, and projects. Partners in the New Standards consortium (the states and

school systems) would then implement all or parts of the New Standards system to reach their accountability goals.

The New Standards Project defined a "standard" as a "criterion for an acceptable outcome" (Wiley & Resnick, 1997). In the education context, goal-based instruction produces learning outcomes. The criterion for a successful outcome incorporates "*what* as well as *how much* is learned" (Wiley & Resnick, 1997). Seeking to avoid associating standards with measurement targets and test blueprints, New Standards sought to specify the "desired contents of learning and exemplify student performances that successfully meet those standards" (Wiley & Resnick, 1997).

In order consider an assessment to be standards- or criterion-referenced, New Standards identified essential four elements that must exist: (a) a set of standards, (b) a definition of measurement targets, or constructs, derived from these standards, (c) a test blueprint for a test that yields scores for estimating the status of test respondents with respect to these constructs, and (d) criteria for successful performance in terms of these scores (Wiley & Resnick, 1997).

Two assessment features of the New Standards Project were (1) a paradigmatic emphasis on performance assessments and (2) the use of portfolios, mentioned earlier as being rarely implemented. Endeavoring to contrast its assessment system with traditional multiple-choice tests, New Standards' goal was to create performance assessments that would require students to engage in tasks "that mirror as closely as possible the conditions under which a particular competence is performed in 'authentic' settings" (Simmons & Resnick, 1993). Otherwise, Simmons and Resnick (1993) argue, the meaning of content standards would be subject to interpretations, which, if allowed to vary, "would undermine efforts to set high standards for the majority of American students."

Lauren Resnick (1994) set out the differences between "traditional tests" and performance assessments. First, performance assessments are intended to function as integrated elements within the overall education system, rather than as external monitors of the system. In this way, performance assessments should maintain their validity even when they are "taught to." Lauren Resnick described this in the following way:

> [W]e got ourselves into a Catch-22 in this country by using forms of assessment that weren't designed to be taught to. Teachers were told: "Raise the scores but don't prepare the kids for the test." A great deal can be done by changing those tests to ones worth teaching to, as we hope ours are…That testing drives instruction is usually pointed to as a negative. I believe it's a positive (O'Neil, 1993).

The second difference emanates from a more theoretical and epistemological framework, namely that traditional testing is rooted in assumptions of associationism, as expressed by the psychological writings of Edward L. Thorndike, where as performance assessment is more aligned with the pragmatism of John Dewey, George Herbert Mead, and theorists of situated cognition (Resnick, 1994).

Associationist epistemology is grounded in the view that knowledge and skill can be fully characterized in terms of collections of separate bits of mental associations or stimulus-response pairs. Competence functions as a consequence of internally represented knowledge. In Thorndike's view, connection making and practice in using right habits help to establish reasoning in arithmetic. Memorization and rigorous drilling may play positive roles (Resnick & Resnick, 1996). Testing under this view identifies "traits" or abilities that are innately unrelated to the context in which one finds the observed performance (Resnick, 1994).

Pragmatic epistemology assumes that the particular context or environment in which performance occurs is inseparable from the concept of competence. To perform well is to meet or surpass an established criterion in a particular environment. Tools, people, and institutional

demands impact an individual's preparation in order to produce an outcome. Performance assessment is therefore focused more on certifying accomplishments than on identifying traits of individuals (Resnick, 1994).

Elizabeth Spalding (2000), who served as onsite coordinator of the New Standards' Literacy Unit from 1991-1996, participated in the development of the English language arts portfolios. Initially, the portfolio development suffered from what appeared to be an impossible task; that is, creating a workable scoring design. As Spalding relates it, many teachers, who were understandably excited by their students' portfolios, discovered that these classroom gems did not translate well into other contexts.

> For example, at one meeting teachers and staff 'oohed' and 'aahed' over the vividly colored an exquisitely detailed botanical drawings produced by a high school student after reading Hawthorne's 'Rappaccini's Daughter' and included in the portfolio as a response to literature. They were beautiful, but…ultimately unscoreable (Spalding, 2000).

In terms of illustrating the processes of reading and writing, showing growth in these areas, and inviting students to reflect on their learning, the classroom was well suited to the value of portfolio work. Yet, at the New Standards national meetings, teachers who were overjoyed with the voluminous output of their students were nearly embarrassed when outside readers found the portfolios incoherent or, worse, boring (Spalding, 2000). A five or ten pound portfolio was understandably resistant to the purposes of large-scale assessment and quantifiable measurement. Despite the setbacks, each partner state and district, at the end of the 1994-1995 field trial year, held meetings at which teachers evaluated student portfolios according to New Standards scoring rubrics. From these meetings came fourth-, eighth-, and tenth-grade samples to be analyzed in a national benchmarking conference (Spalding, 2000).

In the end, the English language arts design emerged as a reference examination consisting of a multiple-choice component of text-editing skills and reading comprehension and a performance component calling for open-ended responses to reading and a response to a writing prompt.

3.0 CHAPTER

3.1 Methodology

3.1.1 Introduction

The literature pointed out that standardized test instruments are being used to annually test students' academic achievement of standards and to determine placement of students or academic needs. The Pennsylvania System of School Assessment (PSSA) is being used as a performance measure to measure reading skills in grades five, eight, and 11. The New Standards Reference Examination (NSRE) English/Language Arts is used as an alternate performance measure for grade ten, The NSRE findings are used to predict achievement on the PSSA in the eleventh grade and to plan for student needs. Empirical assessment of the ability of the NSRE to predict PSSA performance is lacking, implying the need to determine the relationship between the two tests and the predictive ability of the NSRE. This research study is designed to investigate these variables. This chapter presents the research design inclusive of sample, variables and the statistical analysis used to answer the following research question and sub-questions: What is the predictive validity of the Grade Ten New Standards Reference Examination, English/Language Arts in relation to the Grade Eleven Pennsylvania System of School Assessment, Reading Test in the Pittsburgh Public Schools?

- Is there a difference in the relationship between male and female proficiency agreement?
- Is there a difference in the agreement of proficiency levels between student scores of various socioeconomic groups?
- Is there a difference in the agreement of proficiency levels between African-American and White student scores?

3.1.2 Sample Population

The sample population consisted of 1,648 Pittsburgh Public School students who took both tests (NSRE and PSSA) at grades 10 and 11 in 2003 and 2004. The demographics and performance of that sample can only be inferred looking at the grade 10 and 11 demographics and performance in general. There were approximately 32,000 students enrolled in the Pittsburgh Public Schools during the 2002-2003 and 2003-2004 school years. Other District demographics reveal that 78% of its student population receive free and reduced lunch while 67% of its population are African-American.

3.1.3 Variables

- The independent variable will be the NSRE
- The dependent variable will be the PSSA
- Demographic variables
    - Socio-economic
    - Gender
    - Ethnicity

3.1.4 Instrumentation

The Pennsylvania System of School Assessment (PSSA) is being used as a performance measure in several grades. The PSSA is used to measure reading skills in grades five, eight, and 11. The PSSA is a standards-based criterion-referenced assessment that measures the student's

attainment of academic standards and the degree that school programs help students attain proficiency levels (Brunner, 2003). The Pennsylvania Academic Reading Standards are listed as follows:

- Students will learn to read independently;
- Students will learn to read critically in the content areas;
- Students will analyze what they read to interpret the literature;
- Students will be able to identify types of writing (narrative, informational and persuasive);
- Students will identify quality of writing (content, organization and style);
- Students will be able to listen, speak, discuss and present literature;
- Students will identify characteristics and functions of the English language (word origins, variations and applications); and
- Students will be able to research information (location of information, selection and organization) (Pennsylvania Department of Education, 2004).

PSSA Reading scores based on the Pennsylvania Academic Standards will be used for this study. By way of illustration, the Writing Assessment Handbook for the PSSA (2004) includes several kinds of writing prompts, notably a narrative/imaginative prompt, an informational prompt, and a persuasive prompt that are reflective of tasks that students are expected to perform as an indicator of having attained the standards. Figures 1, 2 and 3 are samples of the aforementioned writing prompts respectively.

Prompt 1

We all have memories connected to our experiences. Think about an experience you feel you'll always remember. Try to picture the time, the place, and the people involved. Try to remember everything you can about this experience.

Write about the experience you remember. Be sure to include enough details so that your reader can share your experience. Show why this memory stands out for you.

Figure 3-1: Sample Narrative/Imaginative Prompt

As you write and rewrite your paper, remember to:

- describe what happened
- give details that are specific and relevant to this experience.
- present your ideas clearly and logically.
- use words and well-constructed sentences effectively.
- correct any errors in spelling, punctuation and capitalization.

Prompt 2

Think about discoveries or inventions that have affected our lives. Select one.

Write to inform someone about this discovery or invention. Tell whether it has been good or bad for society.

Figure 3-2: Sample Informational Prompt

As you write and rewrite your paper, remember to:

- give enough information so that the reader will know what the discovery or invention is and why you chose it.
- give details that are specific and relevant to the discovery or invention.
- present your ideas clearly and logically.
- use words and well-constructed sentences effectively.
- correct any errors in spelling, punctuation and capitalization.

Prompt 3

A new principal is contacting all students about changing or adding to the school rules. Think of a rule you would like to change or add.

Write to persuade the principal to use your suggestion.

Figure 3-3 : Sample Persuasive Prompt

As you write and rewrite your paper, remember to:

- state what rule you wish to change or add.
- include enough convincing details so the principal will want to
- use your suggestion.
- present your ideas clearly and logically.
- use words and well-constructed sentences effectively.
- correct any errors in spelling, punctuation and capitalization.

The New Standards Reference Examination (NSRE) English/Language Arts is used to measure reading performance in grades 4, 8, and 10. The test measures English Language Arts (Reading: Basic Understanding; Reading: Analysis & Interpretation; Writing; and Conventions). Test scores tell how well students perform relative to standards. The standards for the NSRE, which are based on national standards are as follows:

- Students read a wide range of print and non-print texts to build an understanding of texts;
- Students build a wide range of literature from many periods and genres;
- Students apply a wide range of strategies to comprehend, interpret, evaluate and appreciate texts;
- Students adjust their use of spoken, written and visual language;
- Students employ a wide range of strategies as they write and use different writing process elements appropriately to communicate to different audiences;

- Students apply knowledge of language structure, language conventions , media techniques, figurative language and genre to create, critique and discuss print and non-print texts;
- Students conduct research on issues and interests by generating ideas and questions and posing problems;
- Students use a variety of technological and information resources;
- Student develop an understanding of and respect for diversity in language use, patterns and dialects across cultures, ethnic groups, geographic regions and social roles;
- Students whose first language is not English make use of their first language to develop competency in the English language and to develop understanding of content across the curriculum;
- Students participate as knowledgeable, reflective, creative and critical members of varied literacy communities;
- Students use spoken, written and visual language to accomplish their own purposes.

Scores from the NSRE test will be used for this study. By way of illustration, the Practice Test for the NSRE (1997) includes several kinds of writing prompts, notably an independent writing prompt, an informational or reading and writing prompt, and a comprehension and editing prompt that are reflective of tasks that students are expected to perform as an indicator of having attained the standards. Below are listed samples 1, 2 and 3 of the aforementioned writing prompts respectively.

3.1.5 Sample of an Independent Writing Prompt

The first section, Preparing to Write, will help you think about what you are to write. The second part, Your Writing Task, tells you exactly what to write. In evaluating your writing, scores will look for evidence that you can:

- Express your ideas clearly
- Organize your ideas and make them easy to follow
- Choose words carefully to express what you want
- Use correct spelling, grammar and punctuation

3.1.6 Sample of a Reading and Writing Prompt

This part of the test is designed to see both how well you understand what you read and how well you can write. You will read a short passage, answer several questions, and finish with a short essay. You will want to save time for the essay because you will receive both reading and writing scores for this part of the assessment When you have finished reading, answer all the questions about the reading passage as fully as you can. The people who evaluate your test will be looking for:

- what you understand about the reading passage
- how you use references to the selection as well as your ideas and experiences to support your interpretation of the passage
- how you present your ideas
- how you use specific details to support your ideas

3.1.7 Sample of a Reading Comprehension and Editing Prompt

This is a test of your ability to understand and interpret what you read. This test will also evaluate your ability to edit a sample of student writing. This section requires you to fill in the circle beside the answer that you choose.

- Which of these best describes the purpose of the article?
    - To show how important medicine can be.
    - To describe the life of John Pemberton.
    - To show the benefits of carbonated water.
    - To describe the origin of Coca-Cola.
- Which of the following words could be taken out of the first sentence without changing the meaning of the sentence?

- Back
- Invented
- First
- Soft (Harcourt Brace, 1997).

3.1.8 Data Collection

The 2002 and 2004 archived assessment data for the NSRE and PSSA will be used for analysis with respect to this study.

3.1.9 Design and Analysis

Quantitative archival research, a non-experimental design will be used for this study. The experimental design allows for the control over variables and threats to validity, while the non-experimental design does not, however both yield empirical results. Empirical research includes the collection of data and the analysis of the data to answer a research question or hypothesis. Student test scores for tenth (NSRE English/Language Arts scores) and eleventh grade students (PSSA Reading scores) will be reviewed to answer research questions and test hypotheses. In this case, the regression analysis enables this researcher to quantify the relationship between the scale scores on the NSRE and the scale scores on the PSSA

The research design used will utilize the mean test of significance and regression analysis with an examination of the inter-correlations of the demographic groups and the outcome variables. The PSSA scale scores will be regressed on the NSRE scale scores. Differential impact of the variables will then be assessed to see if there are differential predictive validity of various subgroups. Another way to examine how well one variable predicts another is to see if the variables identify similar levels of achievement. If there is predictive validity, both the NSRE and the PSSA should place students in the same performance level categories. Crosstabulation analysis can be used for these comparisons. Cross-tabulation is a combination of

two (or more) frequency tables arranged such that each cell in the resulting table represents a unique combination of specific values of cross-tabulated variables. Thus, cross-tabulation allows this researcher to examine frequencies of observations that belong to specific categories on more than one variable. By examining these frequencies, this researcher can identify relations between cross-tabulated variables.

Cross-tabulations of the PSSA and the NSRE proficient and performance levels will be run overall and for all demographic subgroups. An examination of the mapping of performance levels will be necessary to carry out this analysis. To that end, an examination of agreement and non-agreement will aid in this mapping.

Individual student growth or decline on the PSSA relative to the NSRE will be examined overall and for all subgroups. A non-parametric sine test will be used to determine significance. The regression analysis of scale scores will or will not highlight the statistical significance of the predictive validity of the NSRE relative to the PSSA. Even if there is statistically significance agreement in the cross-tabulation cells of proficient/proficient or non-proficient/non-proficient, this analysis will determine whether or not the findings may be due to chance.

4.0 CHAPTER

4.1 Results

4.1.1 Introduction

This study sought to investigate the predictive accuracy of the overall New Standards Reference Examination (NSRE) English/Language Arts for determining performance on the Pennsylvania System of School Assessment (PSSA) test. This study utilized one set of population data from an urban public school district (Pittsburgh Public Schools). The set consisted of performance results from the 2003 NSRE administered in the tenth grade and performance results from the 2004 PSSA administered in the eleventh grade. This study presumed that most of the students that performed on the NSRE also performed on the PSSA the following year. A total of 1,648 students were administered both the NSRE and the PSSA in grades 10 and 11 during the spring of 2003 and 2004 respectively. The purpose of this Chapter is to present the central research questions about the relationship between recent grade 10 achievement tests and grade 11 achievement tests in an urban public school setting, and then present the empirical results that answer these questions. The chapter is organized as follows: Section 4.1 presents the key questions that are restated as hypotheses, Section 4.2 presents the statistical methodology to be

used to answer these questions, and Section 4.3 presents and discusses the results. By way of summary we find that the NSRE is an overall valid predictor of how well students will or will not perform on the PSSA.

4.1.2 Section 4.1 Key Questions

Since most school districts and the district of interest, the Pittsburgh Public Schools, also purchase and administer national standardized tests, it is of interest to enquire if these standardized test results from national tests accurately predict performance on state assessment tests. In particular, it is of interest to investigate whether or not the NSRE accurately predicts the overall reading performance of students on the PSSA. The NSRE is of further interest because its assessment process creates, as a byproduct of the testing, a student-by-student evaluation of each student's strengths and weaknesses in the area of language arts. Since the NSRE accurately predicts PSSA reading performance, it should be possible to utilize the NSRE individual assessments of weaknesses to develop, student by student, a strategy of intervention that will lead to improved PSSA reading performance.

Critical to deciding whether or not this is a sound strategy, is knowing at the outset whether or not performance on the NSRE language arts assessment accurately predicts PSSA performance, and whether or not the statistical relationships are not only statistically significant and in the expected direction, but also large and of consequence. Also of interest is whether or not the relationship between NSRE test results and PSSA test results as strong and reliable is whether or not they vary by gender, ethnicity, and socio-economic status. To that end, the data analyzed are the matched test results of grade 10 students who took the NSRE in Spring, 2003 and the PSSA in Spring, 2004. Since the NSRE seeks to measure academic achievement at the

end of grade 10, and the PSSA examination seeks to measure academic achievement at the end of grade 11, we expect there to be a positive relationship between the two assessments. Since this researcher is viewing the New Standards examination as a predictor and also as a diagnostic, I do not want to include in the analysis the intervention of differences in training in 11th grade before the administration of the PSSA.

To begin our analyses it is important to note that based on national, state, and district level information about the achievement gap, this researcher expects that *within* the NSRE Total Language Arts results and *within* the PSSA Total Reading Test results, to find that students with lower socio-economic status (measured by free and reduced lunch) will score lower on the total test than their counterparts, that African American students will score lower than White students on the total test, but do not expect there to be differences by gender in total test results. As we shall see when we examine the data, this commonly is not true for both tests in the Pittsburgh Public Schools.

4.1.3 Hypotheses

H1: There are no differences in PSSA scores between the group of students who took both tests, (the NSRE in grade 10 and the PSSA in grade 11) and those who just took the PSSA in 11th grade).

H2: NSRE Assessments for Free and Reduced Lunch Students will be lower than for Regular Lunch Students.

H3: PSSA Assessments for Free and Reduced Lunch Students will be lower than for Regular Lunch Students.

H4: NSRE Assessments for White Students will be higher than for Black Students.

H5: NSRE Assessments for Female Students will be the same as for Male Students.

H6: NSRE Assessments by Socioeconomic Status for Regular Lunch Students will be higher than for Free and Reduced Lunch Students.

H7: New Standards Language Arts scores positively predict PSSA test scores.

The second set of hypotheses involves whether or not there is a positive relationship between NSRE Language Arts Assessments and overall PSSA Reading Assessments. Since both examinations measure language arts achievement, although at 10^{th} and 11^{th} grade respectively, it seems most likely that the relationship will be positive, but less than one to one.

4.1.4 Section 4.2: Statistical Methodology

In order to thoroughly investigate hypotheses H1-H7, this researcher shall use several common statistical techniques: the test of the difference between two sample means, and bivariate multiple regression analysis. In general we are interested in examining the NSRE assessment results for the number of observed students, and in the relationship between the NSRE and the subsequent PSSA test results. This researcher shall denote different demographic subgroup results through subscripts. PSSA then divides into $PSSA_{Male}$ and $PSSA_{Female}$.

4.1.5 Section 4.2.1: Testing for Differences Between Means

Hypotheses H1-H7 inquires about whether or not various subgroup assessment results are the same. By "the same" we mean that some characterization of the sample from each of the subgroups shows that they are nearly the same in arithmetic value so that any observed difference could be ascribed to chance rather than systematic difference. The usual way to answer this statistically is to inquire if the sample means of each group are the same or not. Such a test is constructed by stating the negative of what one expects and then to see if it withstands

statistical scrutiny. Thus the *null hypothesis* is that the two sample means are the same, and one tests to see if the observed differences are sufficiently large, taking into account the observed variance in each sample mean, to warrant the inference that they are different in the expected direction. This researcher shall calculate the means for various subgroups and report whether or not the differences are statistically different from zero and at what confidence level or degree of reliability this is believed to be true.

4.1.5.1 Section 4.2.2: Multiple Regression Analysis of Natural Logarithms Models by Sex, Face, and Socio Economic Status

In order to test H7 the researcher will summarize the relationship between the NSRE and the PSSA, by estimating a least squares, linear relationship of the form:

$$PSSA = B_1 + B_2\ NSRE + \epsilon \qquad (1)$$

Where B_1 and B_2 are to be estimated and ϵ is a random disturbance term distributed with a mean of 0 and variance of 1.0. H7 says that we expect the estimated $B_2 > 0$. While both NSRE and PSSA are scaled scores, they have different ranges, means and standard deviations, although both are designed to be normally distributed. Since the researcher is interested in not only whether or not there are positive relationships that are not due to chance between NSRE and PSSA, but also how *large* the effect of NSRE is on PSSA. There is a need to state PSSA and NSRE results that overcome the difference in the way the two tests are measured. One way to do this is by statistically estimating a version of (2) in which both PSSA and NSRE are stated in terms of their *natural logarithms*:

$$\log_e PSSA = B_1' + B_2' \log_e NS + \epsilon' \qquad (2)$$

This transformation of the data allows us to again test whether or not there is a positive relationship between NSRE and PSSA, e.g. $B_2' > 0$. Because (2) is now in natural logarithms, the interpretation of the estimated B_2' is that it will tell the researcher what a 1% *increase* in NSRE will predict in terms of a *percentage change* in PSSA score. Since the researcher intends to examine not only an overall version of (1) and (2), but versions also for subgroups, we can use the difference between means test to determine, for example, whether or not B_2' for Males is statistically different than B_2' for Females, whether or not there are such differences for White vs. Black students and so forth.

Now the B_2' can be viewed as the mean of a sample, with size N, and the estimated standard error, squared, of the regression coefficient is a sample estimate of the variance.

4.8 Section 4.3: Results

Tables 4.1-4,7 report the statistical test for each hypothesis, H1-H7, by statistically testing whether or not the total group and their subgroup sample assessment results are statistically different from each other. Table 4.1 relating to Hypothesis 1 indicates that the students who took both tests scored 18% higher on average than the students who just took the 11^{th} grade PSSA test, and this observed difference is highly significant statistically. With hundreds of observations, the calculated Z of 15.71 is significant at the .001 level or better.

Table 4.1 :Statistics for 11th Grade Students on the PSSA for the 2003-04 School Year

Demographic Variable	#	Mean Score	Standard Deviation	Difference Between Mean	% of Difference	Z Score
11th Grade Students Who Took NSRE in 10th Grade (Matched)	1648	1339.29	254.02		18.3%	15.71
11th Grade Students Who Only Took the PSSA (Non-Matched)	417	1093.44	292.97			

Next, we analyzed the assessments of students by several subgroups who took both the NSRE in 10th grade Language Arts and the PSSA Reading in the 11th grade. Table 4.2 demonstrates that there are statistically significant differences in NSRE assessment results by socioeconomic status. Students who received free or reduced lunch scored on average 2% lower than those students who received regular lunch. The difference, while small, was statistically significant .at the .01 level. By contrast, Table 4.3 indicates the mean PSSA score for students on free and reduced lunch was 14% lower than those students who received regular lunch. Again this difference was highly significant with a Z Score of 14.57.

Table 4.2 : Statistics for 11th Grade Students on the PSSA by Gender for the 2003-04 School Year

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% of Difference	Z Score
11th Grade Males	957	1264.24	289.02	59.29	0.04	4.77
11th Grade Females	1056	1323.53	266.91	59.29		

Table 4.3 : Statistics for 11th Grade Students on the PSSA by Socioeconomic Status 2003-04 Sch. Yr.

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% of Difference	Z Score
Free & Reduced Lunch	852	1196.43	246.02		0.14	14.57
Regular Lunch	1161	1367.94	266.91			

This researcher next examined whether or not there are differences by race for the two assessments. Table 4.4 demonstrates that there are large (17%) statistically significant differences (with a Z Score of 20.95) between white and black students' assessment results on the NSRE. However, Table 4.5 examines the difference in scores between males and females on the NSRE. There is a statistically significant, but small difference of about 1%. Table 4.6 illustrates again, a small difference by socioeconomic status of about 2% which is statistically significant. Finally, Table 4.7 displays a statistically significant difference between white and black student of 3%.

Table 4.4: Statistics for 11th Grade Students on the PSSA by Race for the 2003-04 School Year

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% of Difference	Z Score
White	973	1412	270.44	239.9	17%	20.95
Black	958	1172.1	231.65			

Table 4.5: Statistics for 10th Grade Students on the NSRE by Gender for the 2002-03 School Year

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% Of Difference	Z Score
Female	902	147.24	6.18	1.5	0.01	4.98
Male	760	145.74	6.06			

Table 4.6: Statistics for 10th Grade Students on the NSRE by Socioeconomic Status for the 2002-03 Sch. Yrr

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% Of Difference	Z Score
Free & Reduced Lunch	667	144.51	**5.56**	3.43	0.02	11.79
Regular Lunch	995	147.93	6.18			

Table 4.7: Statistics for 10th Grade Students on the NSRE by Race for the 2002-03 School Year

Demographic Variable	#	Mean Score	Standard Deviation	Difference between Mean	% Of Difference	Z Score
White	850	148.80	5.98	4.93	0.03	17.37
Black	746	143.88	5.35			

Tables 4.8 reports the regression analysis that explains the 11'th grade PSSA scores by 10'th grade NSRE scores, while Tables 4.9 through 4.11 report the regression results for subgroups of students. Note that in the subgroup analysis in Tables 4.9-4.11, the researcher only reports B_2, since it already is evident from the analysis of differences between means, reported above that there are differences in B_1. The focus for the subgroups will be on whether or not the effect of NSRE on PSSA varies by socioeconomic status, gender, and race.

Table 4.8: Regression Analysis of the 11th Grade PSSA by 10th Grade NSRE Reading Scores

Variable	B	SE B	*T* Statistic	Pr > |t|	Adjusted R-Squared	Observations
Intercept	-11.2537	.3221	-34.94	<.0001	.6654	1647
NSRE Reading	3.6969	.0646	57.25	<.0001		

Table 4.9 Regression Analysis of the 11th Grade PSSA by 10th Grade NSRE Reading Scores by Gender

Variable	B	SE B	*T* Statistic	Pr > |t|	Adjusted R-Squared	Observations
Male Students	3.8043	.1065	35.74	<.0001	.6295	752
Female Students	3.6457	.07949	45.86	<.0001	.7017	895

Table 4.10: Regression Analysis-11th Grade PSSA by 10th Grade NSRE Reading Scores by Socioeconomic Status

Variable	B	SE B	*T* Statistic	Pr > |t|	Adjusted R-Squared	Observations
Regular Lunch Students	3.6136	.08538	42.32	<.0001	.6536	950
Free & Reduced Lunch Students	3.5925	.11029	32.57	<.0001	.6272	631

Table 4.11: Regression Analysis of the 11th Grade PSSA by 10th Grade NSRE Reading Scores by Race

Variable	B	SE B	*T* Statistic	Pr > |t|	Adjusted R-Squared	Observations
White Students	3.5290	.09494	37.17	<.0001	.6212	843
Black Students	3.4641	.10467	33.09	<.0001	.5975	738

There are several overall regularities in the regression analysis in Tables 4.8-4.11. First, NSRE explains 60% to 70% of the variation in PSSA scores. Second, the relationship is always positive and highly significant. Thus, as students do better on the New Standards Reading Examination, it is quite likely they will also do better on the PSSA reading exam a year later. Third, a 1% increase in NSRE reading scores is associated with a 3.5% to 3.7% increase in PSSA scores. Overall, a 1% increase in NSRE scaled score is associated with a 3.69% increase in PSSA score. (see Table 4.8). Table 4.12 tests the null hypothesis for each of the subgroup regression results contained in Tables 4.9-4.11 that the *effect* of NSRE on PSSA is *not different*. What emerges from these tests is the conclusion that the estimated B_2 by subgroup is statistically different; however, the differences are quite small, and vary between 5% by race to 1% by gender. These differences in effect are thus smaller than the differences between means observed for subgroups in their PSSA scores.

Table 4.12: Test of Difference Between B_2 by Subgroup

	B	SE B	Observations	Variance	Z

Gender					
Male	3.8043	0.1065	752	0.0113423	
Female	3.6457	0.07949	895	0.0063187	
Diff	0.1586				33.7
% Diff	0.0106443				
Socioeconomic					
Regular	3.6136	0.08538	950	0.0072897	
Free	3.5925	0.11029	631	0.0121639	
Diff	0.0211				4.1
% Diff	0.001464				
Race					
White	3.529	0.09494	843	0.0090136	
Black	3.4641	0.10467	738	0.0109558	
Diff	0.0649				12.8
% Diff	0.0046403				

5.0 CHAPTER

5.1 Conclusions and Implications

This study investigated the question of whether the Grade 10 New Standards Reference Examination (NSRE) held predictive validity for determining student performance on the Grade 11 Pennsylvania System of School Assessment (PSSA). Statistically, the Grade 10 NSRE is a valid predictor of student performance on the Grade 11 PSSA. The relationship between the two tests is always positive and statistically significant. Thus, as students do better on the New Standards Reading Examination, it is expected that they will also do better on the PSSA reading exam a year later.

Toward taking a deeper look at the data on the Tables in Chapter IV, four themes emerged: 1) significant differences among varying socioeconomic levels, 2) significant differences in achievement between White and African American students, and differences between male and female students, and the overall differences between the two assessments. Another important finding was that although there were significant differences related to poverty, race, and gender, those differences were small for the NSRE and large for the PSSA. Each of these will be discussed below.

5.1.1 The Influence of Socioeconomic Levels

With the receipt of free or reduced lunch as a proxy for socioeconomic status, the findings are clear that there were significant differences. This achievement gap has been identified in the literature for decades and the data in this study align with what is already known. The Rand Corporation found that the most important factors associated with the educational achievement of children are socioeconomic in nature. These factors include parental educational levels, neighborhood poverty, parental occupational status and family income. Neighborhood poverty was found to be a predictor of behavior problems among young children – problems that impede school readiness. Children in poor neighborhoods are significantly more likely to exhibit both anxious and aggressive behavior. Furthermore, it was found that improving socioeconomic circumstances of African Americans consistently corresponded to the simultaneous improvements in student achievement relative to Whites. Overall, the research indicates that the combined improvements in socioeconomic measures among African American families (including parents' education, occupation, and income) correlated with a significant decrease in the African American – White reading scores from 1972 to 1992 (Lara-Cinisomo, *et. al.,* 2004). Consequently, Pittsburgh Public Schools services a disproportionate population of low-income African American students whose performance on the district's standardized tests mirror what the research illustrates with achievement and socioeconomic status.

5.1.2 The Influence of Race

Closely related to socioeconomic influences on achievement is the influence of race. These two variables (socioeconomic status and race) are closely related because African American students are over represented in lower economic levels. The results of this study

demonstrated once again that there were significant differences between African American and White students. The results of this study indicate that the district's gap between White and African American students is even larger than the gap between poor and non-poor students.

The literature indicates that the achievement gap between African Americans and Whites has narrowed since 1970, but the typical African American still scores below 75% of American Whites (Jencks & Phillips, 1999). The gap in Pittsburgh Public Schools is large enough to have important social and economic implications. Possible reasons for this gap could be differences in instructional resources available to students in predominantly White versus students who attend predominantly African American schools or how students respond to teachers that have low expectations for African American students who are usually students who read below their grade level. Given equitable resources and higher teacher expectations, work can begin to attempt to close the achievement gap between African American and White students. To that end, eliminating the achievement would allow colleges, professional schools and employers to phase out the racial preferences that have caused so much controversy over the past generation. This study indicates a strong need for Pittsburgh Public Schools to continue its efforts to significantly lessen the achievement gap by specifically generating curriculum resources and instructional strategies that promote African Americans attaining the standards at the same rate as their white counterparts.

5.1.3 The Influence of Gender

The differences between male and females scores were statistically significant and not due to chance or coincidence, the differences were nonetheless relatively small. For gender, the differences between scores varied by 1% and 4% on both the NSRE and the PSSA respectively. In terms of a district prioritizing its goals of raising achievement for all students, gender

differences do not rate on the list for immediate attention as does the need raise student achievement the race and socioeconomic subgroups.

5.1.4 Differences in Performance on the Two Tests

This study's results showed that students who received free or reduced lunch scored on average 2% lower on the NSRE than those students who received regular lunch. By contrast, the mean PSSA score for students on free and reduced lunch was 14% lower than those students who received regular lunch. Reasons for the variance in scores between the two subgroups warrants further study in that the difference in percentages could be related to variables such as the PSSA is a timed test while the NSRE is not timed or that the NSRE is more of an "open book" test while the PSSA does not permit students to reference any other instructional resources during test administration. Moreover, it could be that since the NSRE and the PSSA are scaled differently, there are technical problems in equating scores between the two assessments. The NSRE has five proficiency levels that are not aligned to the four proficiency levels of the PSSA. Perhaps a better approach would be for the state to develop its own preliminary version of the PSSA that is written to the same standards and formatted on the same four proficiency levels as the PSSA.

Finally, the NSRE yields a raw score, but also states in instructional language those specific areas where students have strengths and weaknesses. By contrast, the PSSA only yields a raw score thereby leaving the school administrator and/or teacher with the task of analyzing and interpreting the data in each tested category in an effort to determine students' instructional strengths and weaknesses. This can ultimately lead to an inconsistent interpretation of what must be done to appropriately address the instructional needs of students.

5.1.5 Practical Implications

This study found that the NSRE is a valid predictor of performance on the PSSA. For the subgroups of race, gender, and socioeconomic status, the findings show statistically significant levels of agreement regarding proficiency and predictability between the NSRE and the PSSA. Considering these findings, the NSRE's predictive capability yields several important implications regarding the use and merit of educational assessments. Notably, the NSRE's predictive validity suggests three basic benefits – information, motivation, and goodwill.

First, the NSRE's predictive validity provides researchers and educators with information. At a general level, results on the NSRE and the PSSA provide researchers and educators with diagnostic information, allowing teachers to pinpoint strengths and weaknesses in student performance relative to the established standards. Having two assessments that are linked by proficiency standards enables students to reinforce the skills necessary to succeed on the examinations and allows for an objective measure of accountability regarding a school system's academic progress. Additionally, as discussed in Chapter II, The Elementary and Secondary Education Act of 1965 sought to lessen performance gaps based upon poverty and socioeconomic status. An assessment with predictive validity for performance aligned to the standards, such as the NSRE, informs researchers about whether the original goals of the federal law are being reached.

Next, predictive validity motivates students. While motivation may not be an end in itself, it can lead to desirable outcomes. For instance, the understanding that the NSRE is highly correlative with the PSSA may encourage students to take the NSRE seriously, to pay greater attention in class, and to study more. Students would be motivated to perform well on the NSRE because doing so would indicate high performance on the PSSA, which is critical to the student's graduation. However, use of the PSSA, as a criterion for graduation (as it is used in the

Pittsburgh Public Schools) can be problematic because it does not "add value." If a student cannot graduate because he has "failed" the PSSA, it is unlikely that the backup criterion (the NSRE) will enable the student to graduate with a passing performance. This is an important blemish because if students perform poorly on the NSRE, the data in this study indicates that those same students would also perform poorly on the PSSA. Therefore, it would not be beneficial to use the NSRE and an alternate assessment toward a students' graduation requirement if that student has failed the PSSA. However, teachers, school systems, and administrators are motivated by the NSRE's predictive validity since the goal of meeting state standards on the PSSA is important to school system's compliance with the No Child Left Behind Act of 2001.

The category of goodwill is often overlooked. As earlier discussed, the debate concerning assessments and high-stakes testing has primarily involved the question of whether the nation's students are properly educated and whether the schools are providing adequate educational tools for progress. Certainly, the general public perceives that, since public school authorities exert responsibility over students for substantial amounts of time, the public has the right to objective, impartial information about the student performance. While classroom grades may be unreliable, the predictive validity of an examination like the NSRE demonstrates that students are taking examinations that measure student performance in a quantifiable and meaningful way. Consequently, the NSRE predictive validity can lead to renewed public confidence in the school system, a demonstration of the school system's commitment to uphold standards, and a genuine sense of achievement and accomplishment felt by students who pass important tests like the PSSA.

5.1.6 Implications for Further Study

Further study regarding the relationship between the Grade 10 NSRE and the Grade 11 PSSA should evaluate the testing conditions of the two assessments and incorporate an analysis of school curricula and teacher accountability. Thus, the NSRE should and could be used as an instructional diagnostic tool by teachers to elucidate student academic strengths and weaknesses. This would then allow teachers and administrators to assess what is taught against what is actually being assessed on both the NSRE and the PSSA since the R-Squared data analyzed in Chapter IV surprisingly reveals that a 1% increase in scores on the NSRE would yield a 3.7% increase in PSSA scores. To that end, if the NSRE were utilized as a diagnostic tool across subgroups, the positive results could prove to be phenomenal in closing the achievement gap.

Additionally, further investigations of the Grade 10 NSRE's predictive validity for the Grade 11 PSSA should focus on the variable of time. That is, further attention should be given to the year of difference between Grades 10 and 11 to determine the nature of student growth between the NSRE and the PSSA. The reason for this is that the passage of time, as well as the changes that accompany it, are always present but may not be represented by statistical analysis unless a study specifically attends to it.

In particular, researchers should determine how the testing population itself may have changed or shifted, considering students that move in and out of the area, shifts in demographics, changes resulting from teacher mobility and instructional expertise, overall school performance, and individualistic student growth.

Finally, although it was hypothesized to find statistically significant differences in both assessments between subgroups, it was surprising to find that the differences between subgroups, although significant, were small for the NSRE and large for the PSSA. Further research is needed to ascertain why this was an occurrence in this study. It may be interesting to note if

similar occurrences might develop if the same analysis was conducted using empirical data from the assessments given throughout all grade levels in the areas of both reading and mathematics.

BIBLIOGRAPHY

[1] Academic Standards and Assessment. 22 Pa. Code §4.51 (2004). Beck, M. (1991, July 8). New York meets Lake Woebegone. *Newsweek,* 48-51.

[2] Belsley, D. A, Kuh, E., & Welsch, R. E. (1979). Regression diagnostics: Identifying influential data & sources of co linearity. New York: Wiley.

[3] Bock, R. D. (1996). Open-ended exercises in large-scale educational assessment. In L. B. Resnick and J. G. Wirt (Eds.), *Link school and work: Roles for standards and assessment* (pp. 305-338). San Francisco: Jossey-Bass Publishers.

[4] Bond, L. A. (1996). Norm- and criterion-referenced testing. *Practical assessment, research and evaluation*, 5(2), Retrieved November 27, 2004 from http://PAREonline.net/getvn.asp?v=5&n=2.

[5] Brown v. Board of Education of Topeka, 347 U.S. 483 (1954). Retrieved November 19, 2004, from Lexis/Nexis Academic.

[6] Brunner, B. J. (2003). The right to write? Free expression rights of Pennsylvania's creative students after Columbine. *Dickinson Law Review*, (*107)*, 891-918.

[7] Buckendahl, C. W., Plake, B. S.; Impara, J. C.; & Irwin, P. M. (2000). Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives. Paper presented at the Annual National Conference of the Council of Chief State School Officers on large-scale assessment (ERIC Document Reproduction Service No. ED 442829).

[8] Caro, R. (1982). The years of Lyndon Johnson: The path to power. New York: Alfred Knopf.

[9] Chadsey, M. (2002). Federal courts and school desegregation. *Thurgood Marshall Law Review*, *27*, 149-164.

[10] Chatterjee, S., & Bertram, P. (1997). *Regression analysis by example*. New York: Wiley.

[11] Clinton, W. J. (1999, January 19). State of the Union Address. *New York Times*, *A22*.

[12] Coleman, A. L. (1998). Excellence and equity in education: High Standards for high-stakes tests. *Virginia Journal of Social Policy and Law*, (*6)*, 81-113.

[13] Crehan, K. (1991, October). Performance assessment: Comparative advantages. Paper presented at the Annual Meeting of the Arizona Educational Research Association. (ERIC Document Reproduction Service No. ED 338710)

[14] *Data Recognition Corporation (DRC) Educational Clients*. (n.d.). Retrieved November 24, 2004, from http://www.drc-mn.com/education/clients.html.

[15] Debra P. v. Turlington, 644 F. 2d 397 (1981). Retrieved November 24, 2005, from Lexis/Nexis Academic.

[16] Dougherty, E. (1998). Getting beyond policy: School reform in practice. 6 *Virginia Journal of Social Policy and Law*, (*6)*, 127-153.

[17] Draper, N. R. (1980). *Applied regression analysis*. New York: Wiley. Education Council Act of 1991, Pub. L. No. 102-62, § 404, 105 Stat. 305, 314-15 (1993).

[18] Elmore, R. F. (1996). Policy choices in the assessment of work readiness: Strategy and structure. In L. B. Resnick and J. G. Wirt (Eds.), *Link school and work: Roles for standards and assessment* (pp. 53-78). San Francisco: Jossey-Bass Publishers.

[19] Elul, H. (1999). Making the grade, public education reform: The use of standardized testing to retain students and deny diplomas. Columbia Human Rights Law Review, 30, 495-536.

[20] Gergen, D. R. (1990). Lake Wobegon's schools. *U.S. News & World Report*, *108*, 74-78.

[21] Goals 2000: Educate America Act, 20 U.S.C. § 5801 (1997).

[22] Heise, M. (1994). Goals 2000: Educate America act: The federalization and legalization of educational policy. *Fordham Law Review*, *63*, 345-381.

[23] Hernon, P. (1991). *Statistics: A component of the research process*. New Jersey: Ablex Publishing.

[24] Huitema, B. E. (1980). *Analysis of variance and alternatives*. New York: Wiley (1980).

[25] Jencks, C., & Phillips, M. (1999). The Black-White Test Score Gap. In C. Foreman (Ed.), *The African American Predicament* (pp. 63-66). Washington, DC: The Brookings Institution Press.

[26] Jennings, J, F. (1991). Title I: Its legislative history and promise. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 1-24). Mahwah, N.J.: L. Erlbaum Associates.

[27] Kelly, L. (1998). Yearning for Lake Wobegon: The quest for the best test at the expense of the best education. *Southern California Interdisciplinary Law Journal, 7*, 41-79 (1998).

[28] Khattri, N. & Sweet, D. (1996). Assessment reform: promises and challenges. In M. B. Kane and R. Mitchell (Eds.), *Implementing Performance Assessment: Promises, Problems and Challenges* (pp. 1-21). Mahwah, N.J.: L. Erlbaum Associates.

[29] Kucerik, E. (2002). Hot topic: The No Child Left Behind Act of 2001: Will it live up to its promise?" *Georgetown Journal on Poverty, Law & Policy, (9)*, 479-487.

[30] Lewis, A. C. (1995). An overview of the standards movement. *Phi Delta Kappan, (76)*, 744-750.

[31] Lara-Cinisomo, S. et. al. (Fall, 2004). *A Matter of Class.* Santa Monica, CA: Rand.

[32] Linn, R. L. (1996). Linking assessments. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: promises, problems, & challenges* (pp. 91-105). Mahwah, N.J.: L. Erlbaum Associates.

[33] Linn, R. L. (1996). Work readiness assessment: Questions of validity. In L. B. Resnick & J. G. Wirt (Eds.), *Link school and work: Roles for standards and assessment* (pp. 245-266). San Francisco: Jossey-Bass Publishers.

[34] Milliken, G. A. (1998). *Analysis of messy data*. Boca Raton: Chapman & Hall.

[35] Mitchell (Eds.), *Implementing performance assessment: Promises, problems and challenges* (pp. 23-38). Mahwah, N.J.: L. Erlbaum Associates.

[36] National Commission on Excellence in Education (1983, April). *A nation at risk: The imperative for educational reform* [Electronic Version]. Retrieved November 24, 2004, from http://www.ed.gov/pubs/NatAtRisk/index.html.

[37] Neill, M. (2000, January). Old tests in new clothes. *Instructor,(*109*),* 31-33.

[38] No Child Left Behind Act of 2001, 20 U.S.C. § 6311 (2002).

[39] O'Neil, J. (1993). On the new standards project: A conversation with Lauren Resnick and Warren Simmons. *Educational Leadership, 50*, 18-21.

[40] Pennsylvania Department of Education, Division of Evaluation & Reports. (n.d.). *Writing assessment handbook.* Retrieved November 24, 2004, from

http://www.masd.k12.pa.us/programs/Classroom_Connections_Math_LA/HTML/writing/wrihand/wrihand.htm.

[41] Pennsylvania Department of Education. (2004). Academic Standards for Reading, Writing, Speaking and Listening [electronic version]. Retreived July 8, 2005, from *http://www.pde.state.pa.us*

[42] Reckase, M. D. (2002). Using NAEP to confirm state test results: An analysis of issues [Electronic Version]. In Thomas B. Fordham Foundation (Ed.), *No Child Left Behind: What Will It Take?* Retrieved November 24, 2004, from http://www.edexcellence.net/doc/NCLBreport.pdf.

[43] Reichbach, A. M. (2004). The power behind the promise: Enforcing No Child Left Behind to improve education. *Boston College Law Review, (45)*, 667-704.

[44] Resnick, D. P. & Resnick, L. B. (1996). Performance assessment and the multiple functions of educational measurement. In M. B. Kane & R. Resnick, L. B. (1994). Performance puzzles. *American Journal of Education*, (*102)*, 511-526.

[45] Resnick, L. B., Nolan, K. J., & Resnick, D. P. (1995). Benchmarking education standards. *Educational Evaluation and Policy Analysis, (17)*, 438-461.

[46] Rosenberg, G. N. (1991). *The hollow hope: Can courts bring about social change*. Chicago: University of Chicago Press.

[47] Simmons, W., & Resnick, L. B. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50(5),* 11-15.

[48] Spalding, E. (1995, March 13). The New Standards Project and English Language Arts portfolios: A report on process and progress. *The Clearing House,* (*68)*, 219-225.

[49] Spalding, E. (2000). Performance assessment and the new standards project: A story of serendipitous success. *Phi Delta Kappan,(81)*, 758-764.

[50] Stecher, B.M., Rahn, M. L., Ruby, A., Alt, M.N., Robyn, A.E., & Ward, B. (1997). *Using alternative assessments in vocational education* (1997), Retrieved November 27, 2004, from http://www.rand.org/publications/MR/MR836/MR836.chap4.pdf.

[51] Title I of the Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1995).

[52] Wiley, D. E. & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems and challenges* (pp. 61-89). Mahwah, N.J.: L. Erlbaum Associates.

[53] Wiley, D. E. & Resnick, L. B. (1997). *The New Standards Reference Examination: Standards-referenced scoring system: CSE Technical Report 470*. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[54] Wong, K. K. (2002). Federal educational policy as an anti-poverty strategy. *Notre Dame Journal of Law, Ethics & Public Policy, 16,* 421-446.

[55] Zamora, P. (2003). Children in poverty: In recognition of the special educational needs of low-income families: Ideological discord and its effects upon Title I of the Elementary and Secondary Education Acts of 1965 and 2001. *Georgetown Journal on Poverty, Law and Policy (10),* 413-447.